

A Burden of Proof: Stereotype Relevance and Gender Differences in Math Performance

Ryan P. Brown and Robert A. Josephs
The University of Texas at Austin

Three studies explored gender differences in mathematics performance by investigating the possibility that men and women have different concerns when they take standardized math tests, and that when these gender-specific performance concerns are made relevant, performance may suffer. Results of 3 studies supported these hypotheses. In Study 1, women who believed a math test would indicate whether they were especially weak in math performed worse on the test than did women who believed it would indicate whether they were exceptionally strong. Men, however, demonstrated the opposite pattern, performing worse on the ostensible test of exceptional abilities. Studies 2 and 3 further showed that if these gender-specific performance concerns are alleviated by an external handicap, performance increases. Traditional interpretations of male–female differences on standardized math tests are discussed in light of these results.

[Women] should be taught all sorts of breeding suitable both to their genius and quality . . . in particular, Music and Dancing.

—Daniel Defoe, “The Education of Women”

Despite its sexist implications, the quote above by Daniel Defoe (1661–1731) actually represents a rather liberal and radical position for his era, in that Defoe was arguing that women should, in fact, be educated. Yet, even his well-meaning support of women’s education connoted a belief that has pervaded Western thought for centuries and continues unabated in many non-Western cultures even today: Men and women differ naturally in their “geniuses and qualities,” with men more suited for the harder sciences and mathematics than are women. It is with such stereotypes and their performance-related sequelae that the present research is concerned. In this article we present evidence that men and women have different performance concerns when they take standardized math tests, and that these gender-specific concerns can dramatically affect how men and women perform on such tests.

Gender Differences in Mathematics

Although Western beliefs about gender differences are less severe today than they were in Defoe’s time, stereotypes about the abilities of women and men still support Defoe’s implicit message of female inferiority in certain domains. Research over the last 3 decades suggests that such beliefs about gender differences, par-

ticularly in math abilities, are alive and well, both in the United States and abroad. For example, Ernest (1976) surveyed teachers and found that 63% believed boys were naturally better at math than were girls. None thought girls were naturally better. More recently, Yee and Eccles (1988) reported that parents similarly believed mathematics to be more difficult for their daughters than for their sons. With such parental and institutional support for stereotypes of gender differences in math abilities, it is no wonder that young women are typically less confident in their math abilities (Catsambis, 1994), take fewer math courses (Elmore & Vasu, 1986; Yee & Eccles, 1988), and generally have more negative attitudes toward math (Hyde, Fennema, Ryan, Frost, & Hopp, 1990).

According to statistics provided by the National Education Association (cited in Halpern, 1992), parents’ and teachers’ stereotypes about female deficiencies in math are, on the surface, supported by sex differences in scores on the math subtest of the Scholastic Achievement Test (SAT-M), favoring boys by about 50 points (half a standard deviation). Despite attempts by the Educational Testing Service to eliminate gender-biased test questions, this difference in SAT-M scores has remained relatively constant over the last several decades. Similarly, gender differences on the quantitative section of the Graduate Record Exam (GRE-Q) have favored male test-takers by about 80 points, slightly more than half a standard deviation (Wah & Robinson, 1990). Thus, although some researchers have reported a gradual reduction in gender differences in math over the past several decades (e.g., Feingold, 1988), such encouraging trends have not been apparent on the SAT-M and the GRE-Q, the two most widely used and influential standardized tests of mathematics abilities and achievement in the United States.

Halpern (1992) has recently provided an excellent summary of research on cognitive sex differences, including research on differences in mathematical abilities. Her review describes the nature and degree of such sex differences, dividing causative theories about sex differences into biological and psychosocial theories. Although noting the important role that sex stereotypes may play

Ryan P. Brown and Robert A. Josephs, Department of Psychology, The University of Texas at Austin.

This work was supported in part by Grant SBR-9616526 from the National Science Foundation. We thank Shana Bratcher, Frances Ellington, Marisa Chaves, and Taryl Gillum for their invaluable help in conducting these studies. We also thank Kelli Keough, Bill Swann, and James Pennebaker for their helpful comments on an earlier version of this article.

Correspondence concerning this article should be addressed to Robert A. Josephs at the Department of Psychology, The University of Texas, 330 Mezes Hall, Austin, Texas 78712. Electronic mail may be sent to rbrown@mail.utexas.edu or to josephs@psy.utexas.edu.

in producing sex differences in performance (through differential self-confidence and motivation to take math classes), Halpern lamented the limitations of psychosocial explanations for these differences, particularly in the failure of psychosocial theories to explain why girls typically earn grades in math classes that are at least as high, or higher, than the grades of boys but consistently underperform on standardized math tests (see Kimball, 1989, for a review of research on this discrepancy). Halpern thus sees psychosocial explanations of sex differences in mathematics as incomplete and insufficient.

Psychosocial Explanations Reborn: The Impact of Stereotype Threat

Recent social-psychological research by Steele and colleagues on race differences in standardized testing has shed new light on minority members' underperformance on these tests and may also help to explain gender differences in math from a psychological, rather than a biological, perspective. Steele and Aronson (1995) showed that what they referred to as "stereotype threat" can radically undermine the test performances of Black students. In one study, Steele and Aronson had Black and Caucasian students take a test of verbal intelligence but told half the students that the test was not really diagnostic of intelligence. Even though every student took the same test, Black students performed much better when they believed that the test was nondiagnostic than when they believed it would truly measure their intelligence; no such difference was observed among the Caucasian students. In a related study, Steele and Aronson told students that the test they were taking was nondiagnostic of intelligence and primed the racial identities of half these students prior to testing simply by having them indicate their race on a biographical information form. Congruent with previous results, the Black, race-primed students underperformed as compared with their Black, non-race-primed peers. However, Steele and Aronson observed no such effects of race priming among the Caucasian students. Thus, it appeared that simply reminding Black students of their racial affiliation was sufficient to initiate the racial stereotype threat of inferiority, producing decrements in performance among these academically stigmatized students. Recently, other researchers have also shown that similar stereotype-threat effects can occur when lower socioeconomic status (Croizet & Claire, 1998) or negative stereotypes about elderly people (Levy, 1996) are primed.

The effects that Steele and Aronson demonstrated were theoretically predicated on the vulnerability of their Black students to the evaluation apprehension produced by racial stereotypes about intelligence. The researchers argued that because of cultural stereotypes concerning group differences in abilities, Black students and other minorities have to face substantial self-evaluative anxiety in standardized-test-taking situations. This anxiety, as found in decades of research on general test anxiety (e.g., Wine, 1971), may then impede performance on already stressful standardized tests, creating a self-fulfilling prophecy: The salience of a negative stereotype about one's group leads to self-doubt or concern that one may confirm the group stereotype, which then leads to poorer performance. Ultimately, such vicious cycles may cause stigmatized students to devalue and disidentify with the stigmatized performance domain (Crocker & Major, 1989; Steele, 1997), a

result we have already alluded to in connection with women in mathematics (e.g., Hyde et al., 1990; Yee & Eccles, 1988).

