Effects of Gender Stereotype Threat on Physiological Stress Response

Joey Janz, Caitlin Kuckes, Linda Mei, Hasan Nadeem, Emerald Reeg, Rachel Smitz

University of Wisconsin-Madison, Department of Physiology

Physiology 435

Lab 602, Group 5

Word Count: 3,897

Keywords: ElectroDermal Activity (EDA), Exam, Galvanic Skin Response (GSR), Gender, Mathematics, Physiological Arousal, Stereotype Threat, Stress, Working Memory

Abstract

Stereotype threat is observed in many settings, but one of particular interest is gender. A classic gender stereotype is that females perform worse on mathematical assessments as compared to males, which can have significant implications on women today looking to pursue careers in math or science. Previous research has found that performance on standardized tests was hindered by the presence of significant negative stereotype, which affects monitoring processes, such as working memory, and induces several physiological responses. This study explored the relationship between induced gender stereotype threat during a mathematical examination and the physiological stress response that resulted. This was accomplished through measurements of heart rate, respiration rate and depth, and electrodermal activity (EDA) during a two portion mathematical exam, halfway through which an article was administered to induce gender stereotype threat. The statistical analyses did not show a significant increase in physiological stress response linked to math exam performance after induced stereotype threat. Likely reasons for this include small sample size, time constraint for data collection, and experimental equipment inconsistencies. Thus our hypothesis that increased levels of physiological stress response would result from induced stereotype threat in a female cohort was not supported.

Introduction

A phenomenon widely studied in psychology and sociology is stereotype threat: "the situation in which there is a negative stereotype about a person's group, and he or she is concerned about being judged or treated negatively on the basis of this stereotype." (Spencer *et al.,* 2016). While stereotype threat can be observed in many settings, the scope of this paper focuses on stereotypes against gender; specifically, the stereotype that women perform worse on mathematical assessments as compared to men. In a seminal study by Spencer, it was found that

women risk judgement based on this stereotype - a risk men do not face (Spencer *et al.,* 1999). For instance, in a study by Ryan *et al.,* 2008, it was found that when group members take standardized tests, their performances were partially hindered by the presence of significant, negative stereotypes. This can have serious implications for women attempting to pursue careers in fields with high standards of proficiency in mathematics. Furthermore, solutions to counteract stereotype threat prove to be elusive, as the mechanisms through which stereotype threat affects individual performance are not clearly defined.

Spencer theorized three mechanisms through which stereotype threat may take place: underperformance due to extra pressure to succeed, threats to self-integrity and belonging, and priming the stereotype. In the first mechanism (extra pressure to succeed), Spencer noted several studies that posited increased sympathetic nervous system arousal and the misattribution of this arousal that undermined the performance of individuals (Spencer *et al.,* 1999). It is important to note, however, that physiological arousal mainly affected the performance of individuals on complex tasks, as opposed to easy tasks. This is because the prepotent response of individuals is often correct for easy tasks, but often incorrect for difficult tasks (Ben-Zeev *et al.,* 2005). This corresponds with the Yerkes-Dodson law of physiological arousal, which states that task-performance rises with increased physiological arousal up to a certain point, implying that an optimal level of arousal is associated with an optimal level of performance (Broadhurst, 1959).

In a recent analysis by Schmader, a process model was proposed that attempted to explain stereotype threat in the context of stress response, vigilance, working memory, and self - regulation. Process models are used extensively by psychologists to describe mental processes and they generally share units of information input, processing, storage, and output. Stereotype threat effects monitoring processes as well as induces a physiological response. The stress

response is shown to both directly and indirectly affect thoughts/emotions, suppression processes, and working memory. The impairment of working memory is what leads to poorer performance on cognitive tasks (Schmader *et al.,* 2008). Given the results of Spencer *et al.* (1999), Schmader *et al.* (2008), and the Yerkes-Dodson law of physiological arousal, we expect females to experience heightened levels of physiological stress after the induction of a stereotype threat before a difficult math test. This stress response would occur due to greater motivation to disprove the stereotype against them and the difficulty of the given exam.

Considering that the literature supports links between physiological arousal, stress, and stereotype threat, we believe that sympathetic arousal as a stress response is a significant determinant of decreased performance due to stereotype threat. To test this, the physiological stress responses were monitored as stereotype threat was induced to participants midway through a two-portion math exam. Stereotype threat was induced in between portions of the exam using a short article that conveys the message of female inferiority in the mathematical sciences. All participants had their heart rate, respiration, and electrical skin conductivity monitored continuously throughout the experiment. Based on our pilot study, it was determined that data was to be evaluated using an average for each experimental group, derived from averages of the data for individual participants in a specified time interval. As a negative control, some female participants were not exposed to the gendered media, but gender-neutral media, along with the math exam. Under these controlled conditions, no change in the physiological variables between the first and second half of the test was expected, due to the absence of overt stereotype threat.

