Modern Psychological Studies

Volume 18 | Number 2 Article 3

2013

Switching genders: identifying the evaluator in stereotype threat for men and women in a math context

Mirjana A. Antonic
University of Illinois at Chicago

Mary C. Murphy
University of Illinois at Chicago

Katherine T.U. Emerson
University of Illinois at Chicago

Lara D. Mercurio
University of Illinois at Chicago

Switching Genders: Identifying the Evaluator in Stereotype Threat for Men and Women in a Math Context

Mirjana A. Antonic, Mary C. Murphy, Katherine T. U. Emerson, Lara D. Mercurio
University of Illinois at Chicago

Abstract

The current study seeks to identify the source of evaluation that causes stereotype threat for men and women in a math context. In a 2 (participant gender: male vs. female) X 3 (gender label: Match, Mismatch, Control) factorial design, male and female participants that identified highly with math were asked to take a math test. Throughout the test, participants' ostensible gender was displayed on the computer screen. The displayed gender was either the correct gender, the opposite gender, or "Alabama." Although our results were unable to determine if stereotype threat is a self- or an outside evaluator-threat, we did observe a strong gender-math relationship in which being labeled with the opposite gender disrupted both men and women's math performance. However, women were more affected in that they not only performed significantly lower on the math test, but also took a longer time, attempted fewer problems, and significantly disidentified from math.

Keywords: stereotype threat, math performance, gender

Introduction

Stereotype threat has been used to explain the underperformance of stigmatized groups in various domains. Stereotype threat is defined as the fear or concern of confirming a negative stereotype about one's social group (Steele, 1997; Steele, Spencer & Aronson, 2002). This threat is linked to underperformance when a member of the stereotyped group who identifies with the domain in question is in a context where the negative stereotype is salient (Steele, 1997). Individuals are afraid that if any of their actions align with the negative stereotype, the stereotype is more likely to be seen as a self-characteristic (Steele & Aronson, 1995). Past research shows that the negative stereotypes that exist in our society about women's math ability create stereotype threat that causes women to underperform on difficult math tests compared to men (Spencer, Steele, & Quinn, 1999; Ben-Zeev, Fein, & Inzlicht, 2005). However, when this threat is removed, women perform just as well as their male peers.

Despite the general consensus that stereotype threat is the fear or concern of confirming a negative stereotype, it is much less clear to whom threatened individuals are afraid of confirming the stereotype. In the present study, we wanted to elucidate the definition of stereotype threat and examine whether the underperformance it causes for women in math is a self-threat or an outside evaluator-threat. To address this question, we created an experimental situation in which the self-threat was different from the outside evaluator-threat in a way that is relevant to the gender stereotype. Therefore, we created three conditions in which men and women took a math test. In the Match condition, participants completing a math test were labeled with their correct gender. In this condition, because both the participants and the experimenter were aware of the gender of the participant, decreased performance would lead to the confirmation of the stereotype to both the participant and to an outside evaluator. This condition mirrors the typical stereotype threat experiment manipulation. In contrast, in the Control condition, participants were labeled with a nonsense gender, "Alabama." This mislabeling allowed participants to be aware of their gender without this information being provided to the outside evaluator. Lastly, in the Mismatch condition, participants were labeled with the opposite gender. This final condition built upon the experience provided in the Control condition while also exploring how evaluation based on the opposite-gender stereotypes might affect the participant.

