Journal of Social Sciences 7 (4): 538-549, 2011
ISSN 1549-3652

Gender, Stereotype Threat and Mathematics Test Scores

[1]Ming Tsui, [2]Xiao Ying Xu and [3]Edmond Venator
[1]Department of Sociology-Anthropology,
Millsaps College Jackson, MS 39210-0001,
[2]Department of Physics Wuhan University of Technology Wuhan, China,
[3]Emeritus Professor of Psychology, Millsaps College USA

Abstract: Problem statement: Stereotype threat has repeatedly been shown to depress women's scores on difficult math tests. An attempt to replicate these findings in China found no support for the stereotype threat hypothesis. Our math test was characterized as being personally important for the student participants, an atypical condition in most stereotype threat laboratory research. **Approach:** To evaluate the effects of this personal demand, we conducted three experiments. **Results:** Experiment 1, where in Chinese students were tested with the added independent variable of test importance. Our results produced only marginally significant stereotype threat effects. Experiment 2, a replication of experiment 1, yielded completely different results, with no threat effects at all. Math-test scores were significantly higher in the threat condition for both men and women, consistent with the phenomena of stereotype lift and stereotype reactance. Experiment 3, which did not include the test-important variable, yielded no significant effects. **Conclusion:** Stereotype threat, in the mathematics domain, does not seem to be a problem for women in China. We discuss our results in terms of factors which moderate stereotype threat and societal differences in the U.S. and China.

Keywords: Gender gap, stereotype threat, societal differences, gender parity, graduate education, laboratory setting

INTRODUCTION

Since 1982, women have surpassed men in American college enrollment and graduation every year and are rapidly achieving gender parity in many traditionally male-dominated academic fields (Halpern *et al*., 2007). Since 1986, at the graduate school level, female enrollment has grown 2% each year, compared with 1% for men. In 2008, 61% of graduate students in the U.S. were women. Women outnumber men in all major fields of graduate education, except math, computer sciences, engineering, physical sciences and business (Snyder and Dillow, 2009).

However, despite these successes, women still score lower than men on the math section of the high-stakes standardized tests used for admissions to college and graduate school including the SAT and the Graduate Record Examination (Halpern *et al*., 2007). Although women generally receive higher grades than men in high school and college, women underperform men in math and the physical sciences when tests are not closely related to material that has been previously taught (Willingham and Cole 1997; Halpern *et al*., 2007). The mean difference between men and women on the math portion of the SAT (SAT-M) has remained virtually unchanged for the past 35 years, with men outscoring women by an average of 38 points (College Board, 2009). In 1972, the mean SAT-M score for women was 38 points lower than that for men; by 2009, this difference was 35 points. Among the top scorers (700 points or higher out of a possible 800) on the 2009 SAT-M, men outnumbered women almost 2-1. This math-gender gap may be one reason behind the persistent sex segregation at the doctoral level in graduate education. Although the numbers of earned doctorates awarded to women went from 14% in 1971 to nearly 50% in 2006, in the fields of math, engineering and economics, the percentages of all doctoral degree recipients who were women were only 21, 12 and 30% respectively in 2006 (England, 2010).

In a meta-analysis of 100 studies published between 1963 and 1987, Hyde *et al*. (1990a) observed a complex pattern regarding gender differences in math performance. At the elementary and middle school levels, girls are superior to boys in computation and equal to boys in understanding mathematical concepts. Gender differences favoring boys emerge in high school on problem-solving tasks and the differences persist on SAT-M. The magnitude of gender differences in math

Corresponding Author: Ming Tsui, Department of Sociology-Anthropology, Millsaps College Jackson, MS 39210-0001 USA

performance grows larger with more selective samples; while gender-math differences were moderate for samples of college students (d = 0.33), the differences were much larger (d = 0.54) for samples of students from highly selective colleges and graduate students.

Among mathematically gifted children, boys have consistently outscored girls in math. Data collected from 1972 through 1991 in the Study of Mathematically Precocious Youth (SMPY) showed that among the intellectually talented 12 and 13 year-old American youths who scored 700 or more on the SAT-M, there were 13 boys for every 1 girl (Lubinski and Benbow 1992). Since then, this gender-math gap has narrowed considerably, but there is still a 4:1 boy-girl ratio among those scoring 700 or more on the SAT-M (Halpern *et al*., 2007).

While some scholars have suggested that biological factors underlie the gender gap in math (e.g., Benbow 1988; Halpern *et al*., 2007), most researchers have argued that the gender gap is a function of socio-cultural factors (e.g., Hyde *et al*., 1990b; Keller, 2001; Stage and Maple 1996; Tiedemann, 2000).

Stereotype threat and group differences on standardized tests: Steele and Aronson (1995) introduced an intriguing phenomenon, stereotype threat, to explain racial differences on standardized test scores. In four laboratory experiments using Stanford University undergraduate students, Steele and Aronson (1995) showed that "making African American participants vulnerable to judgment by negative stereotypes about their group's intellectual ability depressed their standardized test performance relative to White participants, while conditions designed to alleviate this threat, improved their performance, equating the two groups once their differences in SATs were controlled". To explain this phenomenon, Steele and Aronson (1995) posited that a well-known negative stereotype can become a self induced threat to the members of the stereotyped group, depressing their performance on tasks that are the target of this specific stereotype. In their experiments, Steele and Aronson (1995) found that the stereotype threat was more evident when the threat was made salient by instructions telling participants that the test measured their cognitive ability. Pointing to the fact that the African-American participants in their study were strong students who identified with the material on the test, Steele and Aronson (1995) speculated that stereotype threat may have a particularly negative effect on the academically more able members of the stereotyped groups. Identifying with the domain in question, strong students may be more anxious to not confirm the stereotype than weak students; and such fear may lead them "try hard with impaired efficiency," resulting in low test scores (Steele and Aronson 1995). Based on these findings, they suggested that stereotype threat may offer at least a partial explanation for the persistent gap in standardized test scores between black and white students.