In a series of unpublished studies, Spencer, Steele, and Quinn (cited in Steele, 1997) have verified that similar performance effects can occur with women taking math tests. Spencer et al. argued that because of stereotypes about female math deficiencies, women have to contend with the same kind of evaluation apprehension in math that Blacks contend with in general academics. Furthermore, Steele (1997) has argued that such stereotype-threat effects may occur regardless of direct expectancies for performance; explicit performance expectations may be high for men and women in a given sample, and yet the same women who report high expectations may still suffer from stigma-induced apprehension and consequent poorer performances. Indeed, the populations from which Steele and his colleagues have sampled in their investigations of stereotype threat have arguably been populations high in identification with the performance domains tested, as well as high in expectations for personal performance. Moreover, stereotype-threat effects may occur primarily in people who are highly identified with the performance domain and for whom performance is highly self-relevant. Those for whom the performance domain is not central to their self-concepts may be unlikely to experience self-evaluative concerns when performing in that domain, such that stereotype threat for them is relatively moot (see Brunstein & Gollwitzer, 1996, for a relevant review). To date, these propositions about the moderating role of domain identification in stereotype-threat phenomena have not been systematically investigated, but they do seem reasonable.

Gender-Specific Stereotypes and Their Consequences

The work of Steele and his colleagues has underscored the previously unappreciated impact of social factors in producing group differences in test performances. Hence, this research has been pivotal in refocusing researchers' attention on psychosocial explanations for apparent race and gender differences in cognitive abilities. The next step, however, should be to identify the specifics of the stereotype-threat experience for particular groups (i.e., the nature and content of their performance concerns) so that we may best design strategies for overcoming the effects of stereotype-based threats. The studies we report here are an attempt to accomplish this goal within the domain of gender differences in mathematics.

We hypothesized that gender-specific stereotypes about math abilities might lead to differential performance concerns in men and women. Because women in the United States live with the stereotype that women are bad at math, this gender-specific stereotype might cause women to be concerned about the possibility of confirming their group's negative stereotype when they take standardized math tests. In other words, women may be primarily focused on not performing poorly on standardized math tests rather than on performing very well. Men, however, live with a complementary stereotype—that men are good at math—although this positive stereotype has not been construed heretofore as a burden or stigma because of its positive nature. Nonetheless, consistent with our hypothesis about women, we suspect that men may be concerned about upholding their group's positive stereotype when they take standardized math tests. In other words, men may be primarily focused on performing very well on standardized math

tests rather than on not performing poorly. Thus, both men and women may be concerned about their performances on standardized math tests, but the outcomes with which they are concerned may be qualitatively different.

Several surveys that we have conducted support these predictions. In one such survey we asked undergraduate students in psychology classes to indicate whether they were primarily concerned with the possibility that they might perform poorly, that they might not excel, or neither when they take standardized math tests (e.g., the math subtest of the SAT). The order of the first two responses was counterbalanced across participants. Using this answer format, participants were able to indicate whether the gender-specific performance concerns we have described matched their own experiences, or if they were not particularly concerned about their performance at all. The results revealed a statistically significant difference in the pattern of responses for men and women, with fully half the women indicating a primary concern with the possibility of performing poorly but less than 25% of the men indicating this concern (about a third of the men indicated being primarily concerned with not excelling), $\chi^2(1, N = 208) = 35.27, p < .001$. It is interesting that the most common response among men was "neither of the above." However, we have found that when we asked these questions of men and women at The University of Texas at Austin who were actually about to take a standardized math test, the percentage of respondents who indicated "neither of the above" dropped dramatically—by over 50%—among all respondents. We suspect that part of the reason so many men indicate that they have no performance concerns at all is that doing so allows them to appear confident and competent (a self-presentation strategy), although it could also reflect the truly greater confidence that men tend to have in their math abilities. These results thus offer some support for the validity of our hypotheses, but more so for women than for men.

It is important to note that this pattern of performance concerns seems to be specific to the math domain. When respondents indicated their primary performance concerns about taking standardized tests of their verbal abilities, men and women did not significantly differ in their responses, $\chi^2(1, N = 208) = 3.30, ns$. Furthermore, the interaction between sex and test type (math vs. verbal abilities) was significant in predicting participants' primary performance concerns, $\chi^2(6, N = 208) = 29.50, p < .01$. Although men were again more concerned with the possibility of not excelling than of performing poorly on tests of their verbal abilities, women were also slightly more concerned with not excelling, contrary to their concerns about standardized math tests.

Overview of Studies

In the present studies, we further investigated the validity of our hypotheses about gender-specific performance concerns and, more important, the potential impact of these performance concerns on individuals' actual test performances. Inspired in part by a clever paradigm developed by Tice (1991) for investigating differences between the self-handicapping strategies of people with high or low self-esteem, we targeted the hypothesized differential performance concerns of men and women by manipulating the implications of a math test that we administered to college students. Essentially, to some people we described an upcoming math test as being specially designed to indicate whether they were "excep-

tionally strong" in their math abilities, and to others we described the test as indicative of whether they were "especially weak" in their math abilities.

We predicted that men would be more concerned with their performance on a test that would ostensibly indicate whether they were exceptionally strong in their math abilities relative to a test that would ostensibly indicate whether they were especially weak in their math abilities. Whereas failure on the former test would disconfirm the positive male stereotype that we have described, the possibility of failure on the latter would not be a likely concern for them. For women, on the other hand, we predicted greater concern about performance on a test that would ostensibly indicate whether they were especially weak in math. Failure on such a test would confirm the social stereotype about women in this academic arena and presumably threaten their self-esteem. Although failing a test ostensibly designed only to indicate whether they were exceptionally strong in math might still be somewhat threatening, we nonetheless predicted that, relative to men, women would perceive such failure as being less threatening than failure on a test designed to indicate whether they were deficient in math. For both genders, we expected that the test targeting people's specific performance concerns would pose the greatest threat, which would then manifest itself in their test performance. In Study 1, we investigated these hypotheses using a manipulation of the implications of a math test. In Studies 2 and 3, we examined whether giving individuals an external excuse (a handicap) for poor performance might minimize these performance concerns on either an ostensible test of weak abilities (Study 2) or an ostensible test of exceptional abilities (Study 3).

Study 1: The Impact of Performance Concerns

In this study, we examined the basic hypothesis that the relevance and salience of gender-specific performance concerns might inhibit individuals' performance on a math test that was described to participants as indicative of either very weak math abilities or very strong math abilities. We expected that women would perform less well on a test described as indicative of weak abilities than on a test described as indicative of exceptional abilities, whereas men would show the reverse pattern.

Method

Participants. One hundred twenty-six introductory psychology students (65 women and 61 men) at The University of Texas at Austin participated in this study in return for credit toward a research requirement for their class.¹

Instructions and procedure. Participants were seated in individual cubicles for the entire experiment. After having them read and sign consent forms, the experimenter explained that they would be taking a math test that was in its final stages of development prior to being utilized in the psychology department for placement and tracking purposes. The experimenter then provided participants with a written description of the test and its purposes, which the experimenter later summarized verbally and made

¹ Two additional students participated in this study, but their data were not included in analyses because they failed the manipulation check (i.e., they could not successfully recall the purpose of the test they had taken). Both of these participants were men: 1 in the weak-abilities condition and 1 in the exceptional-abilities condition.

sure that every participant understood. The specific description of the test's purposes served as the experimental manipulation.

Half the participants learned that the math test was designed to indicate whether students were exceptionally strong in their math abilities. The precise description of this test, adapted from Tice (1991), was as follows:

The particular test you will be taking today is designed to help us identify people who are exceptionally strong in their mathematical reasoning abilities. Your performance on this test will not be scored like most normal tests, but rather will be classified as either above or below a predetermined cutoff score. If you score above the cutoff score, this suggests that you are exceptionally strong—in other words, well above average in your mathematical reasoning abilities. However, scoring below the cutoff score tells us little or nothing about your math abilities and math potential. You may be below average, average, or even above average. Thus, this test and the scoring method used with it are designed only to separate those who are especially strong from everyone else.