Our study intends to provide evidence that links a physiologically manifested stress response and an induced stereotype threat. Furthermore, if stress induced by gendered stereotype threat is associated with lowered exam scores, these findings will have a much broader societal

implication. If the findings hold true, we as a society are now accountable to take measures in reducing the presence of such threats.

To quantitatively measure the presence of a stress response, physiological data was measured in the form of heart rate, respiratory rate, and electrodermal activity. Elevated levels of these metrics would be indicative of a stress response linked with stereotype threat induced by reading a gendered article. This experiment is novel because findings in the literature are yet to associate physiological data regarding stress response and stereotype threat. Does the induction of gender stereotype threat lead to physiological arousal in women that is absent in men? It was hypothesized that a significant increase in measures of physiological stress would be found as a result of induced stereotype threat in a female cohort, in accordance with Spencer *et al*. (1999), Schmader, *et al.* (2008), and the Yerkes-Dosdon law of physiological arousal.

Methods and Materials

Participants

Male and female participants from the University of Wisconsin - Madison (age = 20 – 22) were recruited from an undergraduate physiology course. All subjects provided written informed consent prior to the commencement of the study.

Materials

The experiment utilized two mathematical exams, one of two educational articles, and *BIOPAC transducers and software*. Exam questions were culled by the study team and selected from the Kaplan GRE Math Workbook, 9th edition (Kaplan Publishing, 2013) and Kaplan GRE: Premier 2015 with 6 Practice Tests (Kaplan Publishing, 2015). Questions were pulled equally from sections of algebra, arithmetic, geometry, and quantitative reasoning.

To record physiological data, researchers used a respiratory transducer, pulse oximeter, and EDA transducer. The equipment was connected to the BIOPAC® Analog to Digital Converter (MP36, BIOPAC Systems, Inc., Goleta, CA, USA) and data was recorded throughout the experiment. Setup and placement of devices followed the Biopac Student Lab Manual 4.0. A Respiratory Transducer (SS5LB, BIOPAC Systems, Inc., Goleta, CA, USA) was fastened just below the participant's left and right armpits. To measure electrodermal activity, EDA Transducers (SS3LA, BIOPAC Systems, Inc., Aero Camino Goleta, CA, USA) with Isotonic Recording Electrode Gel (GEL101) were wrapped around the index and middle finger of the participant's left hand. A pulse oximeter (9843, Nonin Medical, Inc., Plymouth, MN, USA), placed on the participants left ring finger, was used to continuously measure the pulse of the participant.

Experimental Design

Upon completion of the consent form, participants entered a testing room with two exams, an article, a TI-83 calculator, pencil, and two data recorders present. Subjects were randomly assigned to an experimental or control group using a random number generator. Odd numbered participants experienced experimental treatment, whereas even numbered participants served as controls.

After completing exam #1, experimental groups were given a variation of an article from The Telegraph, a British daily newspaper, titled **"***OECD education report: boys 'pulling ahead of girls' in math tests*." (Paton and Graeme, 2013) This article was used to induce gendered stereotype threat against female participants. Controls read a gender-neutral and educational article from National Geographic titled "*Dogs Are Even More Like Us Than We Thought*." (Wei-Haas and Maya, 2015). Articles were controlled for similarity in comprehension difficulty and

length. The articles provided can be seen in Appendix 1. Participants then began taking exam #2 and physiological data were recorded.

Baseline EDA, respiration, and heart rate measurements were obtained by starting data collection one minute prior to starting the first exam. Participants were provided with as much time as necessary to conclude both exams. After completion of the first exam, participants were instructed to read the article assigned to them. The second exam was administered subsequently after reading the article presented. A timeline of experimental design is shown in *Figure 1*.