Spencer *et al*. (1999) were the first to study the effects of a math-gender stereotype threat on women's math performance in a laboratory setting. Using female students from elite universities, they found that women scored significantly lower than equally qualified men on a difficult math test when they were told that there were gender differences on the test, but performance differences "could be eliminated" when the test givers "lowered stereotype threat by describing the test as not producing gender differences" (Spencer *et al*., 1999,). Because this experiment demonstrated that the math-gender stereotype threat can dramatically impair the test performance of high-math-ability women, Spencer *et al*., suggested that stereotype threat may underlie the consistent gender differences in advanced math performance. Inasmuch as their study (and many subsequent laboratory studies) was conducted at elite American colleges and universities with participating students who were good at math, stereotype threat seems to offer a plausible explanation for the gender differences among high-math-ability students. To further test the nature and strength of the gender-math stereotype threat, several laboratory studies used procedures to explicitly reject the negative math-gender stereotype and found significant improvement of female participants on difficult math tests (McIntyre *et al*., 2003; Pronin *et al*., 2004; Walton and Cohen, 2003).

Carr and Steele (2009) utilized two classic psychological problem-solving situations-the Luchins Water-jar task and the Wisconsin Card Sorting Test (Berg, 1948)-to show that stereotype threat engenders inflexibility. Under threat conditions their participants exhibited significantly more maladaptive perseverance than did participants under reduced threat conditions.

Gender and mathematics in China: In China, the belief that women are weaker than men in math has a long history. Despite a continuous official "women holding up half of the sky" campaign since the 1950s and a consistent government effort promoting equal education for women and men, most Chinese, both men and women, still see math and science as a male domain (Broaded and Liu, 1996). Today few top mathematicians, engineers and natural scientists in China are women. In 2004, among all university faculty members and scientists who served as doctoral-program advisors, only 9% were women Educational Statistics

Yearbook of China. In academic senior high schools, boys outnumber girls in the science track, while in the humanities track, 80% of the students are girls. This gender gap is surprising given the fact that there are no gender differences in the mean math score on the Chinese College Entrance Examination, an equivalent of the SAT (Tsui, 2007).

Testing stereotype threat hypothesis in China: Venator (2008) conducted a study in China to examine the cross-cultural generalizability of stereotype threat involving gender and math among Chinese students. Because this study was published in China, in Chinese, the descriptions which follow are somewhat more detailed than typical accounts of previously published research.

In a 2 (gender)×3 (threat) complete factorial design, the participants worked on a difficult math test (our dependent variable). Details concerning the math test can be found in the method section of experiment 1 presented below. The first page of each test booklet included instructions and questions. Embedded in the instructions was one of three statements concerning gender norms for the math test (boys better than girls, no gender difference, girls better than boys), our threat manipulation. Our participants (196 men and 84 women) were biology, physics and computer science majors from three universities in Wuhan, China.

The experimenter told the students they would be taking a math test with questions taken from the GRE math subject test, which is taken by students who apply for admission to study mathematics at the graduate level. They were also told that the test would be part of their term evaluation and would be compared with students from other universities. The experimenter then read aloud the gender norms that were printed on the first page of the test booklet. At the completion of the test, the students were told that the test they had just taken was part of a research project and that the statements about gender were not true. They were also told that they would receive course credit for their participation, but that their individual scores would not be counted toward their term grades and thanked for their help with the research.

A two-way (Gender x Threat condition) Analysis Of Variance (ANOVA) was used to evaluate the participants' math-test performance (number of items correctly solved). The predicted Gender x Threat interaction was not significant, nor was the main effect of gender. The only significant factor was the main effect of threat. Multiple comparisons revealed a significant difference between the no-gender-difference and the girls-better-than-boys groups. Separate analyses for males and females revealed that the mean score for female participants in the girls-better-than-boys group was significantly lower than the mean score for females in the no-gender-difference group. There were no significant differences among the males. An analysis of covariance, using scores from the math portion of the Chinese College Entrance Exam as the covariate, produced the same pattern of significance.

We were surprised by the overall lack of gender differences in math scores; female students did just as well as male students on our challenging test and on their College Entrance Exam. The absence of support for the predictions generated by stereotype threat theory, a theory with consistent support in dozens of experiments, was especially striking.

Stereotype threat in high-stakes settings: Sometime after the acceptance of our (2008) article, a colleague directed our attention to stereotype threat studies that, unlike most stereotype threat research, had been conducted in operational (real world), high personal-stakes situations. The quasi-experimental research of Stricker and Ward (2004) reported on two independent studies: one using a nationwide sample of classes that include White, Black, Asian and Hispanic students, of both genders, taking either the AP Calculus or AB Examination, the other using a large sample of community college students who were taking the Computer Placement Tests (CPTs). Their experimental manipulation involved the placing of questions concerning ethnicity and gender either before or after the relevant test. Stricker and Ward concluded that their data, analyzed in terms of both statistical and practical significance, offered no support for stereotype threat theory. A re-analysis of the data by Danaher and Crandall (2008) disputed this conclusion and presented several specific criticisms of the criteria and analyses used by Stricker and Ward. In a rejoinder to Danaher and Crandall, Stricker and Ward (2008) answered each of the criticisms and stated that their original conclusion is justified. Readers interested in stereotype threat theory, or in the more subtle, controversial aspects of data analysis and decision making will find this sequence of articles informative.

A different approach to evaluate the generalization of stereotype threat from laboratory to operational (high-stakes) settings was taken by Cullen *et al*. (2004). They used the differential prediction paradigm and two sets of archival data to evaluate stereotype threat theory. From the College Board they obtained SAT scores and freshmen grades and the U.S. Army provided them with predictor and criterion scores from

a large-scale project. Neither of these data sets was consistent with predictions derived from stereotype threat theory and the authors suggested caution in generalizing stereotype threat effects.

A concern that domain identification (Steele, 1997) was not individually ascertained in the (2004) study led to a follow-up study. Cullen *et al*. (2006) used the same College Board data set that was used in the Cullen *et al*. (2004) study, but classified students as math-identified or non-math-identified on the basis whether or not their intended college major was in a math-related field. Again, their results did not support predictions derived from stereotype threat theory and they restated their caution about generalizing from laboratory to real-world settings.

