Can Test Anxiety Interventions Alleviate a Gender Gap in an Undergraduate STEM Course?

Rebecca B. Harris,[†‡] Daniel Z. Grunspan,[†‡] Michael A. Pelch,[‡] Giselle Fernandes,[‡] Gerardo Ramirez,[§] and Scott Freeman[‡*]

[†]Center for Evolution & Medicine, Arizona State University, Tempe, AZ 85281; [‡]Department of Biology, University of Washington, Seattle, WA 98195; [§]Department of Psychology, University of California, Los Angeles, Los Angeles, CA 90095

ABSTRACT

Gender gaps in exam scores or final grades are common in introductory college science and engineering classrooms, with women underperforming relative to men with the same admission test scores or college grade point averages. After failing to close a historically documented gender gap in a large introductory biology course using interventions targeted at training a growth mindset, we implemented interventions designed to reduce student test anxiety. We combined evidence-based exercises based on expressive writing and on reappraising physiological arousal. We also used a valid measure to quantify test anxiety at the start and end of the course. This instrument measures an individual's self-declared or perceived test anxiety—also called trait anxiety—but not the immediate or "state" anxiety experienced during an actual exam. Consistent with previous reports in the literature, we found that women in this population declared much higher test anxiety than men and that students who declared higher test anxiety had lower exam scores than students who declared lower test anxiety. Although the test anxiety interventions had no impact on the level of self-declared trait anxiety, they did significantly increase student exam performance. The treatment benefits occurred in both men and women. These data suggest that 1) a combination of interventions based on expressive writing and reappraising physiological arousal can be a relatively easy manner to boost exam performance in a large-enrollment science, technology, engineering, and mathematics (STEM) course and encourage emotion regulation; 2) women are more willing than men to declare that they are anxious about exams, but men and women may actually experience the same level of anxiety during the exam itself; and 3) women are underperforming in STEM courses for reasons other than gender-based differences in mindset or test anxiety.

INTRODUCTION

Achievement gaps are a prominent issue in undergraduate science, technology, engineering, and mathematics (STEM) education because of their negative effects on retention (Cromley *et al.*, 2016; National Science Foundation, 2017). Gaps that impact low-income and underrepresented minority students have received a great deal of attention (e.g., Malcom, 1996; Haak *et al.*, 2011), and recent work has highlighted the importance of achievement gaps based on gender. In the life sciences, for example, Eddy *et al.* (2014) compared performance between women and men in a three-course introductory biology sequence and found that, controlling for college grade point average (GPA), women performed worse than men on course examinations. Similar work using data on academic preparation and ability has shown the same pattern of gendered underperformance across STEM courses and institutions (Creech and Sweeder, 2012; Ballen *et al.*, 2017, 2018; Matz *et al.*, 2017). Closing gender gaps, increasing retention, and boosting the number of women who complete STEM degrees could be an important way to meet calls for an additional one million STEM graduates

Rebecca Price, *Monitoring Editor*
Submitted Jun 6, 2018; Revised Apr 9, 2019; Accepted Apr 25, 2019
CBE Life Sci Educ September 1, 2019 18:ar35
DOI:10.1187/cbe.18-05-0083
*Address correspondence to: Scott Freeman (srf991@uw.edu).

per year (President's Council of Advisors on Science and Technology, 2012).

Gaps in performance metrics are sometimes mirrored by gaps in latent traits, such as a student's sense of belonging, science identity, and self-efficacy (Eddy and Brownell, 2016). To reduce persistent gender gaps in science, researchers have posed several socio-affective hypotheses—each rooted in a different explanation for why gendered underperformance gaps might occur in STEM courses. We briefly review four of these hypotheses and how they have been used in an effort to reduce or eliminate the persistent gender gap in one of the large-enrollment introductory biology courses studied by Eddy *et al.* (2014) at the University of Washington (referred to hereafter as UW biology).

1. *Stereotype threat*: The cognitive load imposed by having to cope with the fear of confirming a negative stereotype about your demographic or cultural group—for example, that females are bad at math or not as brilliant as males—can reduce performance in evaluative situations (Steele, 1997). Stereotype threat has been invoked to explain achievement gaps across a wide array of undergraduate STEM courses (Eddy *et al.*, 2014; McPadden and Brewe, 2015; Matz *et al.*, 2017). Experimental interventions have shown that affirming students' self-integrity using a short affirmation exercise can reduce stereotype threat (Cohen, 2006; Cohen *et al.*, 2009). Although Jordt *et al.* (2017) showed that the affirmation exercise reduced achievement gaps experienced by underrepresented minorities in UW biology, the intervention had no impact on the gender gap.