For participants who received this test description, not scoring above the exceptional cutoff level only meant that a person was not exceptional, not that he or she was necessarily deficient or a failure.

The remaining participants received a different description of the test they were about to take. These participants learned that the test would indicate whether they were especially weak in their mathematical reasoning abilities. The description these participants read was as follows:

The particular test that you will be taking today is designed to identify people who are especially weak in their mathematical reasoning abilities. Your performance on this test will not be scored like most normal tests, but rather will be classified as either above or below a predetermined cutoff score. If you score below this cutoff score, this suggests that you are especially weak—in other words, significantly below average in your mathematical reasoning abilities. However, scoring above the cutoff score tells us little or nothing about your math abilities or math potential. You may be slightly below average, average, or even above average. Thus, this test and the scoring method used with it are designed only to separate those who are particularly weak from everyone else.

For participants who received this test description, scoring above the cutoff level only meant that a person was not especially weak, not that he or she was necessarily strong or highly competent in mathematics.

After the experimenter gave participants one of these two test descriptions and verbally reiterated them, she administered a multiple-choice manipulation check form. Participants indicated on this form what they understood the test was designed to measure. After completing this measure, half the participants also completed a word-fragment completion test modeled after Steele and Aronson (1995). This measure was designed to capture participants' possible test anxiety through 6 word fragments (embedded among 14 filler fragments) that could be completed in esteem-relevant ways (e.g., *idiot*, *flunk*, *dumb*). The remaining participants completed this measure after taking the math test. Finally, the experimenter provided each participant with a test booklet. The test booklets were composed of 20 multiple-choice math problems that we had compiled from old GRE-Q tests. On the basis of national norms for each of these problems, we included a variety of difficulty levels and problem types to minimize the possibility of either ceiling or floor effects on performance in this study. The experimenter advised participants not to guess blindly and informed them that they would have exactly 20 min to complete the test. She then provided them with scratch paper and a scannable form on which they were to indicate their answers to each problem and instructed them to begin. Every participant was timed using a digital stopwatch for accuracy.

After taking the test, each participant completed a brief questionnaire that included a final manipulation check on his or her understanding of the purposes of the test and that asked what their SAT-M score was. The experimenter then questioned each participant for suspicion and described the true purposes of the experiment. No direct performance feedback was ever given to participants, as in all of the studies we report here.

Results

Performance measure. Because we expected that participants' SAT-M scores would be highly related to their performance on the GRE-Q problems used in this study, we wanted to use SAT-M scores as a covariate in our analyses. However, if standardized-testing environments are subject to the effects that Steele and Aronson (1995) have demonstrated in their research, then the SAT-M scores of women may tend to underrepresent these students' actual abilities. Thus, using SAT-M scores as a covariate across gender to adjust mean performance levels for prior ability or preparation differences may be inappropriate, as it would conflate any true differences between men and women with the effects of stereotype threat. For these reasons, we could not reasonably use SAT-M scores to adjust performance levels across gender in an omnibus analysis of covariance (ANCOVA). One way to avoid this conceptual confusion was to test our predictions by first regressing participants' test scores on their SAT-M scores, separately for each gender, and then analyze the residuals from these SAT-M controlled regressions by a standard, general linear model analysis of variance (ANOVA).² This ANOVA thus analyzed the difference between how participants actually performed and how they should have performed, given their SAT-M scores, as a function of their test description condition.

The residualized performance of each participant was analyzed using a 2 (participant gender) \times 2 (test description) ANOVA. This analysis revealed no main effects, but the expected Gender \times Test Description interaction did emerge, $F(1, 122) = 10.82, p < .005, MSE = 7.89$. More important than this overall interaction, though, was that the within-gender comparisons across conditions supported our predictions. As we hypothesized, women who believed that the test would indicate whether they were especially weak in math performed less well than did women who believed the test would indicate whether they were exceptionally strong, $F(1, 122) = 3.97, p < .05$. The reverse was true for men, who performed less well when they believed the test might indicate whether they were exceptionally strong, $F(1, 122) = 7.01, p < .01$. The mean residualized performance levels for men and women in each condition, adjusted within-gender with SAT-M scores, are shown in Table 1, as are estimates of the effect sizes (calculated by Hedge's g). Because these residualized scores may not be as meaningful to readers as actual raw means, Figure 1 also graphically displays participants' raw scores on the test, adjusted within-gender for

² Although we believe this residualized ANOVA analysis is the best way to analyze these data, we also performed a standard ANCOVA at the request of an anonymous reviewer, using participants' SAT-M scores as a simple covariate. This analysis revealed a pattern of results similar to what we found using the residualized ANOVA, and the Gender \times Test Description interaction was still present, $F(1, 109) = 10.65, p < .002$.

Table 1
Mean Residualized Performance in Study 1 as a Function of Gender and Test Description

Gender	Test description		Effect size
	Weak abilities	Exceptional abilities	
Women			
<i>M</i>	-0.68	0.71	.49
<i>n</i>	33	32	
Men			
<i>M</i>	0.97	-0.94	.68
<i>n</i>	30	31	

Note. These residualized means represent the difference in performance from what would be predicted from participants' scores on the math subtest of the Scholastic Achievement Test (SAT-M) as a function of the test description. Separate regressions of test scores on SAT-M scores were conducted for men and women to produce these residualized means. Effect-size estimates are calculated using Hedge's *g*. $MSE = 7.89$.

SAT-M scores. The order of administration of the word-fragment completion test did not affect the test performance results.³

Word-fragment measure. We had also predicted that participants' performance concerns might be reflected in their completion of valenced word fragments. Specifically, we predicted that women would complete more negatively valenced fragments than positively valenced fragments if they believed the test they were about to take or had taken would indicate if they were especially weak in math, relative to women who believed the test would indicate if they were exceptionally strong, and the reverse for the men. However, floor effects on the word-fragment completions yielded a null result of gender, test description, and the interaction of these variables, $F_s < 1$.

Discussion

Despite a floor effect on the word-fragment completion measure, the performance data exactly matched our hypotheses. Specifically, women performed poorly if they believed that they were taking a test designed to indicate whether they were especially weak in math, but men performed poorly if they believed they were taking a test designed to indicate whether they were exceptionally strong in math. Thus, members of each gender suffered in their performance when they believed that the relevance of their group's math-related stereotype was being assessed. For women, this stereotype was that they were deficient in math. For men, this stereotype was that they were strong in math. For both women and men, the salience of their gender-specific performance concerns produced by the test descriptions resulted in a corresponding decrement in performance. These results support our contention that the social stereotype of male-female differences in math abilities may be experienced somewhat differently by men and women, with each group focusing on a different end of the performance continuum because of their group's stigma.