Beilock *et al*. (2007) conducted a series of experiments investigating the role of working memory in the performance decrement associated with stereotype threat. In the fifth experiment of their series they showed that the deficits in working memory experienced by women due to stereotype threat when taking a math test can spill over and negatively impact performance on a subsequent unrelated task. They concluded that their findings have "important implications for the sequencing of sections on the GRE and SAT. Walker and Bridgeman (2008) took advantage of the "nonconstant ordering of subject matter tests on the SAT" (p. 2) to conduct a quasi-experimental evaluation of spillover effects in an operational, high-stakes setting. The SAT form they used had math, reading, or writing as the second section; writing is always the first section. The examinees they selected all had a critical reading test as their third section. Thus they were able to compare scores on a critical reading test that had been immediately preceded by a test of mathematics, reading, or writing. In order to fulfill the criterion of math identification, the analyses relevant to the phenomenon of spillover used only those participants who were confident that they would pursue a math-related major in college (N = 19,507). Walker and Bridgeman (2008) found no support for the spillover hypothesis.

Good *et al*. (2008) conducted a stereotype threat experiment using college students enrolled in an upper-level calculus class. These student participants took a difficult practice test containing questions covering the same content as an upcoming course examination and were told by their professor that they would receive extra credit on their examination based on their performance on the practice test. This contingency was repeated by the researcher just before the practice test was administered. Thus the students believed their course grade could be affected by their scores on the practice test, creating a personal-stakes test setting. Actually, all students in the experiment received the same number of extra credit points. All participants read a statement informing them that the test was a measure of math abilities, the stereotype threat manipulation. Those participants randomly selected to take the test under the no-threat condition read additional information stating that men and women have performed equally well on the test. Under the threat condition men and women performed equally well, while women in the non-threat group scored significantly higher than women in the threat group and higher than men in both the threat and non-threat groups. These findings, in contrast with the other studies cited above document stereotype threat effects in a real-world, personal-stakes setting.

More recently, to demonstrate the latent ability of African-American students, Walton and Spencer (2009) conducted two meta-analyses using data obtained in real-world testing situations and found that "under conditions that reduce stereotype threat, stereotyped students performed better than no stereotyped students at the same level of past performance.

In our Chinese study (2008) the tested students had been told by the test administrators that the obtained scores would be used in computing their grades at the end of the semester. So, from the students' point of view they were in a real-world, personal-stakes setting, rather than the typical laboratory setting we had envisioned when designing the experiment. Some of the findings cited above (Cullen *et al*., 2004; Cullen *et al*., 2006; Stricker and Ward, 2004; Walker and Bridgeman, 2008) suggested to us that the lack of stereotype threat effects obtained with our Chinese students may have been a function of the personal importance of the scores. One explanation for the lack of stereotype threat effects observed in high-stakes settings is that students are more motivated to do well, masking any effects of stereotype threat (Cullen *et al*., 2004; Cullen *et al*., 2006). To explore the dimension of test importance as it relates to stereotype threat we designed an experiment in which both of these variables were manipulated. Our hypothesis was that only those females in the test unimportant condition would be affected by stereotype threat.

Test importance and stereotype threat in China:

Experiment 1: Design. We used a 2 (gender)×2 (threat)×2 (test importance) complete factorial experimental design. Importance in this experiment pertains to personal stakes, that is, whether or not the test scores have an impact on students' grades. Manipulations of the threat and test importance variables were via instructions read to students. The dependent variable was a difficult math test.

Participants: Our subject pool consisted of 188 (54 women and 134 men) sophomore physics and chemistry majors who were taking the same required physics class at Wuhan University of Technology (WUT) in Wuhan, China. Due to curricular differences between China and the United States, the math problems in the GRE general examination are too easy for Chinese college students. Based on our pilot test results, we decided that even advanced GRE math may not be challenging enough for Chinese math majors. So we decided to use sophomores majoring in physics or chemistry, who had completed one year of college calculus.

At the undergraduate level in WUT there are approximately three times as many male physics majors as there are female and our original design called for randomly selecting males to create experimental groups with about the same number of men and women. Our Chinese colleague (overseeing the data collection) informed us that it would be very awkward to exclude students, so all students were tested. Because math ability is crucial for physics majors and important for chemistry majors, our presumption was that the students in our pool were good at math and had a personal identification with math.

Materials: Explicit in stereotype threat theory is a boundary condition related to task difficulty; stereotype threat deficits are obtained only with difficult tests (Spencer *et al*., 1999; Steele and Aronson, 1995). For this reason, we decided to use questions from the advanced math test of the GRE (GRE Math Subject Test). We selected 20 questions, ranging from easy to difficult; from the third edition of Princeton review's cracking the GRE Math Subject Test (Leduc, 2005). We first translated all 65 practice-test questions in the preparation book into Chinese. Based on the recommendation of a first-year computer-science graduate student and 5 college sophomores in various science and business majors (all six of them took the entire test), we divided the test questions into "easy," "moderately difficult," and "difficult." The questions we used in the study consisted of 6 easy questions, 11 medium-difficult questions and 3 difficult questions.

To preclude having test floor or ceiling effects in the experiment, we are tested a group of 146 college sophomores majoring in information technology. On this pilot test, students, who were told to attempt every problem and given 20 m. to work, obtained scores (number correctly solved) ranging from 3-17.

Procedure: The participants were randomly assigned to the four Threat-importance treatment combinations in such a way that each cell contained approximately one-fourth of the men and one-fourth of the women. Group lists were typed on a transparency. When the students arrived for the experiment they were told to check the projected transparency and assemble at the location indicated for their group. Four junior physics professors, each randomly assigned to one of the four treatment groups, took the students to their appointed rooms. These professors distributed the testing materials and read instructions to the participants. These instructions, presented below, constituted the manipulation of our independent variables: test importance and gender norms.

Instructions for threat-test not important: A new high-school math competition examination is being developed. Today you will be taking a test to help us evaluate the appropriateness of these questions. These questions are of special interest to us because in the past, boys have done better than girls on them.

Instructions for no threat-test not important: A new high-school math competition examination is being developed. Today you will be taking a test to help us evaluate the appropriateness of these questions. These questions are of special interest to us because in the past, there have been no gender differences on the performance of these questions.

Instructions for threat-test personally important: Today you will be solving math problems. Your score will be included in the calculation of your term grade, so please do them very carefully.

As you may know, there has been some controversy about whether these are gender differences in math ability. Previous research has sometimes shown gender differences and sometimes shown no gender differences. So far, male students have done better than female students with these questions.