2. Bias in test design or in course implementation due to instructor gender: Persistent gaps have also been ascribed to a negative effect of having a male instructor (Kapitanoff and Pandey, 2017). Gender gaps have also been linked to test design, with evidence suggesting that males and females perform differently based on the number of gender stereotypes in content (McCollough, 2004; McPadden and Brewe, 2015; Day *et al.*, 2016), the types of items on a test, such as multiple choice or constructed response (Gamer and Engelhard, 1999; Weaver and Raptis, 2001), and the cognitive complexity of questions (Bielinski and Davison, 1998; Wright *et al.*, 2016), though the literature shows some inconsistencies (Madsen *et al.*, 2013; Federer *et al.*, 2016). Similarly, some studies found an effect of instructor gender on reducing achievement gaps (e.g., Dasgupta and Asgari, 2004; Carrell *et al.*, 2010), while other work found limited (Eddy *et al.*, 2014) or no impact (McPadden and Brewe, 2015).

3. *Mindset*: Dweck and colleagues established that students generally hold one of two views about the nature of intelligence. Those with a growth mindset believe that intelligence is malleable and grows through hard work and effort, whereas those with a fixed mindset believe that intelligence is a stable ability that some students have and others do not (Dweck, 2000). These mindsets serve as "meaning systems" that individuals use to make sense of their educational experiences (Dweck and Molden, 2017). Individuals who have a fixed mindset about intelligence tend to react negatively in the face of cognitive challenges and interpret failure as a sign that they are not competent enough to succeed (Dweck, 2000); individuals with a growth mindset about intelligence tend to view challenging situations and failure as an opportunity to escalate effort, often resulting in higher persistence and greater achievement (Dweck, 1975). Some evidence suggests that early cultural influences can produce a gendered difference in mindset, with females more likely to adopt fixed mindsets than males (Leggett, 1985; Dweck and Licht, 1984; Roberts, 1991). Interventions that help students see intellectual growth as a consequence of effort rather than inherent ability have had mixed results. While many individual studies have found mindset interventions to have positive outcomes (Aronson *et al.*, 2002; Blackwell *et al.*, 2007), a recent meta-analysis found an overall weak effect of mindset interventions on academic achievement (Sisk *et al.*, 2018). While this meta-analysis was unable to include the role of gender in mindset interventions, individual studies have shown some promise: mindset interventions performed with junior high school math students have closed gendered achievement gaps (Good *et al.*, 2003; Blackwell *et al.*, 2007), and female high school students with growth mindsets performed better in math classes than their male counterparts (Degol *et al.*, 2017). Work in our lab, however, showed no impact of a growth-mindset intervention on the gender gap observed in UW biology—even though the intervention worked in the sense of leading to a more growth-oriented mindset (for full results on these findings, see Supplemental Material, Appendix 1, Tables S1–S6, and Figures S1 and S2). These results suggested that gendered underperformance in this context was not being driven by differences in mindset.

4. *Stress and anxiety during evaluations*: Women are more prone to clinical anxiety than men and are more willing to self-declare generalized anxiety and test anxiety (Chapell *et al.*, 2005; McLean *et al.*, 2011; Hannon, 2012; Núñez-Peña *et al.*, 2016). Although the biological and cultural basis of these gendered differences is not well understood, the impact of stress and anxiety affecting student performance is well documented (Hembree, 1988; Zeidner and Schleyer, 1999; Cassady, 2004; Goetz *et al.*, 2013). Reduced performance occurs when normal symptoms of physiological arousal, such as increased heart rate and breathing rate, are interpreted as a sign of potential failure (Hembree, 1988). Anxiety also creates intrusive thoughts and concerns that disrupt thinking and contribute to a diminished capacity to recall information. Two distinct types of interventions have been shown to improve performance by reducing the arousal and worry components of anxiety: emotional reappraisal and expressive writing. Emotional reappraisal is an intervention approach that addresses the arousal component of anxiety, by redirecting students to interpret the physiological arousal present during stressful situations as potentially helpful for thinking (Jamieson *et al.*, 2013, 2016). Students are led to read short passages describing arousal as improving alertness and performance under the premise that this new interpretation will help them to adopt a more adaptive interpretation of physiological arousal during stressful circumstances (e.g., during tests). Research has found that students who engage in emotional reappraisal perform better on challenging in-lab tests, course exams, and high-stakes standardized exams like the Graduate Record Examination (Jamieson *et al.* 2010, 2012, 2016). Expressive writing, in contrast, is an activity that targets the worry component of anxiety. It prompts students to write about their emotional states immediately before taking a high-stakes exam, with the goal of clearing working memory of worrisome thoughts

(Klein and Boals, 2001; Ramirez and Beilock, 2011). In previous investigations, writing about one's thoughts and worries immediately before an exam has increased test performance (Frattaroli *et al.*, 2011; Ramirez and Beilock, 2011; Park *et al.*, 2014).

On the basis of the previous success of the emotional reappraisal and expressive writing interventions and the well-established gender difference in self-declared test anxiety, we implemented both approaches in UW biology—a course in which Eddy *et al*. (2014) had documented a persistent gendered achievement gap. We hypothesized that arousal reappraisal and expressive writing could work synergistically to mitigate the physiological and cognitive components of stress during exams. To the extent that the gender gap in science courses is due to stress and anxiety during tests, helping students regulate the arousal and worry components of the experience may offer an advantageous strategy for reducing the gender gap in science exams.