We were interested in how well men and women will perform on a math task under each of the gender label conditions. We had three hypotheses about math performance for this study. Firstly, we hypothesized that, in line with current stereotype threat research, men will outperform women in the Match condition because our highly math-identified participants were performing a math task where gender (and therefore gender stereotypes about math) was highly salient. Secondly, we had a competing hypothesis that men will outperform women in the Control condition if stereotype threat is a self-threat; however, we predicted that the difference in performance between men and women would be smaller if stereotype threat is an outside evaluator-threat. In other words, because the outside evaluator would be unaware of the participant's gender, if stereotype threat emerges as a result of fear of confirming a negative stereotype to an outside evaluator, women should be less susceptible to stereotype threat than if their gender is known to the evaluator. Our third and final hypothesis about math performance was also a competing hypothesis for each gender. For men in the Mismatch condition, we hypothesized they will show performance decrements if stereotype threat is an outside evaluator-threat because they would now be susceptible to the stereotype threat concerns typically experienced by women; if stereotype threat is a self-threat, we expected them to be unaffected. For women in the Mismatch condition, we hypothesized that if stereotype threat is a self-threat, they will have decreased performance regardless of their gender label. However, if stereotype threat is an outside evaluator-threat, women will be most protected by being judged in line with male stereotypes and therefore will not underperform.

In addition to test performance, we were also interested in seeing if men or women will challenge the identity mislabel more. We had three hypotheses about challenging the label. Firstly, we hypothesized that men labeled as women (Mismatch) would be motivated to try to inform the experimenter of the error (hereafter referred to as challenging the label) more often than they will challenge the label in the Control condition because there is a negative stereotype about women in math. Secondly, we hypothesized that men and women will challenge the label to a similar amount in the Control condition. Thirdly, we hypothesized that women labeled as men would challenge the Mismatch condition less often than men because a male identity does not carry the stigma of a negative stereotype in the math domain.

Lastly, we were interested in how men and women's overall experience, affect, and math identification would be affected by each gender label. Although we did not have any clear a priori hypotheses about these outcomes, research has demonstrated that overall experience, affect (Schmader, 2010; Rivardo, Rhodes, Camaione, & Jegg, 2011), and math identification (Aronson, Fried, & Good, 2002; Pronin, Steele, & Ross, 2003) are all impacted by stereotype threat. The extent to which these outcomes would be impacted by differences in the perceived evaluator was an empirical question.

Method

Participants

Ninety-five students were recruited for this study, but 25 students were excluded from analysis due to reporting a Math ACT score below the 66th percentile or because of technical issues at the time of participation. The final analyses were conducted on the

remaining 70 highly math-identified undergraduates, 28 males and 42 females, who participated in exchange for course credit and/or $7-10. The experimenters for the study were all female.

Research Design

The experiment consisted of a 2 (gender: male, female) X 3 (gender label: Match, Mismatch, Control) factorial design. Undergraduate students who were highly identified with math were recruited to participate in the study from the University of Illinois at Chicago. Math identification was determined via a pretesting measure consisting of two questions: "I am good at math tasks," and "It is important to me to do well on math tasks," (Murphy, Steele, Gross, 2007). Responses were given on a scale from 1 (strongly agree) to 8 (strongly disagree), and only students with a combined score of 3 or lower were eligible for the study.

The study was completed in an online survey administered in the lab. First, participants were asked to complete the Positive and Negative Affect Schedule (Watson, Clark, & Tellegan, 1988) as a baseline measure of their affect. Next, a tutorial introduced modular arithmetic to the participant by explaining that the participant needed to judge if each modular arithmetic statement is True or False as quickly and as accurately as possible, and providing one method that can be used to determine if the statement is true or false (Beilock, Rydell, & McConnel, 2007). For example, when presented with the statement: $17 \equiv 5 \pmod 6$, the statement is true if the mod number divides into the difference of 17 and 5 with 0 as the remainder, and false otherwise. In this example, the answer can be derived by subtracting 5 from 17 and dividing by 6. Since 6 divides into 12 with 0 as the remainder, this particular item is True. Proceeding the tutorial, the participant was given 6 unscored practice problems with comprehensive feedback after each item to ensure the participant understood how to solve the modular arithmetic problems prior to beginning the test. For instance, if the participant answered $5 \equiv 2 \pmod 2$ with True, the feedback would state: "**Incorrect!!** $5 \equiv 2 \pmod 2$ is actually **False** because $5 - 2 = 3$ but 2 does not go into 3 with a remainder of 0." After the tutorial and practice problems, the participant completed a 79-item modular arithmetic test, which interspersed difficult and easy items, with feedback indicating whether or not the participant had chosen the right answer and also reinforcing the correct answer. Performance on the math test was measured using the total time spent on the test, the number of items the participant provided an answer for (number attempted), and the number of items they answered correctly. Upon completing the math test, participants completed additional survey items about their experience including the Positive and Negative Affect Schedule and math identification scale used previously. The study was video recorded and for participants who gave permission for the video recording of their study to be used, the video recording was coded on a 2 point scale, with a 1 if the participant got up from their seat to try to find the experimenter in response to their gender label, and 2 if the participant remained in their seat until the task was over.