Instructions for no threat-test personally important: Today you will be solving math problems. Your score will be included in the calculation of your term grade, so please do them very carefully.

As you may know, there has been some controversy about whether there are gender differences in math ability. Previous research has sometimes shown gender differences and sometimes shown no gender differences. So far no gender differences have been found when using these questions.

Additionally, all participants were given the following instructions: There are a total of 20 questions and you have 20 m. to complete the test. You are not allowed to leave the room early; if you finish the test early, use the remaining time to carefully check your answers. You are required to write down your solutions on the spaces provided under each question. Additional scratch paper is also provided. You must also write down your name, gender and your course number.

At the completion of the test, the students were told that the test they had just taken was part of a research project and that the statements about gender differences were not true. They were also told that they would all receive course credit for their participation, but that their individual scores would not be counted toward their term grades and thanked for their help with the research.

RESULTS

A 2×2×2 between-groups analysis of variance (ANOVA), using problems correctly solved as the dependent variable, yielded no significant factors or interactions, though there were effects of marginal significance. Women in the Threat Group achieved lower scores than did women in the No-Threat Group and men exhibited no threat-related differences, producing a marginally significant Gender x Threat interaction. Both men and women scored higher in the test-not-important condition resulting in a marginally significant main effect of Importance. A 2×2×2 analysis of covariance, using the Chinese College-Entrance Exam math scores as the covariate also failed to reveal any significant effects, though again there was a marginally significant Gender-Threat interaction and a marginally significant main effect of Importance.

To further explore the relationships among gender, threat and test importance we ran separate analysis of variance for each level of importance. In addition to the marginally significant effects already reported, there was one significant simple effect; men in the test-not-important characterization had significantly higher scores under the threat condition (M = 9.78) than the no-threat condition (M = 8.63), F (10.61) = 4.587, p = 0.036.

The scores of the women on our math test and the math portion of the Chinese College-Entrance Exam were higher than those of the men, significantly so (p = 0.038) with the College-Entrance scores. The correlation (r) between the math portion of the College-Entrance Exam and our math-test scores is 0.197, p = 0.008.

Clearly the results of this experiment are inconsistent with our hypotheses; we expected to observe threat effects among the women students in the test-not-personally-important condition. Certainly all of our participants, men and women, had strong math backgrounds and abilities. Certainly, given that they were all physics and chemistry majors, math competency was important to them. But these are precisely the qualities, in women, that have been shown to be correlated with stereotype threat effects (e.g., Inzlicht and Ben-Zeev, 2003; Martens *et al*., 2006; Spencer *et al*., 1999).

Because of the marginally significant results, especially the Gender x Threat interaction, we decided to replicate the experiment. The replication was carried out in the fall of 2009. This was two years after the data were collected for experiment 1; we were concerned that a replication in 2008 would encounter too much risk of students in the second study being aware of what had been done the year before.

Experiment 2: The same design, materials and procedures used in experiment 1 were used in experiment 2. The only difference was, of course, the participants. The participants were WUT sophomore physics majors, 96 men and 45 women.

A2×2×2 between-groups ANOVA yielded an outcome pattern radically different from that observed in experiment 1. The main effect of stereotype threat was statistically significant: F (1,133) = 12.549, p = .001. The scores were higher (M = 8.89) under the threat condition than under the no threat condition (M = 7.41), d = 0.59. The main effect of importance was also significant: F (1,133) = 4.597, p = 0.034. The mean test scores were 8.58 under the important condition and 7.70 under the not important condition (d = .34). There were no significant or marginally significant interactions.

A 2 × 2 × 2 ANCOVA, using College-Entrance Test scores as the covariate, revealed significant main effects of stereotype threat: F (1,132) = 16.853, p<0.001 and importance: F (1,132) = 8.201, p = 0.005. The directions of difference for the main effects were the same as those observed with the ANOVA. The ANCOVA also revealed a significant Gender x Stereotype Threat interaction: F (1,132) = 4.061, p = 0.046. Both men and women achieved lower math-test scores under the no threat condition, but the threat-no threat difference was larger for the women than the men. Analyses (ANCOVA) of simple effects showed that the difference for women was significant: F (1, 42) = 10.238, p = 0.003; d = 0.79, while the difference for men was only marginally significant: F (1, 93) = 3.062, p = 0.083; d = 0.50. Thus the interaction is a function of differing magnitudes of effect rather than differing directions of effect.

The pattern of results is presented in Fig. 1. The correlation (r) between the experimental math-test scores and scores on the math portion of the Chinese college entrance exam is 0.477 (p<0.001). Women had a slightly higher mean math-test score, while men had a slightly higher mean college entrance test score.

Once again, we were surprised by the results of our study; there is not even a hint of stereotype threat. We had expected a pattern of results similar to those obtained in experiment 1, intending to combine the data from the two experiments into a single analysis for increased power.

Obviously, given the completely different outcomes, there is nothing to be gained by combining the two data sets.

The data from experiment 2 are consistent with two phenomena related to stereotype threat: stereotype lift and stereotype reactance. Stereotype lift refers to

enhanced performance produced by the activation of a stereotype that is positive for members of the target group (for example, Shih *et al*., 1999; Shih *et al*., 2002; Walton and Cohen, 2003). If, for the men in our experiment, the explicit males-better-than-females characterization functioned as a lift manipulation then we would expect the higher scores observed under the threat condition. In experiment 1 man exhibited a significant stereotype lift pattern, but only in the test unimportant condition. There is not even a suggestion of stereotype lift for men in the test important condition.

Stereotype reactance refers to a condition wherein members of a negatively stereotyped group exhibit enhanced performance when the stereotype is activated (for example, Kray *et al*., 2001; Kray *et al*., 2004; Wei, 2009). The results for the woman participants in experiment 2 are consistent with stereotype reactance, but the results from experiment 1 are in the opposite direction.

Because of the inconsistency between the data collected in 2007 and 2009, we conducted a third experiment in January of 2011. Because we had access to a limited number of women students and the overall lack of effects of the Test Importance variable, we decided to exclude test importance.

Experiment 3:

Design: We used a 2 (gender) ×2 (threat) complete factorial design in which Threat was again manipulated via instructions read to the students. The dependent variable was the same math test used in experiments 1 and 2.

Participants: The participants were 113 WUT physics majors: 84 men and 29 women.