METHODS

We implemented the stress and anxiety interventions during the Autumn quarter of 2016 in one of the largest introductory biology courses in the nation (1140 students) at a large research-intensive university in the northwestern United States (University of Washington). The course was divided into two sections that met back-to-back, took identical exams, and were cotaught by the same two male instructors. Each section met four times per week for a total of 10 weeks and included weekly lab sections and intensive use of active-learning strategies (Freeman *et al.*, 2011).

Points earned throughout the course can be delineated as either high-stakes (exam) or low-stakes (non-exam) points. Low-stakes points were awarded for weekly activities (e.g., laboratory assignments, in-class clicker questions, daily reading quizzes, weekly online practice exams, a field trip, completing surveys) and were designed to reward participation and effort. For instance, many of the low-stakes points were awarded for either a combination of correctness and participation or participation only; involved minimal time pressure; and included access to peer interaction, the assigned textbook, or other resources. Exam points represent the sum total of four exams taken throughout the quarter, each worth 100 points.

The two course sections were assigned to either the treatment or control condition based on a coin flip. This strategy provided quasi-randomized treatment groups (Shadish *et al.*, 2002), as students register to a particular section without a priori knowledge of the experiment. Before each of the four course exams, the students in the treatment group completed two interventions designed to alleviate stress and anxiety during evaluative situations, while the control group conducted similar exercises focused on professional development. Students were not told that an intervention was taking place.

Emotional Regulation Intervention

Part 1: Reappraisal. Students in the treatment section completed variations of the reappraisal intervention developed by Jamieson *et al.* (2012) through an online assignment (Supplemental Material, Appendix 2). These assignments were made available on the Thursday afternoon before a Monday exam, with completion due by early Monday morning. Students were awarded a small number of course points for participation. Each exercise highlighted a different physiological response and how it affected physical and cognitive performance. Students in the control section did an analogous exercise online, at the same time and for the same number of participation points, focused on an aspect of professional development such as participating in undergraduate research or career options for individuals with a degree in the life sciences.

Part 2: Expressive Writing. On the day of the exam, the expressive writing intervention (Ramirez and Beilock, 2011) was assigned in the treatment section as a "pre question." After exams had been passed out and placed facedown on student desks, the prompt was presented on a PowerPoint slide. The prompt ended with a sentence stating that if all students appeared to make a good faith effort, each student would be awarded 3 exam points. Students were allowed to write in response to the prompt for 3 minutes on the blank last page of the exam and were then asked to rip the page off, crumple it up, and throw it away. Only then were students instructed to flip the exam over and begin answering the graded questions. Although the 3-minute interval is much shorter than the writing time allowed in the original experiments, data in Doherty and Wenderoth (2017) indicate that college students write extensively and meaningfully in the time we provided. In the control section, students were also given 3 minutes to answer a pre question on the back of the exam for 3 participation points before starting the actual exam. In this case, however, the writing prompts asked students how they might act on the information about professional development provided in the reappraisal control exercise they had completed previously. In sum, students were assigned to either a treatment condition (in which they were asked to engage in emotional reappraisal and expressive writing) or an active control condition.

To the best of our knowledge, this study is the first time that an embodied cognition component of expressive writing (crumpling up the sheet and throwing it away) was used when an impact on course performance was evaluated. Cognition is embodied when it is shaped by attributes of the physical body (Foglia and Wilson, 2013). The complete prompts for all of the exercises are provided in the Supplemental Material, Appendix 2.

Data Sources

Trait-based test anxiety refers to an individual's perceptions of his or her emotional "habits," while state-based test anxiety refers to immediate feelings during the actual exam (Goetz *et al.*, 2013). Although our goal was to reduce students' state anxiety, we also examined individual differences in trait anxiety. To quantify trait test anxiety at the beginning and end of the term, we asked students to respond to the same 17 questions taken from the Cognitive Test Anxiety Scale (Cassady and Finch, 2014). These were distributed via the online course management system, and students earned participation points for each. We avoided asking students to report their level of state anxiety during exams, as we feared this would reveal the purpose of the intervention.

Anxiety surveys contained no missing data. Thus, Likert-scale responses were converted to numeric values and summed to calculate pre and post anxiety scores for each student. Course performance data were collected from course instructional staff and merged with demographic data from the Office of the