However, an alternative interpretation for the results of Study 1 is possible because of the design of this study. Specifically, the lack of a control condition in which no information about the implications of the math test was provided allows for

the alternative that the performance differences among women was not due to a performance decrement on the weak-abilities test but to a performance enhancement on the exceptional-abilities test, perhaps because the latter description entailed a challenge that caused women to work harder. Similarly, the performance differences among men may not have been due to a performance decrement on the exceptional-abilities test but rather to a performance enhancement on the weak-abilities test. A possible rationale for this alternative among men is that the weak-abilities test led to increased effort by men because of their strong desire not to look deficient in their math abilities. Unfortunately, an internal analysis of the number of problems attempted would not adequately resolve this issue, because this measure could decrease because of effort withdrawal or because evaluation apprehension or self-doubt interfered with performance, thus reducing the number of problems that could be solved in the 20-min time limit.

However, this alternative explanation begs the question of why women would also not want to avoid appearing deficient in their math abilities and thus try harder on the ostensible weak-abilities test, which would result in only a main effect of test description in Study 1. This issue is especially problematic for this alternative explanation for the results of Study 1, given that women typically report lower math self-esteem than men do. Because low math self-esteem would seemingly lead individuals to worry more about avoiding failure on the weak-abilities test (which should be the most likely outcome for them) than about not performing like math geniuses, we do not feel that this alternative explanation for our results is especially compelling. Nonetheless, to disambiguate these possible explanations for our results, we conducted Studies 2 and 3, which further support our hypothesis of gender-specific performance concerns and our interpretation of Study 1.

Study 2: The Computer Crashes

More than 2 decades of research on the phenomenon of self-handicapping (Berglas & Jones, 1978; Jones & Berglas, 1978) suggests that individuals who experience self-doubt in their ability to perform well may put up roadblocks to their own success to minimize the evaluative threat of a potentially poor performance. In essence, such preemptive roadblocks minimize performance threat by providing individuals with an external excuse for failure; they also allow individuals to augment the value of a successful performance, because it would be achieved despite a debilitating handicap. Unfortunately, these self-inflicted wounds also increase the likelihood of failure and thus may create a self-fulfilling prophecy. For example, a student who is afraid of doing poorly on a psychology test might stay up late partying the night before the test. That way, if failure ensues the student has a ready-made excuse: "I didn't fail because I'm incompetent. I failed because I had a hangover."

³ To reduce the impact of outliers, we eliminated responses that were 2 standard deviations or more from their group mean. On the basis of this criterion, performance data from 3 participants (1 female participant in the weak-abilities condition and 1 female and 1 male participant in the exceptional-abilities condition) were excluded from analyses.

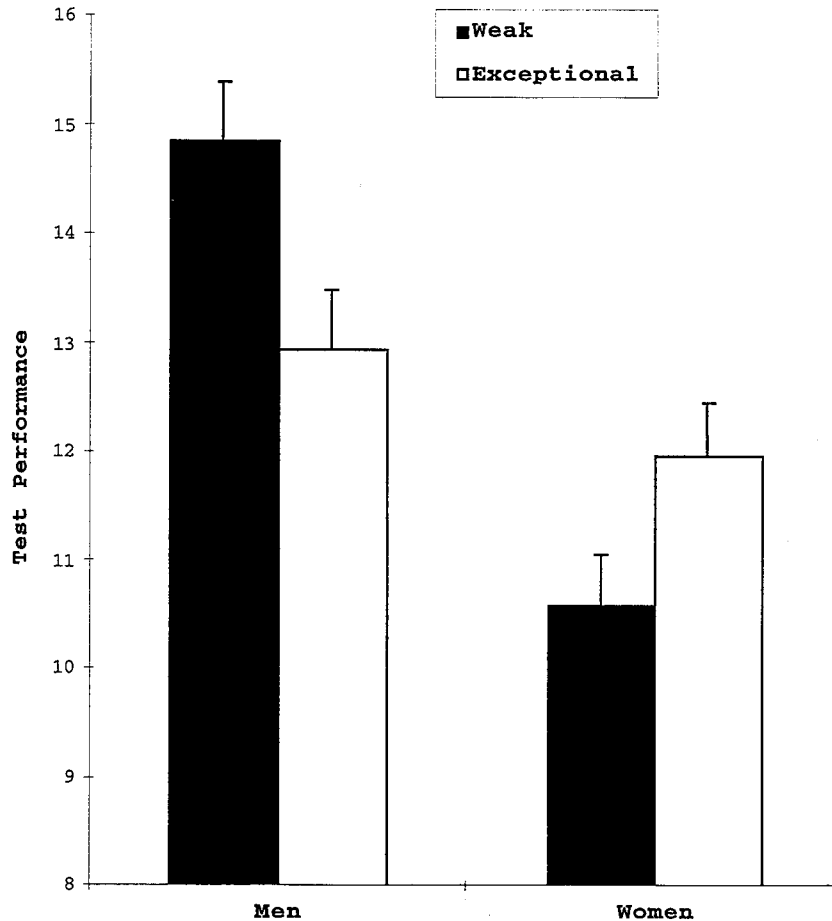


Figure 1. Mean performance levels as a function of gender and test description in Study 1, adjusted within-gender for previous differences in scores on the math subtest of the Scholastic Achievement Test. Vertical lines depict standard errors of the means.

Behavioral self-handicapping has been demonstrated in numerous studies through drug and alcohol ingestion, lack of preparation, a choice of distracting surroundings, and the like (see Higgins, Snyder, & Berglas, 1990, for a review). Researchers have also argued that the claim of personal handicaps, even after a performance, serves the same self-protective, attributional function of behavioral self-handicaps, and thus these verbal excuses should be categorized along with preemptive roadblocks to success (Snyder & Smith, 1982). These claimed handicaps should protect the self-esteem of individuals if their performance proves inadequate but without also increasing the likelihood of a poor performance. Thus, such handicaps would seem to provide the same benefits of behavioral handicaps without all of the immediate costs of true, external roadblocks to success.

In the domain of stereotype-threat research, Steele and Aronson (1995) have shown that stigmatized individuals may claim certain handicaps, such as lack of sleep and lack of concentration, when their stigmatized status is made salient in the context of a standardized test. In their study, Steele and Aronson argued that such claimed handicaps reflected the stereotype-threatened individuals' concerns about the upcoming test. Rather than examining self-handicapping as a response to an experienced

threat, Study 2 was designed to explore the attributional benefits of externally provided handicaps as a function of gender-specific performance concerns in math. Specifically, Study 2 attempted to assess whether actual performance on a math test could be enhanced by giving participants easy access to an attributional escape clause through an externally provided handicap.

Study 2 included a manipulation of the presence of an external handicap within the constancy of a weak-abilities test description. To examine the ability of an external handicap to alleviate the gender-specific performance concerns already outlined, we led some participants to believe that they would be given the opportunity to practice some warm-up problems prior to taking the test. However, a subsequent, ostensible "computer crash" kept them from doing so, providing them with a ready-made, external performance excuse. We predicted that when taking a test described as capable of determining only whether or not people were deficient in their math abilities, women would perform significantly better in the presence of the external handicap (an excuse for failure), but that such a handicap would not significantly improve men's performances because men would not be particularly concerned about such a test.