Materials: The math test used was the same as in experiments 1 and 2.

Procedure: Students were, by gender, randomly divided into two groups: Threat (41 men and 15 women) and No Threat (43 men and 14 women). The details of the data collection were the same as for experiments 1 and 2, except there were two sets of instructions rather than four.

Instructions for threat: A new high-school math competition examination is being developed. Today you will be taking a test to help us evaluate the appropriateness of these questions. These questions are of special interest to us because in the past, boys have done better than girls on them.

Instructions for no threat: A new high-school math competition examination is being developed. Today you will be taking a test to help us evaluate the appropriateness of these questions.

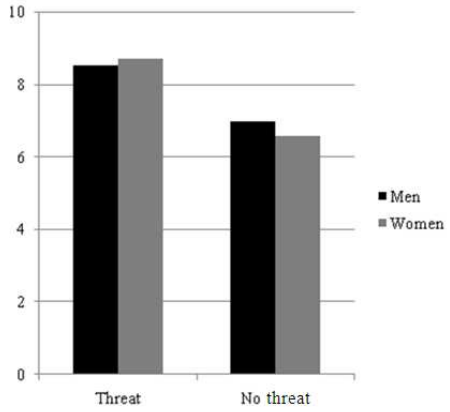

Fig. 1: Experiment mean math-test scores asa function of gender. Stereotype threat and personal importance

These questions are of special interest to us because in the past, there have been no gender differences on the performance of these questions.

Results: A 2×2 between-groups analysis of variance (ANOVA), using number of problems correctly solved as the dependent variable yielded no significant or marginally significant, main effects or interactions. An analysis of covariance, using College Entrance scores as the covariate also failed to reveal any significant or marginally significant effects.

DISCUSSION

Based on the data generated by the three experiments in the present study and our study, a reasonable conclusion would be that the test-score-depressing math-gender stereotype threat, reliably demonstrated in many laboratory studies, is not much of a threat for urban Chinese women. The remainder of our study explores plausible reasons for the lack of stereotype threat effects among our participants.

Math-gender stereotype: The potential for stereotype threat is, of course, predicated on the existence of a widely held stereotype. That men are better at math than women is a widely held belief in China (Broaded and Liu, 1996). Early in the semester during which experiment 1 was conducted students filled out a general questionnaire which contained an item asking about the math-gender stereotype: 50 of the 54 women who participated in experiment 1 indicated that they were aware of the stereotype. Unfortunately, that questionnaire was not administered in 2009 to the participants in experiment 2. After the participants in experiment 3 completed the math test they responded to a multi-item questionnaire which revealed that 25 of the 26 women were aware of the stereotype.

It has been argued that awareness of a stereotype is sufficient to produce stereotype threat; belief in or acceptance of the stereotype is not necessary (Steele, 1997; Aronson *et al*., 2002). However, some recent research has shown that the extent to which women personally endorse the math-gender stereotype is related to social compassion processes (Blanton *et al*., 2002) and stereotype threat (Schmader *et al*., 2004). In reporting the results of their experiment Schmader *et al*., (2004) stated, "An unexpected finding was that women who rejected the stereotype did not just show a weaker effect of the stereotype threat manipulation, but actually showed no effect at all" (2004). While Chinese women are well aware of the stereotype, we have no data suggesting that they believe it to be true. In our informal interviews with Chinese students the answer to our questions concerning the math-gender stereotype inevitably indicated knowledge of the stereotype, but most of the women expressed disbelief. One of the items in the post-test questionnaire completed by students in experiment 3 asked "Do you believe that boys are better than girls in math?" Of the 25 women who were aware of the stereotype only 2 believed it to be true. It seems reasonable to consider lack of belief in the stereotype as an explanation for the lack of stereotype threat effects in the math-test performance among Chinese women.

Amelioration of threat effects: Several empirical studies have clearly shown attenuation, sometimes elimination, of threat effects via pre-test manipulation.

Ability and effort: experimental evidence has shown that the effects of stereotype threat can be significantly reduced by manipulations exposing students to information that focuses on the malleability, rather than the fixedness, of intellectual capacities. Good *et al*. (2008) showed that young women who were encouraged by college-student mentors to view intelligence as a malleable trait earned significantly higher standardized test scores than women in a control condition. Dar-Nimrod and Heine (2006) conducted an experiment demonstrating that the impact of stereotype threat on math-test scores can be significantly reduced by focusing attention on the malleability as opposed to the rigidity of the traits underlying gender differences in math. Thoman *et al*. (2008) designed an experiment to determine whether exposure to different presumed mechanisms underlying gender differences in math would have an impact on stereotype threat effects. Consistent with their hypothesis they found that college women informed, by reading a fictitious article, that the source of male superiority in math is effort rather than ability performed better on a math test than did their classmates who read that male superiority in math is due to innate ability.

While American parents and children tend to believe math ability is innate (Stevenson *et al*., 1990), Tsui and Rich (2002) found that 94 percent of Chinese eighth graders chose "hard work" as the most important factor for academic success. The negative effects of stereotype threat on Chinese women may be reduced by a cultural belief that academic achievement depends on hard work, rather than innate ability.

Counter-stereotype gender roles: Good *et al*. (2010) conducted an experiment with high-school students to evaluate their hypothesis that the presentation of counter-stereotype information reduces the effects of stereotype threat. The students, both male and female, read the same pages from chemistry textbook and then took a 12-item comprehension test. Their stereotype manipulation was via the content of three photographs embedded in the textual material: stereotypic = three lone male scientists; counter-stereotypic = three lone female scientists; mixed = one lone male scientist, one lone female scientist and one image of a male and a female scientist working together. The female students obtained significantly higher scores in the counter-stereotypic condition than in the stereotypic condition. The effect for males, though not statistically significant, was in the opposite direction. Mean test scores under the mixed condition were virtually equivalent (8.37 for females and 8.25 for males).