Procedure

Participants were recruited to participate in individual sessions for a psychology study. Upon arrival at the lab, each participant was informed that they would be working on a computer to complete a study about problem solving that includes:

providing important demographic information that will be linked to their results, completing a math test that involves a new type of math currently being developed, and filling out additional survey items about their experience with the test. The experimenter informed the participant that the experimenter would not be present during the study because she had to attend to another study in a different part of the building. The participant was instructed to complete the task on their own and wait for the experimenter to return once they were finished.

The participant was then left on their own in the lab to complete the task via an online survey. After filling out demographic information, the survey ostensibly displayed the participant's demographic information on a confirmation screen. However, the gender that was displayed on the confirmation screen reflected the condition to which they were assigned (Match, Mismatch, or Control), so for two-thirds of the participants the gender that was displayed was incorrect, either the opposite gender (Mismatch) or "Alabama" (Control). The computer did not allow the participant to return to the previous page to alter any information, so the participant was forced to continue to the tutorial and math test regardless of the content of the demographic confirmation screen. Throughout the math test portion of the study, the participants' ID number and assigned gender label were displayed at the top of the screen (see Fig. 1). The participants were unobtrusively video recorded throughout the experiment to determine whether or not each participant tried to challenge the label by getting up to seek the experimenter at any point. Upon completing the study, participants were compensated and thoroughly debriefed.

Results

Foremost, we were interested in how men and women's test performance would be affected under each gender label condition, as a means of determining if stereotype threat is caused by a self- or an outside evaluator-threat. Test performance was determined by the number of problems completed correctly, the time spent completing the problems, and the number of problems attempted. To examine the test performance, we first conducted a 2 (gender) X 3 (gender label) analysis of covariance (ANCOVA) on the percent of items answered correctly while controlling for Math ACT score. We observed a marginal main effect for gender label condition, $F(2,63) = 2.936$, $p = .06$. Participants in the Mismatch condition ($M = .891$, $SD = .110$) performed significantly worse than those in the Match ($M = .947$, $SD = .081$) and the Control ($M = .947$, $SD = .045$) conditions, with no other significant differences between conditions (See Fig. 2). There was no main effect of gender, $F(1, 63) = 1.741$, ns, and there was no significant interaction, $F(2,63) = .901$, ns. However, because we specifically hypothesized that men would outperform women in the Match condition, we conducted the planned follow-up test between men and women's percent correct in the Match condition. We did not observe a significant stereotype threat between genders in the Match condition as predicted $F(1,63) = 1.310$, ns. This suggests that the Match condition was not experienced as threatening to women, contrary to our hypothesis. Since the Match condition was expected to be the most threatening condition for women, without this threatening condition we cannot make any claims about the source of stereotype threat as a self- or an outside evaluator-threat. However, we did observe an unexpected result, that both men and women

answered significantly fewer questions correctly in the Mismatch condition. Our remaining analyses investigate this difference between the Mismatch condition and the other conditions further.