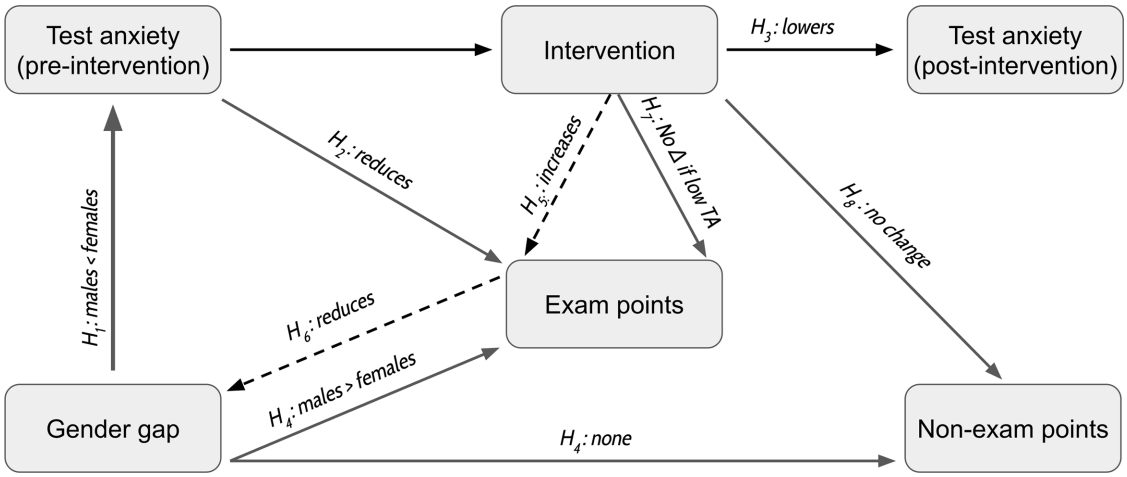

FIGURE 1. Relationships between the hypotheses tested in this study. Solid arrows indicate hypotheses that are drawn from the literature and tested with our data set. Dashed arrows indicate hypotheses that motivated this study. TA, test anxiety.

Registrar and test anxiety scores. The interventions and data collection were performed following a review by the University of Washington Human Subjects Division, application 50071.

Statistical Analyses

Before analyzing the data, we excluded any student who either did not take all four exams, complete the preintervention test anxiety survey, the intervention itself, or the postintervention test anxiety survey.

A schematic of how the hypotheses tested in this study overlap and inform one another is shown in Figure 1. We predicted that, upon entering the course, men would report lower test anxiety than women (hypothesis 1). We next tested whether trait anxiety scores were correlated with course performance. Within the control group, preintervention test anxiety should be significantly associated with performance on high-stakes exam points but should have no association with low-stakes points (hypothesis 2). We expected the interventions to decrease a student's reported trait-level anxiety (i.e., test anxiety) regardless of gender (hypothesis 3). We next reasoned that, if a gender gap in course points exists, such a gap would be more pronounced in high-stakes evaluative testing situations compared with low-stakes (hypothesis 4). These four hypotheses were tested using six separate linear regression models (Supplemental Table S9).

If our data identified an effect of test anxiety on exam performance, supporting hypothesis 4, our next goal was to demonstrate whether the treatment was successful in reducing the maladaptive effect of anxiety on high-stakes exam performance (hypothesis 5). Embedded within this, we asked whether the intervention differentially helped female students (hypothesis 6).

Rather than testing our hypotheses separately, we used a model-selection approach that was designed to identify which variables best predict exam points. Hypothesis 5 was explicitly tested by including the two-way interaction between treatment and test anxiety, while hypothesis 6 was explicitly tested by including the two-way interaction between treatment and gender. Because the intervention could be less beneficial for students with very low test anxiety and because test anxiety often differs by gender, we also included a three-way interaction among gender, test anxiety, and treatment group (hypothesis 7). The full model was

$$\begin{aligned}\text{Exam points: } & \text{gpa} + \text{gender} + \text{test anxiety} + \text{trt} \\ & + \text{gender} * \text{test anxiety} + \text{trt} * \text{gender} \\ & + \text{trt} * \text{test anxiety} + \text{gender} * \text{test anxiety} * \text{trt}\end{aligned}$$

We conducted stepwise backward model selection to find the reduced model that best fit our data using the R package MuMIn (Bartoń, 2018). On the basis of their significance level, we iteratively removed variables from the model until all remaining variables were significant ($p < 0.05$). Using the Aikaike information criterion corrected for small samples (AICc), the best-fit model was chosen as the model with the fewest terms within delta AICc < 2 of the lowest AICc score. Variables that were excluded from the model were considered poor predictors of our data. Thus, if treatment were kept in the best-fitting model, then we knew that the treatment significantly altered course performance. We repeated these analyses using low-stakes non-exam points as the dependent variable, with the expectation that the intervention would not alter performance on these low-stakes questions (hypothesis 8).

We controlled for a student's prior preparation in all analyses that included exam or non-exam points as an outcome. While high-stakes college entrance exam scores are often used as a metric of prior preparation, we suspected they are influenced by test anxiety. The importance of this issue is underlined by our observation that males had higher average Scholastic Aptitude Test (SAT) scores than females in the quarters we did the mindset intervention reported in Appendix 1 in the Supplemental Material ($t = 4.346$, $p < 0.001$; Supplemental Table S6) and the test anxiety intervention reported here ($t = 3.99$, $p < 0.001$; Supplemental Table S8). To prevent circularity in our analyses and create a direct comparison with data in Eddy *et al.* (2014), we used college GPA at the start of the term as an index of academic preparation. First-Fall freshman students were excluded from the analyses, because they did not yet have a college GPA.