Method

Participants. Seventy-four undergraduates (39 men and 35 women) at The University of Texas at Austin participated in this study in return for credit toward a research requirement in their introductory psychology class.⁴

Instructions and procedure. As in Study 1, each participant was seated in a private cubicle for the duration of the study. The experimenter also provided participants with the cover story used in Study 1 but described only the weak-abilities test to all participants, using the same description of this test utilized in Study 1. Half the participants went immediately into taking the test following these instructions and were given exactly 20 min to complete it. The rest of the participants were told that warming up before the test on some practice problems would enable them to perform in a way that would indicate their true ability level, and that everyone in this study would be given the opportunity to do so by working on problems set up on the computer in front of them. However, when the experimenter turned on the computer to allow them to begin practicing, they were unable to read the screen (because the computer monitor had been rigged). After appearing to adjust several buttons and knobs on the monitor in vain, the experimenter announced that the participant would just have to go into the test without the benefit of warming up, and that she would indicate this on their answer sheet so that her advisor could decide what to do with their scores himself. This excuse of not being able to practice was designed to reduce participants' evaluative concerns, if they had any such concerns, and thus improve their performance on the test.

All participants had exactly 20 min to work on the same 20 test problems used in Study 1. At the end of the 20 min, the experimenter collected the test booklets, scratch paper, and answer sheets and left the participants with a brief questionnaire to complete while she ostensibly scored their tests (although she would actually remain blind to participants' performances throughout the experiment). This questionnaire assessed participants' memories for the purpose of the test and their SAT-M scores. After a few minutes, the experimenter returned, probed for any suspicions, and debriefed participants about the true nature of the experiment. The experimenter then thanked the participants and dismissed them with an admonition of confidentiality.

Results and Discussion

As in Study 1, we wanted to adjust individuals' test scores with their SAT-M scores to control for individual differences in math ability across experimental conditions. However, as we described earlier, doing this using an ANCOVA would conflate any true differences in ability with the possible effects of stereotype threat experienced by women on the SAT, as Steele and colleagues have demonstrated (Steele, 1997). A residualized ANOVA (as we described in Study 1) gets around this conceptual mire, at least for testing the predictions relevant to our hypothesis of gender-specific performance concerns. As in Study 1 (see Footnote 2), a traditional ANCOVA approach to the results of Study 2 (and Study 3, for that matter) yields similar patterns of results as either a residualized ANOVA or simple, within-gender comparisons, but the ANCOVA also reduces the male advantage on our test by adjusting across gender for SAT-M differences. Because this main effect of gender was not the focus of our investigation and may reflect a conflation of multiple factors, we elected to perform a residualized ANOVA in this study, controlling for differences in SAT-M scores entirely within gender.

Following our predictions in Study 1, we predicted that men would not be particularly concerned with a test that ostensibly would indicate whether they were especially weak in their math abilities, so we did not expect the presence of an externally

Table 2
Mean Residualized Performance on the Weak-Abilities Test in Study 2 as a Function of Gender and Presence of Handicap

Gender	Presence of handicap		Effect size
	Control	Handicap	
Women			
<i>M</i>	-1.30	1.38	.92
<i>n</i>	18	17	
Men			
<i>M</i>	-0.01	0.01	.01
<i>n</i>	20	19	

Note. These residualized means represent the difference in performance from what would be predicted from participants' scores on the math subtest of the Scholastic Achievement Test (SAT-M), as a function of handicap condition. Separate regressions of test scores on SAT-M scores were conducted for men and women to produce these residualized means. Effect-size estimates are calculated using Hedge's *g*. *MSE* = 8.41.

provided handicap to significantly improve their test performances. However, we predicted that women would be concerned with this weak-abilities test, so we did expect the presence of a handicap to significantly improve their test performances. Results confirmed these predictions. After regressing participants' test scores on their SAT-M scores separately for each gender, we submitted the residuals from these regressions to a 2 (gender) \times 2 (handicap) ANOVA. A main effect of the handicap emerged, such that participants who had the external performance excuse of the bogus computer crash performed better on the test than did participants who did not have this excuse, $F(1, 70) = 3.98, p = .05$. However, this main effect was qualified by an interaction between participants' gender and the presence of the handicap, $F(1, 70) = 3.87, p = .05, MSE = 8.43$. The residualized mean performances of men and women (adjusted within-gender for SAT-M scores) are shown in Table 2 as a function of the presence of an external handicap. As predicted, the mean performance of men did not significantly differ when an external handicap was provided, $F(1, 37) = 0.00$, whereas the mean performance of women was significantly improved by the presence of the handicap, $F(1, 33) = 8.17, p < .01$. Figure 2 also displays participants' raw scores on the test, adjusted within-gender for SAT-M scores, for ease of interpretation.⁵

The results of this study supported the performance difference among women shown in Study 1 as a function of the test description. However, rather than manipulating performance by varying

⁴ Four additional students participated in this study, but their data were not included in analyses because they failed the manipulation check (i.e., they could not successfully recall the purpose of the test they had taken). All of these participants were women: 3 in the handicap condition and 1 in the control condition.

⁵ To reduce the impact of outliers, we eliminated responses that were 2 standard deviations or more from their group mean. On the basis of this criterion, performance data from 5 participants (2 female and 2 male participants in the handicap condition and 1 male participant in the control condition) were excluded from analyses. Most of these participants also scored at or below chance on the test, further supporting their exclusion from analyses.

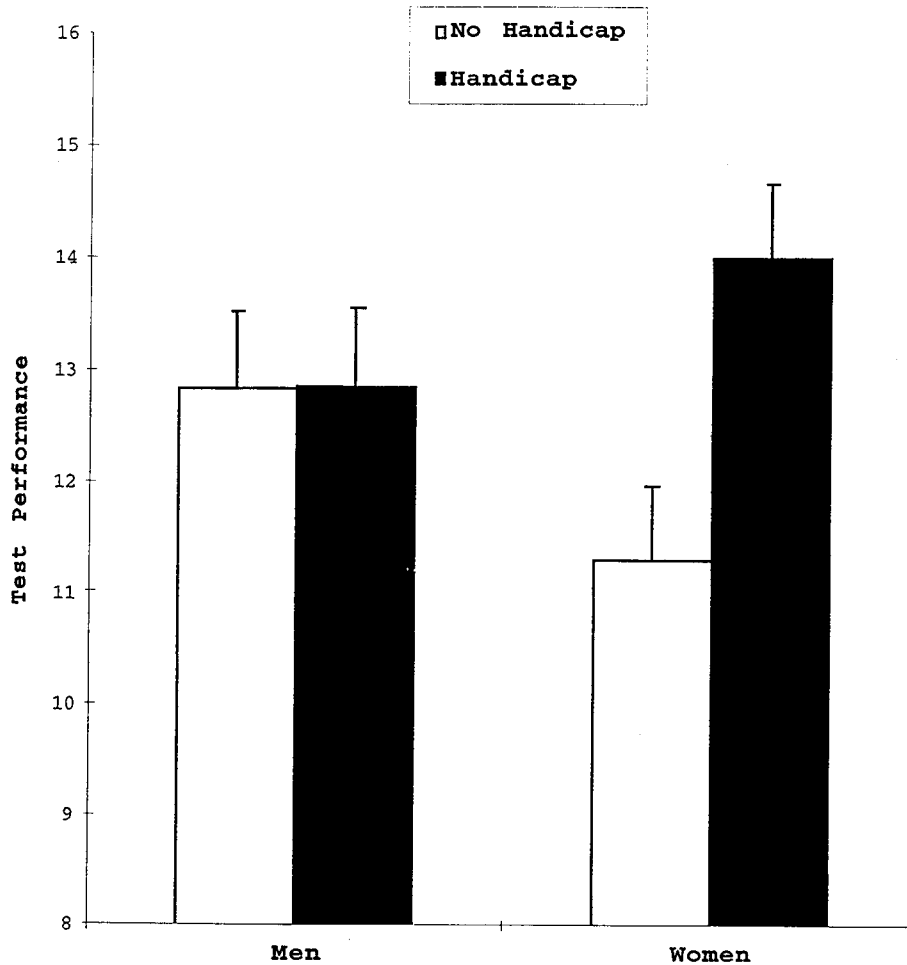


Figure 2. Mean performance levels as a function of gender and the presence or absence of an external handicap in Study 2, adjusted within-gender for previous differences in scores on the math subtest of the Scholastic Achievement Test. Vertical lines depict standard errors of the means.