In addition to providing the correct answers, participants were also instructed to complete each item as quickly as possible. Therefore, the participants knew that the time they spent taking the test was also indicative of their test performance. The differences in the amount of time participants spent on the test may indicate that some conditions were more challenging than others. We conducted a 2 X 3 ANCOVA on the time spent to complete the test while controlling for the percent of items attempted and Math ACT score. Results revealed no significant main effect of gender label condition, $F(2,62) = 1.228$, ns, and no main effect of gender, $F(1,62) = .622$, ns. There was also no significant interaction, $F(2,62) = 1.450$, ns. However, because we were specifically interested in how men and women would react to the different conditions, we conducted planned follow-up tests of the simple effects of gender. There was a marginally significant contrast between conditions for women, $F(2, 62) = 2.535$, $p = .087$ such that women in the Mismatch condition ($M = 670.01$, $SD = 332.73$) took a significantly longer time than women in the Match condition ($M = 479.83$, $SD = 145.83$), $p < .05$, and slightly longer than women in the Control condition ($M = 583.07$, $SD = 155.46$), $p = .130$. There was no effect of condition on time for men, $F(2, 62) = .694$, ns. In terms of the time spent completing the test, men were unaffected, while women's performance is hindered when mislabeled with the male gender (see Fig. 3).

Lastly, we wanted to examine how motivated participants were to complete all of the test items by looking at the percent of the items attempted on the math test. A 2 X 3 analysis of variance (ANOVA) revealed a very marginal main effect of gender label condition, $F(2, 63) = 1.913$, $p = .16$, such that participants in the Mismatch condition ($M = 77.882$, $SD = 1.867$) attempted fewer questions than those in the Match ($M = 78.704$, $SD = .823$), $p = .057$, and Control ($M = 78.539$, $SD = .859$), $p = .143$, conditions. There was no significant main effect of gender $F(1, 63) = 1.455$, ns, and no significant interaction, $F(2, 63) = .265$, ns. However, because we expected women and men to respond differently to the different conditions, we explored the condition contrasts for each gender separately. This follow-up revealed that the difference between conditions was being driven by women, $F(2, 64) = 2.547$, $p = .086$, who attempted fewer problems in the Mismatch condition ($M = 77.636$, $SD = 2.292$) than in the Match condition ($M = 78.647$, $SD = .996$), $p < .05$, and the Control condition ($M = 78.429$, $SD = .852$), $p = .10$. There was no difference between conditions for the men, $F(2, 64) = .296$, ns. Only women in the Mismatch condition are attempting fewer problems. Furthermore, participants in this condition are only attempting approximately one problem fewer, 77.5 out of 79 problems instead of 78.5 out of 79 problems, suggesting that they may not be skipping problems that are too hard or because they are not motivated, but that they may instead be accidentally skipping a problem by clicking "Next" multiple times while distracted or rushing.

In addition to test performance, we were interested to see if and how men and women's math identification will change in the different conditions after completing the task. Controlling for pre-test math identification, we conducted a 2 X 3 ANCOVA on the difference score in math

identification. Results showed a marginal main effect for gender label condition, $F(2, 62) = 2.368$, $p = .102$, such that participants in the Mismatch condition ($M = -.781$, $SD = .856$) showed a greater decrease in math identification than participants in the Match ($M = -.222$, $SD = .670$) and Control ($M = -.400$, $SD = .520$) conditions, with no other differences between conditions. There was no main effect for gender, $F(1, 62) = .069$, ns, and no interaction, $F(2, 62) = 1.207$, ns. However, to follow-up on the hypothesis that women and men experienced the conditions differently, we examined the contrasts between conditions for each gender separately. Men had no significant differences in math identification across conditions, $F(2,28) = .114$, ns. In contrast, women's change in math identification was significantly different between conditions, $F(2, 28) = .014$, $p < .05$, such that the drop in math identification was larger in the Mismatch condition ($M = -.950$, $SD = 1.012$) relative to the Match ($M = -.147$, $SD = .343$) and Control ($M = -.346$, $SD = .555$) conditions, $ps < .05$. Being mislabeled as male, but not "Alabama," while taking a math test, caused women to disidentify from math. This effect did not happen for mislabeled men, whose change in math identification was equally small across all conditions (see Fig. 4).