RESULTS

A total of 779 upper-division students passed filtering thresholds and were included in the analyses. Before any hypothesis testing, we tested whether the quasi-random experimental design had indeed created equivalent treatment groups (Supplemental Table S7). Sections, and therefore treatment groups, did not differ significantly in college GPA ($F = 0.14$, $p = 0.71$), combined SAT scores ($F = 1.24$, $p = 0.26$), or gender composition ($X^2 = 1.68$, $p = 0.19$). At the start of the quarter and before the interventions, the control and treatment groups did not differ significantly in test anxiety ($F = 1.13$, $p = 0.29$). These observations defend the claim that the two sections in this quasi-experimental design were indeed equivalent.

Finding 1: At the Start of the Course, Test Anxiety Was Higher in Women versus Men

Before the intervention, trait test anxiety was significantly elevated in women compared with men across the entire sample ($\beta = 5.11$, $p > 0.001$).

Finding 2: Test Anxiety Predicts Exam Performance

Within the control group, higher test anxiety predicted lower exam points (ignoring gender and controlling for GPA; $\beta = -0.34$, $p = 0.03$). The reverse was true for non-exam points: more-anxious students scored more points ($\beta = 0.21$, $p = 0.02$).

Finding 3: The Intervention Did Not Change End-of-Course Test Anxiety

Controlling for test anxiety at the start of the course, we found that the intervention did not change self-reported test anxiety (main effect of treatment: $\beta = -0.98$, $p = 0.31$). At the end of the class, we also found that neither men nor women showed a difference in self-declared test anxiety as a function of treatment (interaction: $\beta = 0.58$, $p = 0.63$). While the intervention did not appear to change overall perceptions of test anxiety (i.e., trait anxiety), we were still interested in evaluating whether the intervention would change test scores, which may reflect how students deal with test stress and anxiety during evaluation situations (i.e., state anxiety while taking a test).

Finding 4: There Was No Gender Gap in Exam Points, but There Was in Non-exam Points

Controlling for college GPA at the start of the quarter, women in the control group did not differ from men in exam points ($\beta = -1.191$, $p = 0.572$). In terms of non-exam points, we found that women earned significantly more points than men ($\beta = 5.99$, $p = 0.002$).

Findings 5–7: The Intervention Helped Everyone Earn More Exam Points

Exam points were best predicted by treatment, test anxiety, and preparation. None of the interactions were retained in the final model, demonstrating that, contrary to our expectations, the intervention did not differentially help women (hypothesis 6, Supplemental Figure S3) or students who are highly trait anxious (hypothesis 7). That is, model selection reduced the full model to

$$\text{Exam points: trt} + \text{test anxiety} + \text{GPA}$$

However, this means the test anxiety intervention helped everyone in the treatment group. On average, students in the treatment group earned more exam points than equally matched students in the control group (treatment: $\beta = 4.249$, $p = 0.039$). Students who received our easy-to-implement intervention gained a 1.06% boost in points earned on exams.

Finding 8: The Intervention Did Not Affect Low-Stakes Points

Low-stakes, non-exam points were best predicted by gender, test anxiety, and preparation. Treatment group was not retained in the final model. This result provides discriminant validity to our experiment, as it demonstrates that the intervention did not affect low-stakes situations for which students are presumably not experiencing a high level of state anxiety. Again, none of the interactions were retained in the final model. Our reduced final model was:

$$\text{Non-exam points: gender} + \text{test anxiety} + \text{GPA}$$

Interestingly, both women ($\beta = 2.08$, $p = 0.097$) and highly anxious students ($\beta = 0.161$, $p = 0.005$) earned more non-exam points. For AICc scores and final model coefficients for findings 5–8, see Supplemental Table S10 and Table 1, respectively.

DISCUSSION

Despite being a numerical majority, women in many STEM classrooms underperform compared with men with similar academic profiles (Eddy *et al.*, 2014; Ballen *et al.*, 2017; Matz *et al.*, 2017). We sought to close a previously observed gender gap in

TABLE 1. Coefficients of the best-fit model (test anxiety interventions)

Final model	Variable	Beta	Error	*p* value
Non-exam: GPA + anxiety + gender	GPA	22.626	1.44	<0.001***
	Anxiety	0.161	0.058	0.005**
	Gender	2.084	1.254	0.097
	Intercept	213.615	5.812	<0.001***
Exam: GPA + anxiety + group	GPA	49.81	2.45	<0.001***
	Anxiety	−0.313	0.096	0.001**
	Group	4.249	2.052	0.039*
	Intercept	−153.243	9.979	<0.001***

*$p < 0.05$.
**$p < 0.01$.
***$p < 0.001$.

a large introductory biology course using existing psychological interventions. We initially targeted students' implicit theories about intelligence, without success (see Appendix 1 in the Supplemental Material for details). Consequently, we targeted anxiety during evaluative situations.