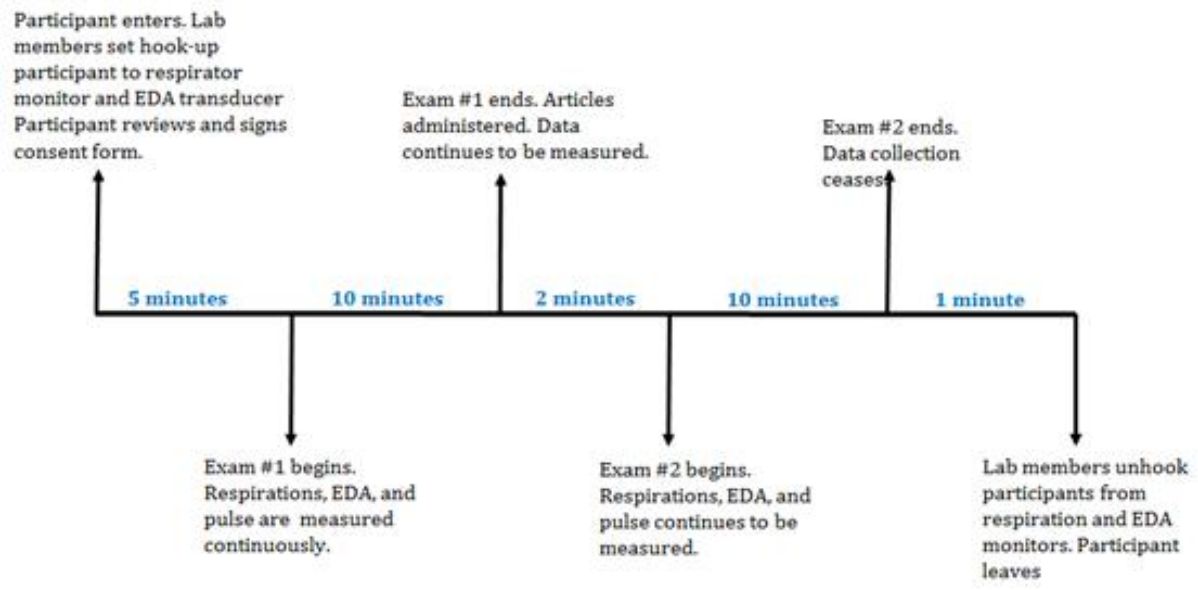

Figure 1. Timeline of a participant completing the study

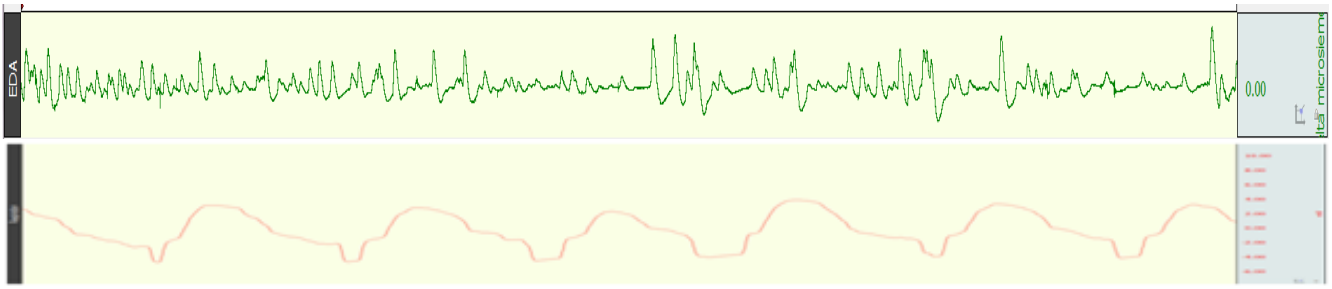

Figure 2. Example data of respirations and galvanic skin conductance (GSR) over time. The green line is representative of galvanic skin conductance changes in μS. The red line is representative of respirations in mV; each wave represents one respiration.

Data Analysis

All measurements were collected using the BIOPAC Systems, Inc. MP36. Data for each physiological measurement was obtained from five time-points from each of the two test trials at the 0, 2.5, 5, 7.5, and 10 minute marks. Heart rate (in beats per minute), respiration (in mV) and electrodermal activity (in microsiemens) data from the first exam were averaged from the five aforementioned time-points. The percent of problems correct on the exams were also recorded in order to account for further identification of stereotype threat induced stress while taking the mathematical exam.

These variables were statistically analyzed through the use of t-tests. T-test analyses were performed both within and between testing conditions. Within conditions, t-tests were performed between the first and second exam for both men and women for all variables. An indication for whether stereotype threat was induced, and subsequent physiological response, would be apparent if there were significant differences in physiological data and percent of correct answers between the two administered exams. It is expected that stereotype threat would only be induced in the female experimental group. Furthermore, ANOVA tests were performed between testing conditions between the first and second exam for both male and females. It is expected that a significant difference between physiological data and percent of correct answers would be found only in the female experimental vs. the female control.

Results

Our study intended to correlate a physiological stress response targeting gender with decreased math performance for female participants. The results do not support this hypothesis.

Respiration

Respiration rates were averaged over 150 second intervals during each exam for each individual. Participants regularly took three to four of these intervals to complete each exam. Mean respiratory rates for each participant were then averaged together by experimental group (Figure 3). Completion of exam 1 provides a baseline to control for stress reactions to math and individual differences in resting respiration rate. Average respiration rate for exam 1 and exam 2 respectively was 18.07 respirations per minute and 17.65 respirations per minute for females in the experimental group; 19.82 and 14.59 respirations per minute for females in the control group, 18.33 and 16.21 respirations per minute for males in the experimental group; and 17.03 and 16.21 respirations per minute for males in the control group. Respiration rate showed no significant change between exam 1 and exam 2 for any group.