Next, we were interested in the participant's overall affect and experience during the study to shed light on any discomfort the mislabel caused or coping strategies employed. First we examined change in positive affect. Results show marginal main effects for both gender label condition $F(2,63) = 2.715$, $p = .074$, such that participants in the Mismatch condition ($M = -.558$, $SD = .895$) showed a decrease in positive affect relative to the Match ($M = .053$, $SD = 1.249$) and Control ($M = .000$, $SD = .937$) conditions, and gender $F(1,63) = 3.098$, $p = .083$, such that women ($M = -.250$,$SD = 1.139$) showed a decrease in positive affect relative to men ($M = .101$, $SD = .951$), but there was no significant interaction, $F(2,63) = .311$, ns. Although women's positive affect stayed the same across conditions, $F(2,63) = 1.057$, ns, there is a trending simple effect for men, $F(2,63) = 1.750$, $p = .068$, to show a decrease in positive affect in the Mismatch condition ($M = -.556$, $SD = .735$) relative to the Match ($M = .417$, $SD = 1.187$) and Control ($M = .167$, $SD = .670$) conditions. Upon completing the math test, participants were also asked to rate their overall experience. There is a marginal main effect of gender label condition, $F(2,60) = 3.141$, $p = .050$, such that participants in the Mismatch condition ($M = 6.94$, $SD = 1.391$) reported a less positive experience than participants in the Match ($M = 7.42$, $SD = 1.391$) and Control ($M = 7.44$, $SD = .961$) conditions, and no main effect of gender, $F(1,60) = .008$, ns. There is a significant interaction, $F(2, 60) = 3.938$, $p < .05$. Follow-up tests indicate that men in the Match condition report a significantly more positive experience ($M = 8.00$, $SD = .000$) compared to women in the same condition ($M = 7.06$, $SD = 1.138$), $p < .05$, and compared to men in the Mismatch ($M = 6.67$, $SD = 1.966$), $p < .01$, and Control ($M = 7.25$, $SD = 1.138$), $p < .05$, conditions. In addition to affect and experience, participants were also asked to report how focused they were on the task. Results show a marginal main effect of gender label condition, $F(2, 63) = 1.893$, $p = .159$, such that participants in the Mismatch condition ($M = 6.18$, $SD = 1.590$) reported less focus than participants in the Match ($M = 7.10$, $SD = 1.595$) and Control ($M = 7.33$, $SD = .888$) conditions, and no main effect of gender, $F(1, 63) = .516$, ns, and no significant interaction, $F(2,63) = 1.252$, ns. Follow-up tests show that women reported no differences in focus across conditions,

$F(2,63) = .271$, ns. In contrast, men did exhibit a difference in focus across conditions, $F(2,63) = 2.359$, $p = .103$, reporting less focus in the Mismatch condition ($M = 5.83$, $SD = 1.941$) relative to the Match ($M = 7.10$, $SD = 1.595$), $p = .08$, and the Control ($M = 7.33$, $SD = .888$), $p < .05$, conditions. Overall, men seem to be reporting focus and affect more consistent with their performance in each condition. Women, however, are reporting consistent affect and focus across all conditions despite a poorer performance in the Mismatch condition. This could indicate a coping strategy similar to that used when under stereotype threat.

Finally, we were interested if men or women will challenge the identity mislabel more. As can be expected because participants in the Match condition were labeled with their correct gender, there is a strong main effect of gender label condition, $F(2,44) = 6.975$, $p < .01$, such that participants in the Mismatch condition ($M = 1.64$, $SD = .497$) challenged the label more often than those in the Match ($M = 2.00$, $SD = .000$) and Control ($M = 1.94$, $SD = .250$) conditions. There was no main effect of gender, $F(1, 44) = .289$, $p =$ ns, and no significant interaction, $F(2, 44) = 1.405$, ns. Follow-up tests support our hypothesis that men in the Mismatch condition would challenge the mislabel more than men in the Control condition. Results show that men in the Mismatch condition ($M = 1.50$, $SD = .548$) did challenge the mislabel more than men in the Control condition ($M = 2.00$, $SD = .000$), $p < .01$. Our hypothesis that women would challenge the mislabel less in the Mismatch condition than men was not supported. Women in the Mismatch condition ($M = 1.75$, $SD = .463$) did not challenge the mislabel less often than men in the same condition ($M = 1.50$, $SD = .548$), ns. When labeled with the mismatched gender, both men and women were equally likely to challenge the identity mislabel.