Test characterization: Alter *et al*. (2010) conducted two experiments in which ability tests were characterized as challenges. Black school children took an age-appropriate standardized math test with racial salience (low vs. high) and test characterization (diagnostic vs. challenge) as independent variables. Under the high salience (threat) condition students did significantly better when the test was characterized as a challenge. In a second experiment, using White college students as participants, Alter *et al*. (2010) compared students from poorly represented high schools with students whose high schools were well represented at a prestigious university and, in a preliminary analysis, found that students from under-represented schools were more anxious and felt more threatened. Participants were randomly assigned to one of four experimental conditions based on salience of high school (high or low) and test characterization (diagnostic Vs challenge) and given a test composed of questions from the quantitative section of the GRE. Participants from under-represented schools did more poorly than students from high-represented school, but only when school salience was high (threat condition) and the test were characterized as being diagnostic.

Field studies: Over the years there have been several field studies implementing manipulations designed to improve the academic performance of African-American students by reducing threat. Steele (1997) presented a series of strategies, situational changes he termed wise schooling, predicted to reduce the stereotype threat experienced by African-American students. These strategies included optimistic teacher-student relationships, challenge as opposed to remediation, a stress on the "expansiveness" of intelligence, affirmation of domain belongingness and the building of self efficacy. In a pilot program, using randomly selected students, he implemented several of these practices and found, using GPAs as the criterion, a small, but not significant, benefit for White students in the program and a significant benefit for Black students in the program. Aronson *et al*. (2002), used a sample of African-American and Caucasian undergraduates (both men and women) to determine whether the belief that intelligence is not fixed, but rather "an expandable capacity" reduces the negative effects of stereotype threat. Their experimental group watched a video describing the way the brain grows in response to intellectual challenge, then wrote pen-pal letters to persuade simulated middle-school students, who indicated they were having difficulties in school, that intelligence is malleable, not a fixed capacity. There were two control groups: members of one wrote pen-pal letters indicating that intelligence is composed of several different talents; members of the second did not compose pen-pal letters. At the end of the semester, African-American, but not White, students in the intelligence-is-malleable group earned significantly higher GPAs than did their cohorts in the two control groups. Cohen *et al*. (2006) conducted two field experiments in which they manipulated the content of a brief, in-class writing assignment. Students (seventh graders) were randomly assigned to write an essay on values most important to them (self-affirmation condition) or on values least important to them (control condition). While there was no reliable effect on European-American students, they found significantly higher GPAs for African-Americans in the self-affirmation condition than for those in the control condition.

Since its initial demonstration stereotype threat has been conceptualized as a situational variable (Aronson *et al*., 1999; Steele, 1997) and the studies cited in this section corroborate its situational nature. Indeed, in much of the early research, safe-from-threat conditions were created by the rather simple expedient of telling participants that test they were about to take had not produced racial/gender differences in the past (McIntyre *et al*., 2003; Pronin *et al*., 2004; Walton and Cohen, 2003) or by characterizing tests as being non-diagnostic of ability (Good *et al*., 2008; O'Brien and Crandall, 2003; Steele and Aronson, 1995). A consistent theme has been that a significant proportion of group differences in standardized tests and academic grades stem from the effects of situational engendered threat rather than from inherent characteristics of group members. Given the range of situational factors that have been shown to either nullify stereotype threat or diminish its effects and individual differences, e.g., in self monitoring levels, that reduce the effects of stereotype threat, it is perhaps not surprising that young women of a culture vastly different from our Western culture might be immune to the threat induced by a negative stereotype.

China: Sociopolitical Milieu. Since 1950 and the emergence of communism there has been a consistent official stress on gender equality. Mao Zedong wrote articles advocating women's rights as early as 1919 (Witke, 1967). This striving for gender equality has been much more successful in urban, as compared with rural, areas. According to Whyte and Parish (1985) while full gender equality had not yet been achieved in China, things were better among urban women than should be expected. They went on to say "…we would argue that a primary reason for this favorable picture is that Chinese cities are so highly bureaucratized and the sorts of modern structural changes that make progress toward equality possible are further advanced than one would expect for a country as poor as China. Though there are exceptions, e.g., the mandatory retirement age for women is younger than for men, Chinese women grow up, are educated and work in a society that sanctions, fosters and even officially demands gender equality.

Chinese education: A demanding and rigorous national curriculum at elementary and secondary school levels and frequent testing may help to explain the absence of stereotype threat for Chinese women and the lack of gender differences in our difficult-math-test scores. In China, math is considered the most important academic subject and the Chinese education system emphasizes math instruction and the training of math teachers. Most Chinese elementary and secondary school math teachers majored in mathematics or math-related fields when in college. When applying for admission to a normal or a teacher's college, applicants in China are required to choose an academic major (in China, education is not a major). Future teachers spend their time learning the content of their major and related subjects and are expected to teach the subject they trained for. Because the national math curriculum demands that everyone teaches the same material at the same time, new and experienced

teachers often prepare lessons together and learn from each other. As a result, math teachers in China are generally competent and skillful. With well trained, experienced teachers and a lot of practice (daily math classes and homework) beginning in the first grade, most Chinese students do not see math as a difficult subject and rank math as one of their most liked subjects (Tsui, 2005).

In China, high standardized test scores are required for admission to academic high schools and universities. The College Entrance Exam is the sole consideration in admission decisions and because entering a top university is the ticket to high-paying jobs, the exam has been called the test that determines one's life. During their college careers scholarships are awarded on the basis of cumulative test scores. Chinese students feel comfortable with math and are accustomed to taking difficult tests.

Inzlicht *et al*. (2006) showed that the typical stereotype threat effects, engendered by being in the minority (gender) when taking a difficult math test, were moderated in women who scored high on a self monitoring scale. High self monitors are "sensitive to the demands of social situations and adept at regulating their expressive behavior and self-presentation to project desired public appearance. Perhaps, young Chinese women who, in the presence of a math-gender stereotype, learn to achieve at the same level as men, develop self monitoring skills that help diminish the effects of threat.

One-child policy: Finally and perhaps most importantly, the lack of stereotype threat effects in our experiments is likely related to China's one-child-per-family policy and its resultant gender-neutral parental expectations and high educational aspirations for both boys and girls. Tsui and Rich (2002) found that, unlike previous generations wherein daughters were less valued than sons, daughters in one-child families are valued as highly as sons. Combined with the advent of capitalism in China, with its attendant elimination of state jobs and security, the one-child policy has produced a generation for which gender plays a markedly diminished role; it is a unique social experiment.