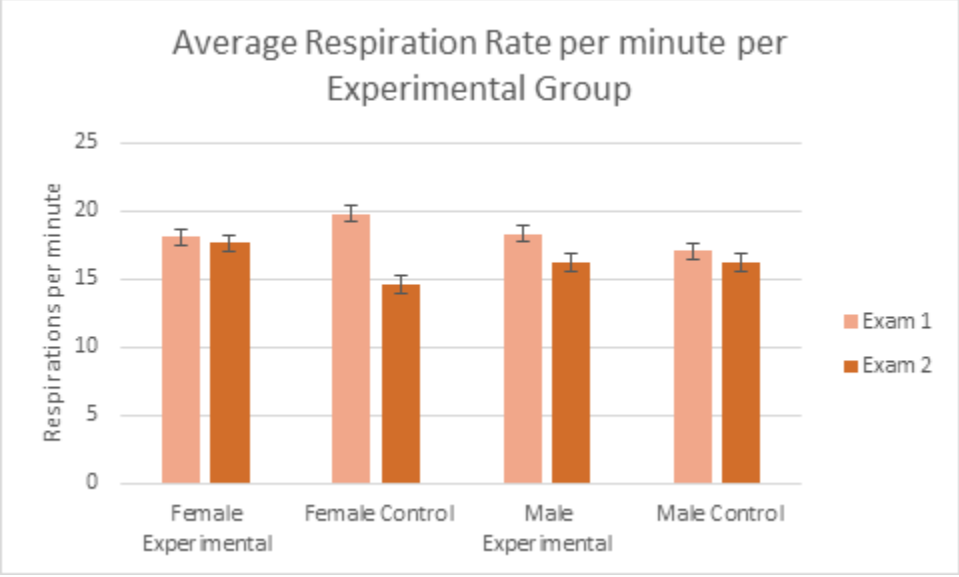

Figure 3. Comparison of average respiration frequency among experimental groups. For female experimental n=5, female control n=4, male experimental n=5, male control n=4.

Heart Rate

Heart rate data was collected and averaged for the same intervals as respirations and aggregated into experimental groups (Figure 4). Exam 1 provides a baseline heart rate to control for reactions to math and individual variation in resting heart rate. Average heart rate for exam 1 and exam 2 respectively was 81.50 bpm and 81.45 bpm for females in the experimental group, 75.76 bpm and 76.29 bpm for females in the control group, 75.38 bpm and 76.58 bpm for males in the experimental group, and 80.40 bpm and 82.67 bpm for males in the control group. Heart rate showed no significant change between exam 1 and exam 2 for any group.

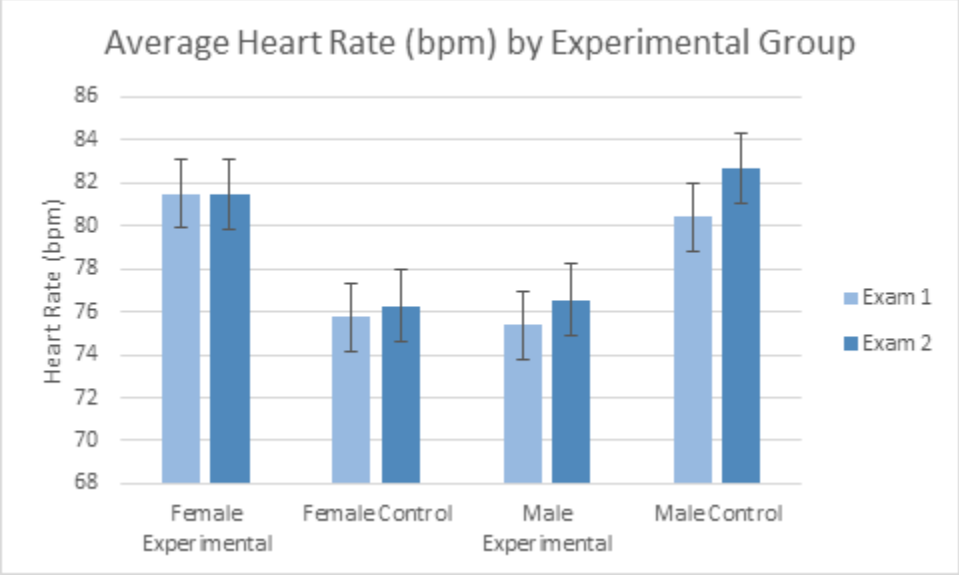

Figure 4. Comparison of average heart rate among experimental groups. For female experimental n=4, female control n=4, male experimental n=4, male control n=5.