Discussion

Because the difference between men and women's performance in the Match condition was not significant, the nature of stereotype threat as a self- or an outside evaluator-threat could not be determined. Results reveal a slight trend of stereotype threat in the Match condition, with women performing on average 4% lower than men, but it is not statistically significant. Our pattern of results does suggest an important gender-math relationship. Both genders respond negatively to being mislabeled with the opposite gender, but it is more problematic for women in a math context. Contrary to our hypothesis, women were not protected from stereotype threat in the Mismatch condition, rather they seem to be more threatened in this condition. When mislabeled with the opposite gender, both women and men completed significantly fewer problems correctly on the math test, but only women also tended to attempt one fewer problem than the other conditions and took a significantly longer time to finish. In fact, women in the Mismatch condition took 18% longer than men in the same condition, and 28% longer than women in the Match condition. These results have strong implications because effects like this would be likely to increase gender disparity in timed testing conditions, such as standardized tests.

Furthermore, only women that were labeled as men while taking the math test tended to significantly disidentify from math. It is not surprising to see a small decrease in math identification in all conditions due to the repetitive nature of the task and some regression to the mean of the extreme starting values on this measure.

However, women's math identification in the Mismatch condition significantly decreased relative to all other conditions, dropping by almost a full point on an 8 point scale. Disidentifying from a domain when under threat serves as an ego-protective strategy (Pronin, et al., 2003). Research also shows that because math is seen as masculine, there is an implicit association between math and male that makes it difficult for women to identify with math (Nosek et al., 2002). Mislabeling women as male may be causing women to react against the male label and also against math through the association, supporting the hypothesis that the math-male association is important in some stereotype threat effects.

Women report equally positive affect and focus on the task across all three conditions, while men report affect and focus consistent with their poorer performance in the Mismatch condition. This indicates that women, but not men, are either unaware of or suppressing the effects of being mismatched, which may reflect a coping strategy similar to that used under stereotype threat. Distraction theories state that when stereotype threat is present, individuals experience reduced working memory capacity and reduced ability to focus on task (Schmader, 2010; Engle, 2002; Beilock, Holt, Kulp, & Carr, 2004). However, one of the coping strategies observed by participants under stereotype threat is to suppress feelings of anxiety (Schmader, 2010), which may explain why women fail to report affect more consistent with their performance. Future research could test this hypothesis.

In sum, this research suggests that mislabeling women as male results in gender differences that are similar to those seen when women are under stereotype threat. In contrast, men who have been mislabeled as women show performance decrements, but are spared from many of the other negative outcomes exhibited by the women in this study.

Limitations and Future Research

As mentioned, it is impossible to conclude whether stereotype threat is a self- or an outside evaluator-threat from this study because the difference in performance in the Match condition was not significant between genders. The set up of this study did not create a threatening condition for women in the Match condition, so there is no reason to believe that the other conditions were threatening because of stereotype threat. The test population may be a possible limitation. Perhaps with a larger test population, the stereotype threat could be significant.

Additionally, stereotype threat may not have been salient enough to the novel task presented. Repeating the current study while making stereotype threat more salient to this novel task by describing it as a test that mirrors the results of IQ or ACT/GRE math tasks might also increase stereotype threat. This could tease out whether it is a self- or an outside evaluator-threat. In addition, increasing the number of participants would give us more statistical power.

Women and men both underperformed on the math test when mismatched with the wrong gender, but their timing and motivation suggest different mechanisms that cause this underperformance. Future studies can be created to explore these different mechanisms. For example, using the current model and asking more in-depth questions about the participant's experience may provide further insight on being mismatched for men and women, informing

research into these gender-specific mechanisms. It is important to identify the source of evaluative threat in order to design stereotype threat interventions that are targeted to address the correct threat source.