the implications of the math test, Study 2 showed that performance on the ostensible weak-abilities test could be enhanced simply by providing women with an external performance handicap. Consistent with our predictions, when participants believed they were taking a test designed to indicate only whether they were especially weak in their math abilities, only women were aided by the presence of a performance handicap. Men, whom we did not expect to be as concerned with such a test, were not significantly aided by this handicap. These results not only substantiate our hypothesis of gender-specific performance concerns in mathematics but also provide a comparison of performance in an experimental condition designed to enhance it with performance in a control condition. Thus, the results of Study 2 suggest that the performance difference among women found in Study 1 was not due to a performance enhancement on the exceptional-abilities test but rather a performance decrement on the weak-abilities test, as we hypothesized.

Study 3: Computer Crash Redux

Study 2 helped to clarify the interpretational ambiguities inherent in the design of Study 1. Nonetheless, Study 2 unlocked only half the design of Study 1, in that it involved only the use of the weak-abilities test description. Although Study 2 was probably the most efficient test of the primary concerns raised by the alternative interpretation of Study 1, Study 3 was designed to complete the picture of performance decrements resulting from the salience and relevance of the hypothesized, gender-specific performance concerns—in particular, the performance concerns of men. As in Study 1, we expected that women would not be especially concerned about performance on a test described as indicative only of exceptional math abilities, but that men would be concerned about such an outcome. Hence, we expected that only men would benefit from an external performance excuse on the exceptional-abilities test because such a performance excuse should alle-

viate the evaluation apprehension that we expected only men would experience.

In addition, previous theorizing by Steele (1997) suggested that the effects we have hypothesized might only occur with people who are somewhat identified with mathematics, for only such people would care about the results of a math test in a psychology study. Thus, as we suggested in the introduction to this article, domain identification may moderate the gender-specific performance effects that we have described, as well as stereotype-threat effects more generally. With this possibility in mind, we created a simple, five-item scale to measure the extent to which people were personally identified with mathematics (see Brown & Josephs, 1997, for a complete description and construct validation of this scale), and we used this instrument to "select-out" people who were especially disidentified with math. Thus, Study 3 used the exceptional-abilities test description from Study 1 and the bogus computer crash manipulation introduced in Study 2, but only with participants who cared somewhat about their math abilities.

Method

Participants. Sixty-nine undergraduates (33 women and 36 men) at The University of Texas at Austin participated in this study in return for credit toward a research requirement in their introductory psychology class. These participants were all prescreened on the basis of having scored above the theoretical midpoint of our math questionnaire, which we administered along with a battery of questionnaires at the beginning of the semester. On this scale, participants responded to five simple questions concerning their personal identification with mathematics, such as how important their math abilities were to them, how important it was to them that others believe they are good at math, and how important math abilities will be to them in their future careers. This five-item scale demonstrated good internal reliability ($\alpha = .82$), and pilot testing showed that people's scores on this scale correlated significantly with their ratings of the importance of performing well on a standardized math test within an actual testing situation (see Brown & Josephs, 1997). The prescreening process eliminated 19% of the men and 24% of the women in the participant pool, who did not appear to care very much about mathematics. In the pool from which the sample in this study was drawn, men ($M = 5.29$, $SD = 1.64$) and women ($M = 5.19$, $SD = 1.70$) did not differ significantly in their self-reported math identification, despite a very large sample size, $F(1, 2415) = 1.77$, $p < .19$. Although this result was somewhat surprising, it is also the case that more women than men take introductory psychology at The University of Texas at Austin (about 50% more women than men), so perhaps many men who are highly math-identified simply were not included in the participant pool.⁶

Instructions and procedure. As in Studies 1 and 2, each participant was seated in a private cubicle for the duration of the study. The experimenter also provided participants with the cover story used in Studies 1 and 2 but described only the exceptional-abilities test to all participants, using the description of this test from Study 1. Half the participants immediately began taking the test following these instructions and were given exactly 20 min to complete it. The remaining participants were told, as in Study 2, that warming up before the test on some practice problems would enable them to perform in a way that would indicate their true ability level, and that everyone in this study would be given the opportunity to do so by working on problems set up on a computer in front of them. After communicating these instructions, the experimenter initiated the same bogus computer-crash ploy described in Study 2, thus providing half the participants with an external excuse for their test performance.

All participants had exactly 20 min to work on the same 20 test problems used previously. At the end of the 20 min, the experimenter collected the test booklets, scratch paper, and answer sheets and left the participants with

Table 3
Mean Performance on the Exceptional-Abilities Test in Study 3
as a Function of Gender and Presence of Handicap

Gender	Presence of handicap		Effect size
	Control	Handicap	
Women			
<i>M</i>	0.59	-0.55	.40
<i>n</i>	16	17	
Men			
<i>M</i>	-0.83	0.93	.62
<i>n</i>	19	17	

Note. These residualized means represent the difference in performance from what would be predicted from participants' scores on the math subtest of the Scholastic Achievement Test (SAT-M), as a function of handicap condition. Separate regressions of test scores on SAT-M scores were conducted for men and women to produce these residualized means. Effect-size estimates are calculated using Hedge's g . $MSE = 7.95$.

the same questionnaire used in Study 2, which they were to complete while he ostensibly scored their tests (although he actually remained blind to participants' performances throughout the experiment). After a few minutes, the experimenter returned, probed for any suspicions, and debriefed participants about the true nature of the experiment. The experimenter then thanked the participants and dismissed them with an admonition of confidentiality.

Results and Discussion

Complementary to Study 2, we predicted that women would not be particularly concerned with a test that ostensibly would indicate whether they were exceptionally strong in their math abilities, so we did not expect the presence of an externally provided handicap to significantly improve their test performances. However, we predicted that men would be more concerned with this exceptional-abilities test, so we did expect the presence of a handicap to significantly improve their test performances. Results confirmed these predictions. Only an interaction between gender and handicap was statistically significant in the residualized ANOVA, $F(1, 65) = 4.55$, $p < .04$, $MSE = 7.95$. Whereas the mean performance of women in the handicap condition did not improve, but actually decreased slightly, $F(1, 31) = 1.05$, *ns*, the handicap did improve the mean performance of the men, $F(1, 34) = 4.75$, $p < .04$. These residualized means are shown in Table 3 as a function of the presence of an external handicap. Participants' raw test scores are shown in Figure 3, adjusted within-gender for differences in SAT-M scores.

General Discussion

Previous research has confirmed that men and women differ in their math self-confidence, perhaps because of internalization of

⁶ An additional 6 students participated, but their data were not included in analyses because either they were suspicious of the experiment (1 male and 1 female participant) or they failed the manipulation check (i.e., they could not correctly recall the purpose of the test they had taken: 2 male participants in the handicap condition and 1 male and 1 female participant in the control condition).