Galvanic Skin Response

Galvanic skin response data was collected and averaged for the same intervals as respiration and heart rate. These results were then averaged to one value for each exam for each participant, and then aggregated into experimental groups. The female control group had the

greatest increase in GSR between exam 1 and exam 2 (Figure 5). Average galvanic skin response for exam 1 and exam 2 respectively was 0.0067 delta microsiemen and 0.0122 delta microsiemen for females in the experimental group, -0.0137 delta microsiemen and 0.0521 delta microsiemen for females in the control group, 0.0250 delta microsiemen and 0.0153 delta microsiemen for males in the experimental group, and 0.0131 delta microsiemen and 0.0128 delta microsiemen for males in the control group. The variation between the exams within each group did not result in a significant difference of the variation within the female experimental group as compared to the other groups.

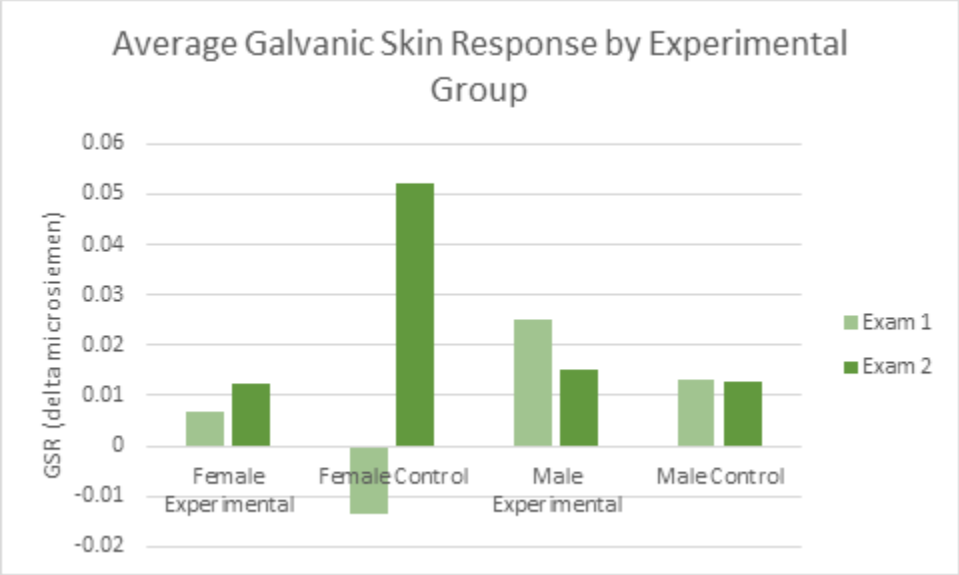

Figure 5. Comparison of average Galvanic Skin Response for exam 1 and exam 2 among experimental groups. For female experimental n=5, female control n=5, male experimental n=5, male control n=4.

Exam Scores

To assess if the physiological responses we measured correlated with a change in exam performance, we averaged the scores for each group (Figure 6). Average number of correct answers for exam 1 and exam 2 respectively were 6.0 and 5.6 for females in the experimental

group, 6.8 and 6.0 for females in the control group, 5.8 and 5.0 for males in the experimental group, and 6.4 and 5.6 for males in the control group. Average scores were lower for exam 2 than for exam 1 for all groups, and there was no statistically significant difference between the groups for change in score. This trend may be accounted for by a difference in difficulty between the two exams.

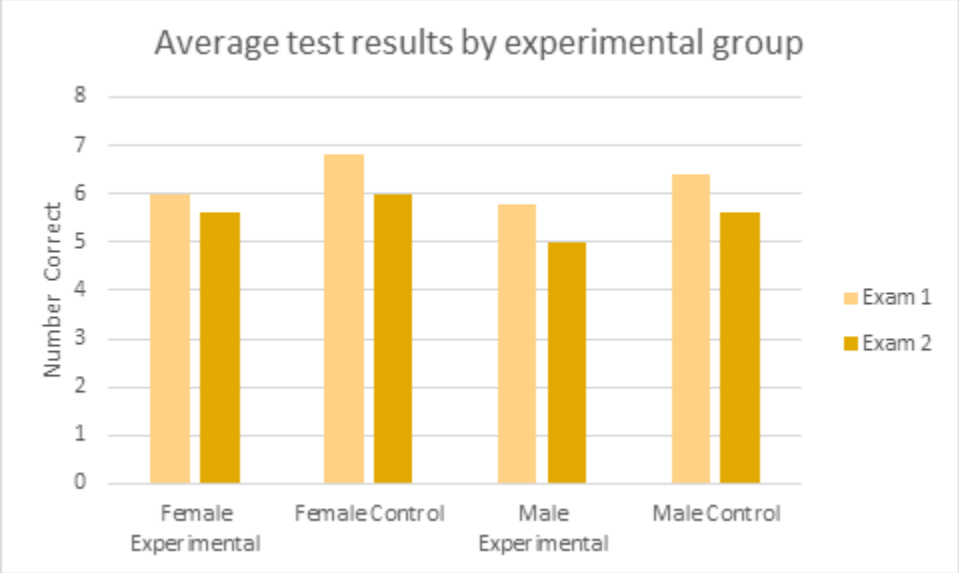

Figure 6. Comparison of average scores for exam 1 and exam 2 for each group. For female experimental n=5, female control n=5, male experimental n=5, male control n=5.