References

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

Beilock, Sian L., Holt, Lauren E., Kulp, Catherine A., & Carr, Thomas H. (2004). More on the fragility of performance: Choking under pressure in mathematical problem solving. Journal of Experimental Psychology, 133, 584-600.

Beilock, Sian L., Rydell, Robert J., & McConnel, Allen R. (2007). Stereotype threat and working memory: Mechanisms, alleviation, and spillover. Journal of Experimental Psychology, 136, 256-276.

Ben-Zeev, Talia, Fein, Steven, & Inzlicht, Michael. (2005). Arousal and stereotype threat. Journal of Experimental Social Psychology, 41, 174-181.

Engel, Randall W. (2002). Working memory capacity as executive attention. Current Directions in Psychological Science, 11, 19-23.

Johns, Michael, Schmader, Toni, & Martens, Andy. (2005). Knowing is half the battle: Teaching stereotype threat as a means of improving women's math performance. Psychological Science, 16, 175-179.

Major, Brenda, Spencer, Steven, Schmader, Toni, Wolfe, Connie, & Crocker, Jennifer. (1998). Coping with negative stereotypes about intellectual performance: The role of psychological disengagement. Personality and Social Psychology Bulletin, 24, 34-50.

Murphy, Mary C., Steele, Claude M., & Gross, James J. (2007). Signaling threat: How situational cues affect women in math, science, and engineering settings. Psychological Science, 18, 879-885.

Nosek, Brian A., Banaji, Mahzarin R., & Greenwald, Anthony G. (2002). Math = male, me = female, therefore math ≠ me. Journal of Personality and Social Psychology, 83, 44-59.

Pronin, Emily, Steele, Claude M., & Ross, Lee. (2003). Identity bi-furication in response to stereotype threat: Women and mathematics. Journal of Experimental Social Psychology, 40, 152-168.

Rivardo, Mark G., Rhodes, Michael E., Camione, Tyler C., & Legg, Jessica M. (2011). Stereotype threat leads to reduction in mumber of math problems women attempt. North American Journal of Psychology, 13, 5-16.

Schmader, Toni. (2010). Stereotype threat deconstructed. Current Directions in Psychological Science, 19, 14-18.

Spencer, Steven J., Steele, Claude M., & Quinn, Diane M. Stereotype threat and women's math performance. Journal of Experimental Social Psychology, 35, 4-28.

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

Steele, Claude M., & Aronson, Joshua. Stereotype threat and the intellectual test performance of African Americans. Journal of Personality and Social Psychology, 69, 797-811.

Steele, Claude M., Spencer, Steven J., Aronson, Joshua. (2002). Contending with group image: The psych-ology of stereotype and social id-entity threat. In M. Zanna (Ed.), Advances in exper-imental social psychology.

Watson, David, Clark, Lee Anna, & Tellegen, Auke. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. Journal of Personality and Social Psychology, 54, 1063-1070

Figures

Figure 1. Screenshot of the modular arithmetic test with participant ID number and gender label condition displayed.