The Test Anxiety Interventions Benefited All Students

Combining expressive writing and arousal reappraisal interventions worked: students in the treatment group had higher exam scores than comparable students in the control condition. These impacts were small but significant, especially given the small amount of time needed to achieve them.

While the interventions did result in significantly higher exam scores overall—indicating that they may have reduced state-based anxiety—there was no impact on self-declared trait anxiety and no evidence that females or students with high trait anxiety benefited disproportionately (Supplemental Table S9). This was surprising, given our observations that women declared higher trait-based test anxiety than men and that higher test anxiety had negative impacts on exam performance. Furthermore, our data also show no change in the gender gap for trait-based test anxiety, despite the interventions working to improve exam scores. One way of interpreting these findings is that women may overreport their trait anxiety relative to men (Goetz *et al.*, 2013). Unfortunately, we did not measure whether student's attitudes toward arousal in stressful exam situations were changed by our intervention. It seems possible that encouraging students to reflect on their emotional reaction and directing students to reinterpret the meaning of arousal may lead to students adopting a more adaptive mindset around the efficacy of arousal.

Can Psychological Interventions Help Solve the Gender Gap?

One of the most attractive aspects of existing psychological interventions is their low cost and relative ease of implementation at scale (Yeager and Walton, 2011; Mavranezouli *et al.*, 2015). Indeed, a variety of studies in which researchers seemingly administered the interventions "as is" have shown added value to student academic outcomes (Yeager *et al.*, 2014; Paunesku *et al.*, 2015).

Other work, however, demonstrates that individual and group differences, as well as social context, serve as boundary conditions for the efficacy of psychological interventions (Hanselman *et al.*, 2014, 2017; Harackiewicz *et al.*, 2016). As leaders in the field have noted, psychological interventions are not magic—their efficacy is highly population and context dependent (Yeager and Walton, 2011).

It is also the case that the achievement gaps these interventions seek to close are not always observed (Stoet *et al.*, 2016; Stoet and Geary, 2018). The gender gap in math, for instance, has been found to vary widely based on culture and economic context (Stoet and Geary, 2018), and students in our control conditions did not show a consistent gender gap in exam points—suggesting that the general pattern documented by Eddy *et al.* (2014) is also context dependent.

Because achievement gaps *and* the success of interventions are context dependent, we suggest that one-size-fits-all interventions may not be the most productive first step for faculty who are struggling with persistent achievement gaps in STEM courses. It may be more productive to begin with student interviews or focus groups to interpret data and diagnose why certain populations of students are underperforming. An important follow-up step would be to use this information to 1) customize interventions to fit the classroom structure and address the salient sociocognitive constructs (see Yeager and Walton, 2011; Yeager *et al.*, 2016) or 2) change other aspects of course design. For example, when Doherty and Wenderoth (2017) piloted the expressive writing intervention in a different introductory biology course, they gained helpful insights about their students via their writing. This information led them to schedule an additional exam, and thus lower the points at stake in each exam, in an effort to improve performance (Doherty and Wenderoth, 2017; see also Cotner and Ballen, 2018).

Future Work on Gendered Underperformance in STEM

Our data suggest that gender gaps in classroom performance and trait-based test anxiety are much more variable than previously thought. More robust studies on the academic achievement gap pattern would require a combination of survey data on trait anxiety with physiological data on state anxiety from 1) wearable devices that record heart rate, breathing rate, and other variables or 2) noninvasive analyses of circulating hormones associated with anxiety. A combination of survey, interview, focus group, and physiological data might also inform the hypothesis that a trait versus state discordance exists because men are less willing than women to admit that they are anxious about exams.

If gendered differences in trait and state anxiety exist, they could be explored with the interventions employed here. The expressive writing intervention is targeted at reducing state anxiety, while Jamieson *et al.* (2016) showed that a reappraisal intervention led to lowered trait anxiety. Jamieson *et al.* (2016) did not analyze their data based on gender, however, and neither that study nor this one quantified state anxiety.

Beyond the many questions that remain about test anxiety in undergraduates, what can instructors do about the pervasive pattern of gendered underperformance in STEM? The data presented here are consistent with earlier work indicating that the pattern is limited to high-stakes exams, as women routinely outperform men on non-exam points in STEM courses (Eddy *et al.*, 2014; Wright *et al.*, 2016; Ballen *et al.*, 2017, 2018; Matz *et al.*, 2017). The data presented here also suggest, however, that the root cause is *not* gendered differences in mindset or test anxiety. And because gender gaps appear only intermittently in the large-enrollment course we studied, our data do not support the hypothesis that large class size is a fundamental underlying mechanism (Ballen *et al.*, 2018). Finally, a recent paper failed to support the "variability hypothesis," which claims that gendered underrepresentation in STEM is due to a dearth of female high achievers (O'Dea *et al.*, 2018).