Data Analysis

For respiration rate, heart rate, galvanic skin response, and exam score change ANOVAs were run to determine if the female experimental group differed significantly from the other groups. The correlation matrix shows the correlation of the exam 1 average and the exam 2 average for each set of physiological data (Figure 7). There was no significant interaction between gender and experimental group except for respiration rate (Figure 8).

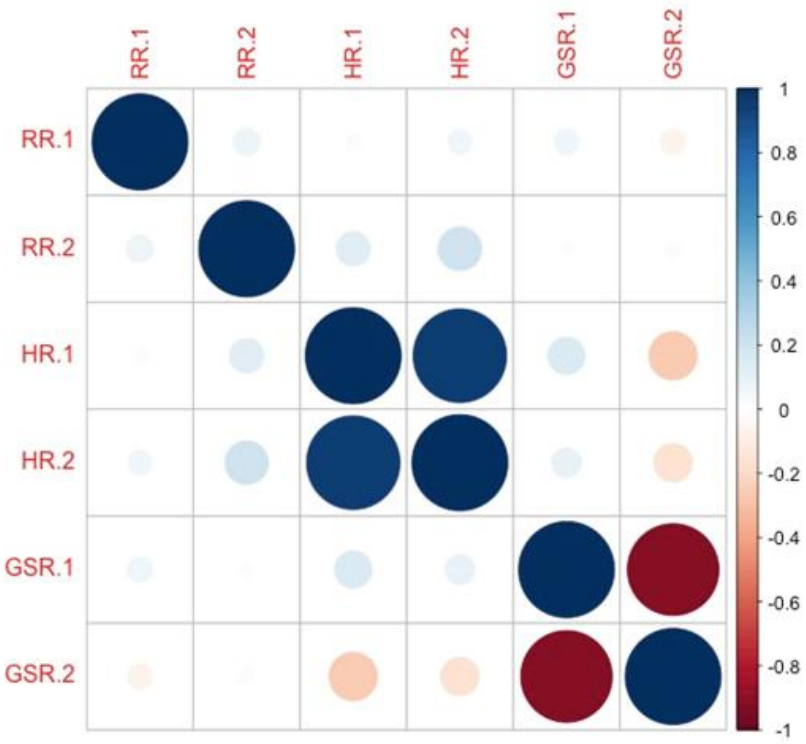

Figure 7. Correlation matrix of physiological data.

```
> fit2 <- aov(RR.C ~ Group * Gender, data=data)
> summary(fit2)
             Df Sum Sq Mean Sq F value Pr(>F)
Group         1  12.13   12.13   1.981 0.1785
Gender        1   4.17    4.17   0.681 0.4213
Group:Gender  1  34.15   34.15   5.576 0.0312 *
Residuals    16  97.98    6.12
```

Figure 8. ANOVA for Respiration Rate data

Discussion

Based on the results of this experiment, we cannot support our hypothesis that physiological stress responses result from gender stereotype threat during the administration of a mathematical exam. Our hypothesis aligned with a similar study looking at gender stereotype threat in female mathematical test performance in which researchers found that the threatened group dropped 10% in accuracy from exam 1 to exam 2 (Beilock *et al.,* 2007). Although no significant change in respiration rate occurred between exam 1 and exam 2 in our experimental groups, as seen in Figure 3, the ANOVA analysis did show a significant difference between group and gender, as seen in Figure 8, suggesting that the articles in between exams may have impacted the subjects' performances. No statistically significant changes were found in the first four experimental groups between exam 1 and exam 2 for heart rate, galvanic skin response, or average exam scores, as seen in Figures 4, 5, and 6, respectively.

Respiration data for the averages for exams 1 and 2 showed no significant difference between experimental group ($p=0.1785$) or between gender ($p=0.4213$). An interesting trend to note, however, is the fact that for each group, regardless of significance, respiration rate actually decreased during exam 2 rather than increased, the latter of which would be expected under conditions of induced physiological stress (Seuss, *et al.,* 1980). A possible explanation for this could be the fact that participants may have adapted to the test taking conditions over time, and therefore relaxed to an extent, potentially steadily decreasing respiration rate, despite the slight increase in difficulty in the second exam (which may be a factor of the decreased scores). The ANOVA test performed between gender and group did, however, show a statistically significant difference ($p=0.0312$), meaning that the female experimental group was significantly different from the other experimental groups.