CONCLUSION

The combination of a government policy that stresses gender equality, an education system that provides teachers highly trained in their subject matter, along with a rigorous national curriculum and a one-child policy that has forced parents and grandparents, to treat their children with diminished regard for gender seems to have substantially weakened, if not eliminated, the ravages of math-gender stereotype threat and even overall math-gender differences in modern, urban China. While there are certainly gender differences rooted in biology, the pervasive under performance in math among women observed in most Western nations may not be one of them.

REFERENCES

Alter, A.L., J. Aronson, J.M. Darley, C. Rodriguez and D.N. Ruble, 2010. Rising to the threat: Reducing stereotype threat by reframing the threat as a challenge. J. Exper. Soc. Psychol., 46: 166-171. DOI: 10.1016/j.jesp.2009.09.014

Aronson, J., C. Fried and C. Good, 2002. Reducing the effects of stereotype threat on African American college students by shaping theories on intelligence. J. experimental, Social, Psychol., 38: 113-125. DOI: 10.1006/jesp.2001.1491

Aronson, J., M. Lustina, C. Good, K. Keough and C.M. Steele et al., 1999. When white men cannot do math: Necessary and sufficient factors in stereotype threat. J. Exper. Soc. Psychol., 35: 29-46. DOI: 10.1006/jesp.1998.1371

Beilock, S. L., R.J. Rydell and A.R. McConnell, 2007. Stereotype threat and working memory: mechanisms, alleviation and spillover. J. Exper., Psychol. General, 136: 256-276. DOI: 10.1037/0096-3445.136.2.256

Benbow, C.P., 1988. Sex differences in mathematical reasoning ability among the intellectually talented. Behav. Brain, Sci. 11: 169-232. DOI: 10.1017/S0140525X00078365

Berg, E.A., 1948. A simple objective technique for measuring flexibility in thinking. J. General Psychol., 39: 15-22. DOI: 10.1080/00221309.1948.9918159

Blanton, H., C. Christie and M. Dye, 2002. Social identity versus reference frame comparisons: The moderating role of stereotype endorsement. J. Exper. Soc. Psychol., 38: 253-267. DOI: 10.1006/jesp.2001.1510

Broaded, C.M. and C. Liu, 1996. Family background, gender and educational attainment in urban China. China Quar., 145: 53-86. DOI: 10.1017/S0305741000044131

Carr, P.B. and C.M. Steele, 2009. Stereotype threat and inflexible perseverance in problem solving. J. Exp. Soc. Psychol., 45: 853-859. DOI: 10.1016/j.jesp.2009.03.003

Cohen, G.L., J. Garcia, N. Apfel and A. Master, 2006. Reducing the racial achievement gap: A social-psychological intervention. Science, 313: 1307-1310. DOI: 10.1126/science.1128317

College Board, 2009. College-Bound Seniors: Total Group Profile Report. College Board Inspiring Mind.

Cullen, M.J., C.M. Hardison and P.R. Sackett, 2004. Using SAT-grade and ability-job performance relationships to test predictions from stereotype threat theory. J. Applied, Psychol., 89: 220-230. DOI: 10.1037/0021-9010.89.2.220

Cullen, M.J., S.D. Waters and P.R. Sackett, 2006. Testing stereotype threat theory predictions for ath-identified and non-math-identified students by gender. Hum. Perform., 19: 421-440. DOI: 10.1207/s15327043hup1904_6

Danaher, K. and C.S. Crandall, 2008. Stereotype threat in applied settings re-examined. J. Applied Soc. Psychol., 38: 1639-1655. DOI: 10.1111/j.1559-1816.2008.00362.x

Dar-Nimrod, I. and S.J. Heine, 2006. Exposure to scientific theories affects women?s math performance. Science, 314: 435-435. DOI: 10.1126/science.1131100

England, P., 2010. The gender revolution: Uneven and stalled. Gender Soc., 24: 149-166. DOI: 10.1177/0891243210361475

Good, C., J. Aronson and L.A. Harder, 2008. Problems in the pipeline: Stereotype threat and women?s achievement in high-level math courses. J. Applied Develop. Psychol., 29: 17-28. DOI: 10.1016/j.appdev.2007.10.004

Good, J.J., J.A. Woodzicka and L.C. Wingfield, 2010. The effects of gender stereotypic and counter-stereotypic textbook images on science performance. J. Soc. Psychol., 150: 132-147. DOI: 10.1080/00224540903366552

Halpern, D.F., C.P. Benbow, D.C. Geary, R.C. Gur and J.S. Hyde et al., 2007. The science of sex differences in science and mathematics. Psychol. Sci. Public Interest, 8: 1-51. DOI: 10.1111/j.1529-1006.2007.00032.x

Hyde, J.S., E. Fennema and S.J. Lamon, 1990a. Gender differences in mathematics performance: A meta-analysis. Psychol. Bull., 107: 139-155. DOI: 10.1037//0033-2909.107.2.139

Hyde, J.S., E. Fennema, M. Ryan, L.A. Frost and C. Hopp, 1990b. Gender comparisons of mathematics attitudes and affect: A meta-analysis. Psychol. Women, Quarterly, 14: 299-324. DOI: 10.1111/j.1471-6402.1990.tb00022.x

Inzlicht, M. and T. Ben-Zeev, 2003. Do high-ability female students underperform in private? The implications of threatening environment on intellectual processing. J. Educ. Psychol., 95: 796-805. DOI: 10.1037/0022-0663.95.4.796

Inzlicht, M., J. Aronson, C. Good and L. McKay, 2006. A particular resiliency to threatening environments. J. Exp. Soc. Psychol., 42: 323-336. DOI: 10.1016/j.jesp.2005.05.005

Keller, C., 2001. Effect of teachers stereotyping on students stereotyping of mathematics as a male domain. J. Soc. Psychol., 141: 165-173. DOI: 10.1080/00224540109600544

Kray, L.J., L. Thompson and A.D. Galinsky, 2001. Battle of the sexes: Gender stereotype confirmation and reactance in negotiations. J. Person. Soc. Psychol., 80: 942-958. DOI: 10.1037//0022-3514.80.6.942

Kray, L.J., R. Jochen, A.D. Galinsky and L. Thompson, 2004. Stereotype reactance at the bargaining table: The effect of stereotype activation and power on claiming and creating value. Person. Soc. Psychol. Bull., 30: 399-411. DOI: 10.1177/0146167203261884