Solutions to the problem of gendered underperformance on STEM exams will remain elusive until context-specific causes are known. Data from a preliminary series of gender-specific focus groups that we conducted may hint at a way forward. The data, summarized in Appendix 3 in the Supplemental Material, Tables S11 and S12, suggest that women in the course that we studied experience or at least declare more negative emotions concerning exams than men. If further research confirms this pattern and links it to either heightened state anxiety during

exams or a decreased willingness to prepare well for exams, it suggests solutions: women might benefit from a reduced emphasis on exams (Ballen *et al*., 2017; Cotner and Ballen, 2018) or early interventions to reduce negative emotions about exam performance (Wäschle *et al*., 2014; Pelch, 2018).

What might these interventions look like? Cheryan *et al*. (2016) recently proposed that promoting participation in STEM depends on increasing self-efficacy and a sense of belonging. For example, instructors who practice the behavior called immediacy—using eye contact, calling students by name, employing appropriate body language, and other cues—have been shown to lower student anxiety and increase performance (Andersen, 1979; Williams, 2010). Student trust in the instructor has also been shown to be correlated with final grade in STEM courses (Cavanagh *et al.*, 2018). Similarly, Kreutzer and Boudreaux (2012) studied five instructors who adopted intensive active-learning techniques in their introductory physics courses. Compared with the same instructors' traditionally taught, lecture-intensive courses, all of the reformed courses achieved stronger student learning gains on a conceptual inventory of relevant content. But in sections taught by one of the five instructors, females also erased the chronic performance gap in the course and achieved gains equal to those of males. The researchers found that this instructor was adapting and implementing key aspects of the "wise schooling" framework proposed by Steele (1997): cultivating optimistic student–teacher relationships, affirming domain belongingness in women, practicing nonjudgmental responsiveness, valuing multiple perspectives, and emphasizing the expandability of knowledge. All of these instructor behaviors should have a positive impact on self-efficacy. If further work shows that women gain disproportionate benefits when instructors employ these types of soft skills in the hard sciences, a generalizable solution to gender gaps in STEM courses may be within grasp.

Limitations

Our study consists of only one replicate conducted at one institution. While there was ample historical evidence for the existence of a persistent gendered achievement gap, we found none, highlighting the variability of achievement gaps among different iterations of the same course.

ACKNOWLEDGMENTS

This research was supported by grant 52008126 from the Howard Hughes Medical Institute to the University of Washington and a grant from the University of Washington's College of Arts and Sciences, Office of the Dean. We thank John Parks for logistical support and advice, Elinore Theobald for guidance on statistical analyses, and the instructional staff of both courses. For comments that improved the article, we thank the Biology Education Research Group at the University of Washington and three anonymous reviewers.

REFERENCES

Andersen, J. F. (1979). Teacher immediacy as a predictor of teaching effectiveness. *Annals of the International Communication Association*, *3*(1), 543–559.

Aronson, J., Fried, C. B., & Good, C. (2002). Reducing the effects of stereotype threat on African American college students by shaping theories of intelligence. *Journal of Experimental Social Psychology*, *38*(2), 113–125.

Ballen, C. J., Aguillon, S. M., Brunell, R., Drake, A. G., Wassenberg, D., Weiss, S. L., … & Cotner, S. (2018). So small classes in higher education reduce performance gaps in STEM? *BioScience*, *68*(8), 593–600.

Ballen, C. J., Salehi, S., & Cotner, S. (2017). Exams disadvantage women in introductory biology. *PLoS ONE*, *12*(10), e0186419.

Bartoń, K. (2018). *MuMIn: Multi-model inference. R package (Version 1.42.1)*. https://CRAN.R-project.org/package=MuMIn/

Bielinski, J., & Davison, M. L. (1998). Gender differences by item difficulty interactions in multiple-choice mathematics items. *American Educational Research Journal*, *35*(3), 455–476.

Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development*, *78*(1), 246–263.

Carrell, S. E., Page, M. E., & West, J. E. (2010). Sex and science: How professor gender perpetuates the gender gap. *The Quarterly Journal of Economics*, *125*(3), 1101–1144.

Cassady, J. C. (2004). The impact of cognitive test anxiety on text comprehension and recall in the absence of external evaluative pressure. *Applied Cognitive Psychology*, *18*(3), 311–325.

Cassady, J. C., & Finch, W. H. (2014). Confirming the factor structure of the cognitive test anxiety scale: Comparing the utility of three solutions. *Educational Assessment*, *19*(3), 229–242.

Cavanagh, A. J., Chen, X., Bathgate, M., Frederick, J., Hanauer, D. I., & Graham, M. J. (2018). Trust, growth mindset, and student commitment to active learning in a college science course. *CBE—Life Sciences Education*, *17*, ar10.

Chapell, M. S., Blanding, Z. B., Silverstein, M. E., Takahashi, M., Newman, B., Gubi, A., & McCann, N. (2005). Test anxiety and academic performance in undergraduate and graduate students. *Journal of Educational Psychology*, *97*(2), 268–274.

Cheryan, S., Ziegler, S. A., Montoya, A. K., & Jiang, L. (2016). Why are some STEM fields more gender balanced than others? *Psychological Bulletin*, *143*(1), 1–35.