Heart rate data between groups and gender were found to be insignificant (p=.851). Averages between exam 1 and exam 2 for the female experimental group were respectively 81.5 bpm and 81.45bpm, showing almost no change. The other groups demonstrated a slight increase in heart rate during the second exam, regardless of which group they were assigned.

The data for the female experimental exhibited our expected trend of an increase between exam 1 and exam 2, however these results were insignificant (p=.496). For exam 1 in the female control group, the GSR was measured to be negative, so these results could perhaps be attributed to incorrect calibration (BIOPAC Systems Inc. 2014). Both male groups had decreased readings from exam 1 to exam 2.

Although there was no statistically significant difference between groups or gender for exam score overall (p=0.675), it is notable to mention that each experimental group on average scored lower on exam 2 than they did on exam 1. It was expected that the female experimental group would score lower on exam 2 due to the induction of overt gender stereotype threat through the gendered-media, while all other groups would score the same on average between exams. Since this was not the case, and every group on average scored lower on exam 2, a potential source of error could have been that exam 2 was unintentionally crafted to be more difficult than exam 1.

After analyzing these sets of data, an additional data set was collected as a positive control in which a time constraint was used to positively induce stress in our participants. Giving participants a time constraint of six minutes to complete each exam was done in order to potentially induce stress, and subsequently exams scores, to indicate an inverse relationship between induced stress and score percentage. The average exam score for our unstressed participants was 5.7 out of 8 (n = 47), and the average exam score for our stressed participants

was 4.0 out of 8 (n = 7). An unpaired t-test revealed this to be a highly statistically significant difference (P = 0.0064). Recording this additional set of data was necessary in order to find out that an induced physiological stress response does indeed cause poorer performance on quantitative exams.

Several sources of error may have influenced our results. Most notably, our limited sample size (n=20) hinders our ability to eliminate differences attributed to individuality and reduces our ability to parse out statistically viable and reliable results. There were also several cases where one individual in one or more of the experimental groups were removed from statistical analysis due to equipment failure. There is also the concern of whether or not our stimulus induced an adequate amount of stress to affect test performance. Our exam questions also may not have been difficult enough, therefore stress would not have affected performance (Broadhurst, 1959).

Errors inherent in the equipment or in the calibration of the equipment may have attributed to inconsistencies in data collection. For the respiratory transducer, participants had varied layers of clothing worn underneath the strap, and the placement of the equipment on the subject could not be consistently controlled, leading to potentially altered respiratory readings. Furthermore, for each participant, the amount of gel applied to their fingers for the galvanic skin responses readings were not equal each time, perhaps hindering how well the EDA transducer could detect changes in the readings. We did not use the exact same equipment for every participant, therefore the intrinsic differences within the equipment available may have provided slightly inconsistent results.

We did not ask for subjects' previous experience with the GRE, from which the test questions were pulled. Those who have had previous experience with the GRE may have been

more familiar with the question contents and felt more relaxed, thus performing better on our assessments compared to those who had not taken or studied for the GRE previously. Of those participants that may have had previous experience with the GRE, time could also be a variable, in that subjects' who had more recently taken it would have the content fresh in their mind and performed better than those who had taken it several months to years ago. This could have been easily controlled for by administering a short questionnaire prior to the test, which is a helpful consideration for future studies that may use a similar approach to this experiment. Additionally, our cohort includes students who were admitted to and have successfully navigated the rigors of a highly ranked research university. As science students, it is reasonable to assume that our physiology 435 classmates have developed better test-taking strategies than the general public. It is also reasonable to assume our cohort have learned to cope with the stress response invoked by exams and therefore might show a diminished response to these exams in comparison with the public at large.

We could not accurately control for the same difficulty for both exams, especially since the exam was made by us, the researchers. Perhaps one exam was harder than the other, potentially skewing the exam scores. As mentioned before, it is a reasonable possibility that exam 2 was made slightly more difficult than exam 1, since scores on average were lower in each group on this second exam. In future studies, exam difficulty should be controlled for by having an experienced test-maker create the exam. Our testing environment could also have affected test performance. Some distractions in our environment that could have led to decreased test-taking ability in participants were knocking, people walking through the room, people talking, and other studies happening in close proximity.

Although our results proved to be inconclusive, we believe that there is a correlation between gendered stereotype threat and math performance in females, as alluded to in other previous studies. For future direction, it is advised to increase the study population and to practice more consistent data collection methods to eliminate confounding errors.