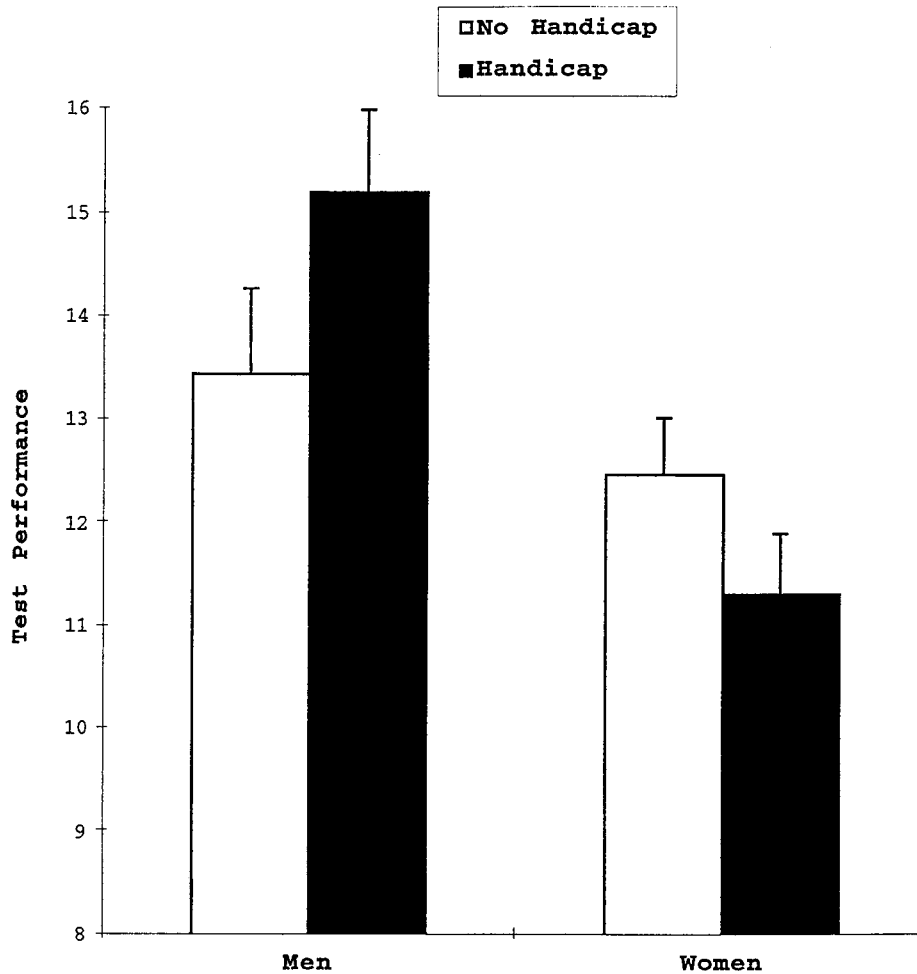


Figure 3. Mean performance levels as a function of gender and the presence or absence of an external handicap in Study 3, adjusted within-gender for previous differences in scores on the math subtest of the Scholastic Achievement Test. Only participants who reported caring somewhat about their math abilities (in an earlier pretesting session) were included in this study. Vertical lines depict standard errors of the means.

social stereotypes about gender differences in math abilities (e.g., Eccles, 1987), or perhaps because social stereotypes and other factors inhibit women from taking as many math courses as men take, leading to relatively poor performance on standardized tests and making women's relative lack of confidence in math somewhat warranted. We suggest that the social stereotype of male superiority in mathematics may be experienced differently by men and women, causing men to focus on their group's positive stereotype in some situations and women to focus on their group's negative stereotype. We further suggest that these different relative foci can lead men and women to differ in their performance-related concerns when they take math tests. Specifically, men may sometimes be more concerned with confirming the positive stereotype that they should excel in math, whereas women may sometimes be more concerned with disconfirming the negative stereotype that they are deficient in math. Thus, both men and women may be concerned with the consequences of their performances, but the outcomes with which they are concerned may differ qualitatively.

The three studies we have reported support our hypotheses

about the effects of gender-specific performance concerns in math. Study 1 demonstrated that the scores of men and women could suffer if their gender-specific performance concerns were made relevant through the description of the implications of their test performance. Studies 2 and 3 supported these results by showing that an external handicap on the particular test that captured their gender-specific performance concerns enhanced participants' performance on that test. In sum, these studies highlight the potentially serious consequences of the gender-specific performance concerns we have suggested.

One question about the effects we have reported here, raised by an anonymous reviewer, is whether these effects might differ as a function of past experience with standardized math tests. Specifically, perhaps previously poor performance on the SAT-M might lead only some women to be concerned with the possibility of performing poorly, whereas others who scored very well on the SAT-M might actually exhibit performance concerns like those of the men in our studies. Similarly, perhaps only men who had previously scored poorly on the SAT-M were concerned about not

excelling in our studies. To test these possibilities, we reanalyzed our data in all three studies, categorizing participants as high or low scorers on the SAT-M by within-gender median splits and entering this past performance variable as a factor in our analyses. In these analyses, the past performance variable exerted a main effect on test scores, such that high scorers performed better in all three of our studies than did low scorers. Past performance, however, did not moderate the Gender \times Condition interactions in any of these studies (all three-way F s $< .30$). Past performance on the SAT-M also failed to moderate the Gender \times Condition interactions we reported when we conducted regression analyses, keeping SAT-M score a continuous variable (all three-way F s < 1). Nonetheless, it may be that something related to previous performance, such as math self-esteem, might moderate (or mediate, within-gender) the effects we have reported here. We are currently investigating this possibility.

The magnitude of our effects is an important aspect of these results, especially given that performance in these studies had no practical consequences for our participants: They knew that their performance on the test would not affect their grades or be reported to anyone outside of the experiment. Effect sizes (measured by Cohen's d) for gender differences on the SAT-M and GRE-Q are typically .50 and .60, respectively (Halpern, 1992; Wah & Robinson, 1990). In the studies we have reported, manipulating the relevance and salience of gender-specific performance concerns or manipulating the presence of an external performance excuse produced within-gender effects as large or larger than these heralded, national gender differences in performance (as large as .92). Thus, not only did we conceptually replicate in Studies 2 and 3 the debilitating effects that Study 1 demonstrated, we also produced effects of an ecologically valid size. Furthermore, as can be seen in Figure 2, our manipulation of an external handicap even eliminated the typical gender difference in performance, such that women who had been given an external performance excuse (the computer crash) performed just as well as men did in the same condition. The reader should note that this between-gender comparison does not control for previous gender differences in SAT-M scores (which, as noted earlier, conflate any real ability differences with the debilitating effects of stereotype threat on SAT scores). Controlling for SAT-M scores across gender does increase the performance difference favoring women in the handicap condition of Study 2, although we believe this ANCOVA approach is problematic for reasons we have already described. Failure to take into account prior performance on the SAT-M results in a nonsignificant difference favoring women, but clearly women performed at least as well as men did in this study.

In terms of ecological generalization, the results we have reported here invoke a host of questions about the nature of gender differences in math. For instance, it may be that the typical standardized-testing environment tends to make most salient and relevant the possibility of very poor performance (i.e., the performance concerns of many women). In this case, at least some of the male-female gap in performance on the SAT-M and GRE-Q could be attributable to the effects of gender-specific performance concerns, as well as to differences in general test anxiety and performance focus (as suggested by the survey results that we have collected, which have consistently shown that men report less concern about their performance in general than women do). Further research should examine the developmental and situational

factors that might enhance or inhibit the differential performance concerns that we have described. For example, some tests and even entire math classes may be perceived by students as "weed-out" tests or classes, which may then lead to performance effects analogous to what we have shown with our weak-abilities test description. Other tests, such as those used in primary schools to qualify students for gifted and talented programs or advanced-level classes may likewise produce performance effects analogous to those of our exceptional-abilities test description.