Leduc, S., 2005. Cracking the GRE Math Subject Test. 1st Edn., Random House, New York, ISBN 0-375-76491-7

Lubinski, D. and C.P. Benbow, 1992. Gender differences in abilities and preferences among the gifted: Implications for the math-science pipeline. Curr. Direct. Psychol. Sci., 1: 61-66. DOI: 10.1111/1467-8721.ep11509746

Martens, A., M. Johns, J. Greenberg and J. Schimel, 2006. Combating stereotype threat: The effect of self-affirmation on womens intellectual performance. J. Exp. Soc. Psychol., 42: 236-243. DOI: 10.1016/j.jesp.2005.04.010

McIntyre, R.B., R.M. Paulson and C.G. Lord, 2003. Alleviating women mathematics stereotype threat through salience of group achievements. J. experimental, Social, Psychol., 39: 83-90. DOI: 10.1016/S0022-1031(02)00513-9

O?Brien, L.T. and C.S. Crandall, 2003. Alleviating women's mathematics stereotype threat through salience of group achievements. J. Exp. Soc. Psychol., 39: 83-90. DOI: 10.1016/S0022-1031(02)00513-9

Pronin, E., C.M. Steele and L. Ross, 2004. Identity bifurcation in response to stereotype threat: Women and mathematics. J. Exp. Soc. Psychol., 40: 152-168. DOI: 10.1016/S0022-1031(03)00088-X

Schmader, T., M. Johns and M. Barguissau, 2004. The costs of accepting gender differences: The role of stereotype endorsement in women experience in the math domain. Sex Roles, 50: 835-850. DOI: 10.1023/B:SERS.0000029101.74557.a0

Shih, M., N. Ambady, J.A. Richeson, K. Fujita and H.M. Gray, 2002. Stereotype performance boosts: The impact of self-relevance and the manner of stereotype activation. J. Person. Soc. Psychol., 83: 638-647. DOI: 10.1037//0022-3514.83.3.638

Shih, M., T.L. Pittinsky and N. Ambady, 1999. Stereotype susceptibility: Identity salience and shifts in quantitative performance. Psychol. Sci., 10: 80-83. DOI: 10.1111/1467-9280.00111

Snyder, T.D. and S.A. Dillow, 2009. Digest of Education Statistics: 2010. 1st Edn., Government Printing Office, Washington, ISBN: 0160829739, pp: 718.

Spencer, S.J., C.M. Steele and D.M. Quinn, 1999. Stereotype threat and women math performance. J. Exp. Soc. Psychol., 35: 4-28. DOI: 10.1006/jesp.1998.1373

Stage, F.K. and S.A. Maple, 1996. Incompatible goals: Narratives of graduate women in the mathematics pipeline. Am. Educ. Res. J., 33: 23-51. DOI: 10.3102/00028312033001023

Steele, C.M. and J. Aronson, 1995. Stereotype threat and the intellectual test performance of African Americans. J. Person. Soc. Psychol., 69: 797-811. DOI: 10.1037//0022-3514.69.5.797

Steele, C.M., 1997. A threat in the air: How stereotypes shape intellectual identity and performance. Am. Psychol., 52: 613-629. DOI: 10.1037//0003-066X.52.6.613

Stevenson, H.W., S. Lee, C. Chen, L. Max and J. Stigler *et al*., 1990. Mathematics achievement of children in China and the United States. Child, Development, 61: 1053-1066. DOI: 10.2307/1130875

Stricker, L.J. and W.C. Ward, 2004. Stereotype threat, inquiring about test takers ethnicity and gender and standardized test performance. J. Applied Soc. Psychol., 34: 665-693. DOI: 10.1111/j.1559-1816.2004.tb02564.x

Stricker, L.J. and W.C. Ward, 2008. Stereotype threat in applied settings re-examined: A Reply. J. Applied Soc. Psychol., 38: 1656-1663. DOI: 10.1111/j.1559-1816.2008.00363.x

Thoman, D.B., P.H. White, N. Yamawaki and H. Koishi, 2008. Variations of gender-math stereotype content affect women vulnerability to stereotype threat. Sex Roles, 58: 702-712. DOI: 10.1007/s11199-008-9390-x

Tiedemann, J., 2000. Parents gender stereotypes and teachers beliefs as predictors of children concept of their mathematical ability in elementary school. J. Educ. Psychol., 92: 144-151. DOI: 10.1037//0022-0663.92.1.144

Tsui, M. and L. Rich, 2002. The only child and educational opportunities for girls in urban China. Gender Soc., 16: 74-92. DOI: 10.1177/089124320201600l005

Tsui, M., 2005. Family income, home environment, parenting and mathematics achievement of children in China and the United States. Educ. Urban Soc., 37: 336-355. DOI: 10.1177/0013124504274188

Tsui, M., 2007. Gender differences in mathematics and Sciences achievement in China and the United States. Gender Issues, 24: 1-11. DOI: 10.1007/s12147-007-9044-2

Venator, E.R., 2008. Stereotype threat and the academic performance of Chinese students. Soc. Chinese J. Soc., 28: 191-202.

Walker, M.E. and B. Bridgeman, 2008. Stereotype Threat Spillover and SAT Scores. 1st Edn., The College Board, New York, pp: 10.

Walton, G.M. and G.L. Cohen, 2003. Stereotype lift. J. Exp. Soc. Psychol., 39: 456-467. DOI: 10.1016/S0022-1031(03)00019-2

Walton, G.M. and S.J. Spencer, 2009. Latent ability: Grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students. Psychol. Sci., 20: 1132-1139. DOI: 10.1111/j.1467-9280.2009.02417.x

Wei, T., 2009. Stereotype threat, gender and math performance: Evidence from the National Assessment of Educational Progress. Harvard University.

Whyte, M.K. and W.L. Parish, 1985. Urban life in contemporary China. 1st Edn., The University of Chicago Press, Chicago, ISBN: 0226895491, pp: 415.

Willingham, W.W. and N.S. Cole, 1997. Gender and Fair Assessment. 1st Edn., Routledge, London, ISBN: 080582331X, pp: 411.

Witke, R., 1967. Mao Tse-tung, women and suicide in the may fourth era. China Quar., 31: 128-147. DOI: 10.1017/S0305741000028733