Appendix I

OECD Education Report: Boys' Pulling Ahead of Girls' in Math Tests

UK Telegraph Education News

Girls are lagging behind boys in math and science amid concerns over a gender gap at the heart of the education system, major international research has found. The gulf between boys and girls is wider in the US than in most other developed nations. According to figures, boys outperformed girls in math by 12 points on average in the latest Program for International Student Assessment (PISA) tests. The gender gap internationally stands at 11 points, with girls performing better in just five nations. The report was compiled after college students sat independently-administered tests in math, reading and science across 65 developed nations. The exams are supposed to test pupils' ability to apply knowledge.

In a report, the OECD said girls felt "less motivated to learn math and have less confidence in their abilities than boys". It comes amid continuing concerns over a shortage of female scientists and engineers and follows the publication of a Government report last month calling for a step-change in support for girls to raise standards in these disciplines. The OECD report said: "Among girls, the greatest hurdle is in reaching the top: girls are underrepresented among the highest achievers in most countries and economies, which poses a serious challenge to achieving gender parity in science, technology, engineering and mathematics occupations in the future.

What is the main idea of this article?

Figure 9. Article given to participants in the experimental group.

Dogs are Even More Like Us than We Thought

National Geographic

It's likely no surprise to dog owners, but growing research suggests that man's best friend often acts more human than canine. Dogs can read facial expressions, communicate jealousy, display empathy, and even watch TV, studies have shown. They've picked up these people-like traits during their evolution from wolves to domesticated pets, which occurred between 11,000 and 16,000 years ago, experts say.

In particular, "paying attention to us, getting along with us, [and] tolerating us" has led to particular characteristics that often mirror ours, says Laurie Santos.

Social eavesdropping—or people-watching—is central to human social interactions, since it allows us to figure out who's nice and who's mean. According to a study published in August in the journal of *Animal Behavior*, our dogs listen in too.

Gaze following is instinctual for many animals—including humans, chimps, goats, dolphins, and even the red-footed tortoise—because it alerts animals to everything from immediate threats to "a particularly tasty berry bush," says Lisa Wallis.

Dogs were previously thought to follow human gazes only when food or toys were involved. Now, a new study suggests dogs also follow human gazes into blank space—but only if they're untrained.

What was the main idea of this article?

Figure 10. Article given to participants in the control group.

References

Beilock, S.L., R.J. Rydell, and A.R. McConnell. 2007. Stereotype threat and working memory: Mechanisms, alleviation, and spill over. Journal of Experimental Psychology: General http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.594.8754&rep=rep1&type=pdf

Ben-Zeev, T., Fein, S., and Inzlicht, M. 2005. Arousal and stereotype threat. *Journal of Experimental Social Psychology* 41 (2): 174-181. http://www.sciencedirect.com/science/article/pii/S0022103104000629

BIOPAC Systems Inc. 2014. Negative EDA (GSR). http://www.biopac.com/knowledge-base/negative-eda-gsr/

Broadhurst, P.L. 1959. The interaction of task difficulty and motivation: The Yerkes-Dodson law revived. Acta *Psycholgica* 16: 331-338. http://www.sciencedirect.com/science/article/pii/0001691859901052

Paton, Graeme. 2013. OECD education report: boys 'pulling ahead of girls' in maths tests. *The Telegraph*. http://www.telegraph.co.uk/education/educationnews/10489916/OECD-education-report-boys-pulling-ahead-of-girls-in-maths-tests.html

Ryan, A.M., and Nguyen, H.D. 2008. Does stereotype threat affect test performance of minorities and women? a meta-analysis of experimental evidence. *Journal of Applied Psychology* 93 (6): 1314-1334. https://www.researchgate.net/profile/Ann_Ryan2/publication/23489223_Does_stereotype_threat_affect_test_performance_of_minorities_and_women_A_meta-analysis_of_experimental_evidence/links/00b7d51a8da2aeacb0000000.pdf

Schmader, T., Johns, M., and Forbes, C. 2008. An integrated process model of stereotype threat effects on performance. *Psychological review* 115 (2): 336–356. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2570773/

Suess, W. M., Alexander, A. B., Smith, D. D., Sweeney, H. W., & Marion, R. J. 1980. The effects of psychological stress on respiration: a preliminary study of anxiety and hyperventilation. *Psychophysiology* 17.6: 535-540.

Spencer, S.J., Logel, C., and Davies, P.G. 2016. Stereotype threat. *Annual Review of Psychology* 67: 415-437. http://www.annualreviews.org/doi/pdf/10.1146/annurev-psych-073115-103235

Spencer, S.J., Steele, C.M., and Quinn D.M. 1999. Stereotype threat and women's math performance. *Journal of Experimental Social Psychology* 35 (1): 4-28. http://www.sciencedirect.com/science/article/pii/S0022103198913737

Wei-Haas, Maya. 2015. Dogs Are Even More Like Us Than We Thought. *National Geographic*. http://news.nationalgeographic.com/2015/07/150720-dogs-animals-science-pets-evolution-intelligence/