Future research should also examine the effects of performance concerns in situations that involve practical consequences for performance (e.g., when grades or access to college are at stake). The minimal practical consequences of participants' performances in the studies we have reported here highlight the potential importance of the effects we have demonstrated. How much greater might these effects be when grades, college entrance, or jobs are at stake? In addition to issues such as these, we are currently examining the relevance of these gender-specific performance concerns for self-handicapping strategies prior to testing. Our preliminary results suggest that men and women may actually prepare less prior to a test that makes relevant their gender-specific performance concerns. Thus, they may be less prepared for the test that concerns them more to provide themselves with a ready excuse for their performance.

The present research extends previous work by Steele and his colleagues in several ways. First of all, these studies demonstrate gender differences in performance using entirely different methods than those employed by Steele and others, methods that more directly invoke the gender-specific performance concerns held by men and women. Second, the studies we have reported here show that stereotype-based performance concerns can affect both women and men. Whereas previous research has offered a limited demonstration of the effects of stigmatization on negatively stigmatized groups (e.g., Blacks, women), our research suggests that stigmatization may be an "equal-opportunity destroyer," such that even individuals who belong to positively stereotyped groups can have their performance suffer if the confirmation of their group's stereotype is put to the test. Ultimately, the effects of stigmatization may boil down to basic evaluation apprehension, to which anyone should be susceptible under the right circumstances. Given this possibility, we are further investigating the effects of positive stereotypes about Asian Americans, who might experience the pressure of stigmatization in a manner similar to that demonstrated by men in general in the studies reported here. The intersection of gender stereotypes and ethnic stereotypes about math abilities (e.g., Asian American females, who perform under the shadow of a negative stereotype about their gender but a positive stereotype about their ethnic group) makes for an intriguing set of possibilities for research on stereotype threat and self-categorization.

Coda

American society, like most Western cultures, has indeed come a long way in what it teaches about cognitive gender differences since the days when Defoe argued for the (limited) education of women. Still, social stereotypes about the abilities of men and women remain and may exert powerful effects on the targets of those stereotypes. Although it is possible, as some have argued (e.g., Benbow, 1988), that biological factors may account for part

of the variance underlying gender differences in performance in certain domains, it is clear that within-group differences are much larger than between-group differences in most domains. In addition, some have made similar biological arguments about ethnic differences in academic performance (e.g., Herrnstein & Murray, 1994), but recent research on stereotype threat by Steele and Aronson (1995) suggests the likelihood of such explanations is greatly exaggerated. Whether or not biological differences underlie gender differences in mathematics, the mere suggestion of between-group differences may lead to a self-fulfilling prophecy in which the threat of failure promotes poor performances among the stigmatized. If educators and policy makers are ever to level the academic playing fields in this country, it is paramount that we understand the nature of stigmatization and of the different performance concerns that individuals may have about standardized tests. Only then will society begin to assess with any accuracy a person's real "genius and quality."

References

- Benbow, C. P. (1988). Sex differences in mathematical reasoning ability in intellectually talented preadolescents: Their nature, effects, and possible causes. *Behavioral and Brain Sciences, 11*, 169–232.
- Berglas, S. C., & Jones, E. E. (1978). Drug choice as a self-handicapping strategy in response to noncontingent success. *Journal of Personality and Social Psychology, 36*, 405–417.
- Brown, R. P., & Josephs, R. A. (1997). *The importance of importance: Measuring identification with mathematics*. Manuscript in preparation, The University of Texas at Austin.
- Brownstein, S. C., Weiner, M., Green, S. W., & Hilbert, S. (1988). *How to prepare for the Graduate Record Examination: GRE general test* (8th ed.). New York: Barron's.
- Brunstein, R. F., & Gollwitzer, P. M. (1996). Effects of failure on subsequent performance: The importance of self-defining goals. *Journal of Personality and Social Psychology, 70*, 395–407.
- Catsambis, S. (1994). The path to math: Gender and racial-ethnic differences in mathematics participation from middle school to high school. *Sociology of Education, 67*, 199–215.
- Crocker, J., & Major, B. (1989). Social stigma and self-esteem: The self-protective properties of stigma. *Psychological Review, 96*, 608–630.
- Croizet, J. C., & Claire, T. V. (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.
- Eccles, J. S. (1987). Gender roles and women's achievement-related decisions. *Psychology of Women Quarterly, 11*, 135–172.
- Elmore, P. B., & Vasu, E. S. (1986). A model of statistics achievement using spatial ability, feminist attitudes and mathematics-related variables as predictors. *Educational and Psychological Measurement, 46*, 215–222.
- Ernest, J. (1976). *Mathematics and sex*. Berkeley: University of California Press.
- Feingold, A. (1988). Cognitive gender differences are disappearing. *American Psychologist, 43*, 95–103.
- Graduate Record Examinations Board (1990). *Practicing to take the GRE general test* (No. 8). Princeton, NJ: Educational Testing Service.
- Halpern, D. F. (1992). *Sex differences in cognitive abilities*. Hillsdale, NJ: Erlbaum.
- Herrnstein, R. J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life*. New York: Free Press.
- Higgins, R. L., Snyder, C. R., & Berglas, S. (1990). *Self-handicapping: The paradox that isn't*. New York: Plenum.
- Hyde, J. S., Fennema, E., Ryan, M., Frost, L. A., & Hopp, C. (1990). Gender comparisons of mathematics attitudes and affect: A meta-analysis. *Psychology of Women Quarterly, 14*, 299–324.
- Jones, E. E., & Berglas, S. C. (1978). Control of attributions about the self through self-handicapping strategies: The appeal of alcohol and the role of underachievement. *Personality and Social Psychology Bulletin, 4*, 200–206.
- Kimball, M. M. (1989). A new perspective on women's math achievement. *Psychological Bulletin, 105*, 198–214.
- Levy, B. (1996). Improving memory in old age through implicit self-stereotyping. *Journal of Personality and Social Psychology, 71*, 1092–1107.
- Snyder, C. R., & Smith, T. W. (1982). Symptoms as self-handicapping strategies: The virtues of old wine in a new bottle. In G. Weary & H. Mirels (Eds.), *Integrations of clinical and social psychology* (pp. 104–127). New York: Oxford University Press.
- 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.
- Tice, D. (1991). Esteem protection or enhancement? Self-handicapping motives and attributions differ by trait self-esteem. *Journal of Personality and Social Psychology, 60*, 711–725.
- Wah, D. M., & Robinson, D. S. (1990). *Examinee and score trends for the GRE general test: 1977–78, 1982–83, 1986–87, and 1987–88*. Princeton, NJ: Educational Testing Service.
- Wine, J. (1971). Test anxiety and direction of attention. *Psychological Bulletin, 76*, 92–104.
- Yee, D. K., & Eccles, J. S. (1988). Parent perceptions and attributions for children's math achievement. *Sex Roles, 19*, 317–333.

Received July 16, 1998

Revision received August 20, 1998

Accepted August 24, 1998 ■