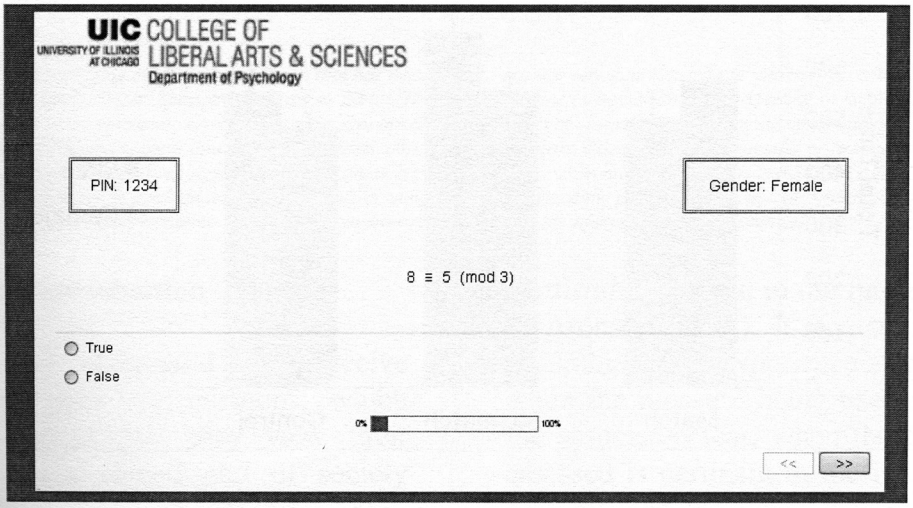

Figure 2. Percent of items answered correctly on math test as a function of gender and gender label condition controlling for Math ACT score.

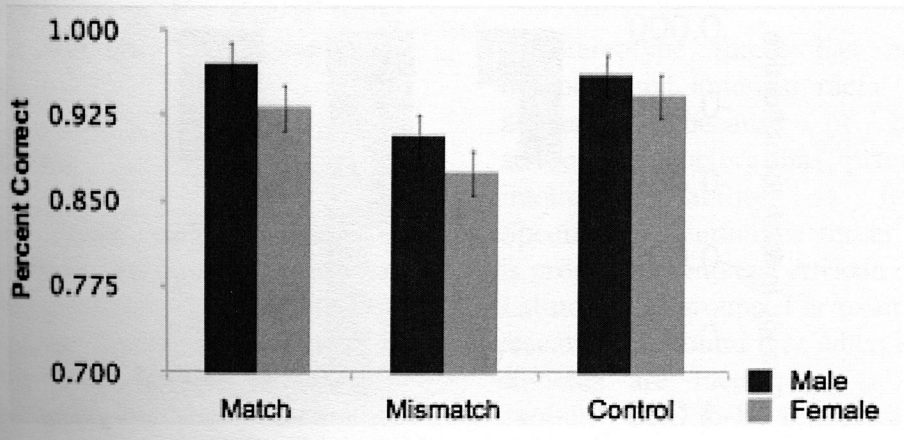

Figure 3. Total time spent taking the math test as a function of gender and gender label condition controlling for percent of items attempted and Math ACT score.

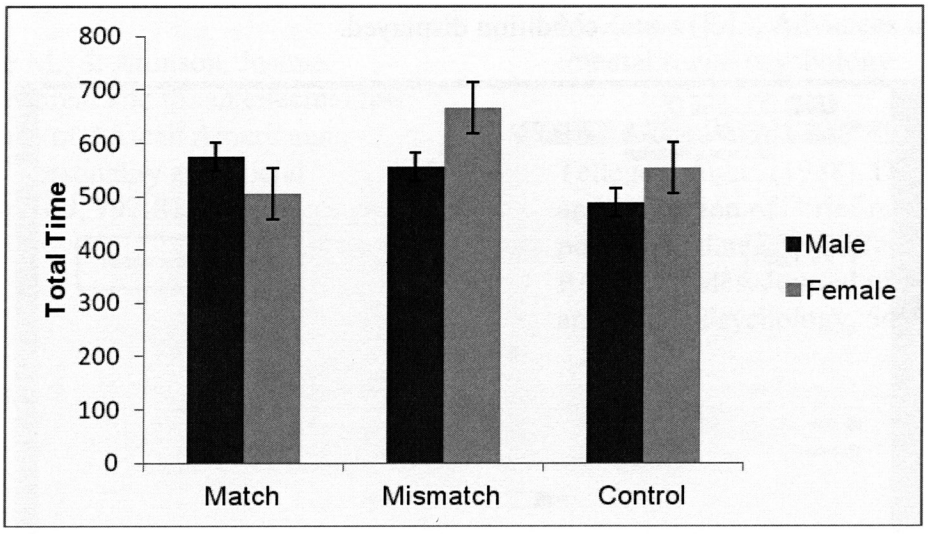

Figure 4. Change in math identification on an 8 point scale as a function of gender and gender label condition.