Cohen, G. L. (2006). Reducing the racial achievement gap: A social-psychological intervention. *Science*, *313*(5791), 1307–1310.

Cohen, G. L., Garcia, J., Purdie-Vaughns, V., Apfel, N., & Brzustoski, P. (2009). Recursive processes in self-affirmation: Intervening to close the minority achievement gap. *Science*, *324*(5925), 400–403.

Cotner, S., & Ballen, C. J. (2018). Can mixed assessment methods make biology classes more equitable? *PLoS ONE*, *12*(12), e0189610.

Creech, L. R., & Sweeder, R. D. (2012). Analysis of student performance in large-enrollment life science courses. *CBE—Life Sciences Education*, *11*(4), 386–391.

Cromley, J. G., Perez, T., & Kaplan, A. (2016). Undergraduate STEM achievement and retention: Cognitive, motivational, and institutional factors and solutions. *Policy Implications of the Brain and Behavioral Sciences*, *3*, 4–11.

Dasgupta, N., & Asgari, S. (2004). Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender and self stereotyping. *Journal of Experimental Social Psychology*, *40*(5), 642–658.

Day, J., Stang, J. B., Holmes, N. G., Kumar, D., & Bonn, D. A. (2016). Gender gaps and gendered action in a first-year physics laboratory. *Physical Review Physics Education Research*, *12*(2).

Degol, J. L., Wang, M., Zhang, Y., & Allerton, J. (2017). Do growth mindsets in math benefit females? Identifying pathways between gender, mindset, and motivation. *Journal of Youth and Adolescence*, *47*(5), 967–990.

Doherty, J. H., & Wenderoth, M. P. (2017). Implementing an expressive writing intervention for test anxiety in a large college course. *Journal of Microbiology & Biology Education*, *18*(2). 18.2.39.

Dweck, C. S. (1975). The role of expectations and attributions in the alleviation of learned helplessness. *Journal of Personality and Social Psychology*, *31*(4), 674–685.

Dweck, C. S. (2000). *Self-theories: Their role in motivation, personality, and development*. New York: Psychology Press.

Dweck, C. S., & Licht, B. G. (1984). Sex differences in achievement orientations: Consequences for academic choices and attainments. In *Sex differentiation and schooling*. London: Heinemann.

Dweck, C. S., & Molden, D. C. (2017). Mindsets: Their impact on competence motivation and acquisition. In Elliot, A. J., Dweck, A. J., & Yeager, D. S. (Eds.), *Handbook of competence motivation: Theory and applications* (pp. 135–154). New York: Guilford.

Eddy, S. L., & Brownell, S. E. (2016). Beneath the numbers: A review of gender disparities in undergraduate education across science, technology, engineering, and math disciplines. *Physical Review Physics Education Research*, *12*(2), 020106.

Eddy, S. L., Brownell, S. E., & Wenderoth, M. P. (2014). Gender gaps in achievement and participation in multiple introductory biology classrooms. *CBE—Life Sciences Education*, *13*(3), 478–492.

Federer, M. R., Nehm, R. H., & Pearl, D. K. (2016). Examining gender differences in written assessment tasks in biology: A case study of evolutionary explanations. *CBE—Life Sciences Education*, *15*(1), ar2.

Foglia, L., & Wilson, R. A. (2013). Embodied cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(3), 319–325.

Frattaroli, J., Thomas, M., & Lyubomirsky, S. (2011). Opening up in the classroom: Effects of expressive writing on graduate school entrance exam performance. *Emotion*, *11*(3), 691–696.

Freeman, S., Haak, D., & Wenderoth, M. P. (2011). Increased course structure improves performance in introductory biology. *CBE—Life Sciences Education*, *10*(2), 175–186.

Gamer, M., & Engelhard, G., Jr. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education*, *12*(1), 29–51.

Goetz, T., Bieg, M., Lüdtke, O., Pekrun, R., & Hall, N. C. (2013). Do girls really experience more anxiety in mathematics? *Psychological Science*, *24*(10), 2079–2087.

Good, C., Aronson, J., & Inzlicht, M. (2003). Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Journal of Applied Developmental Psychology*, *24*(6), 645–662.

Haak, D. C., HilleRisLambers, J., Pitre, E., & Freeman, S. (2011). Increased structure and active learning reduce the achievement gap in introductory biology. *Science*, *332*(6034), 1213–1216.

Hannon, B. (2012). Test anxiety and performance-avoidance goals explain gender differences in SAT-V, SAT-M, and overall SAT scores. *Personality and Individual Differences*, *53*(7), 816–820.

Hanselman, P., Bruch, S. K., Gamoran, A., & Borman, G. D. (2014). Threat in context: School moderation of the impact of social identity threat on racial/ethnic achievement gaps. *Sociology of Education*, *87*(2), 106–124.

Hanselman, P., Rozek, C. S., Grigg, J., & Borman, G. D. (2017). New evidence of self-affirmation effects and theorized sources of heterogeneity from large-scale replications. *Journal of Educational Psychology*, *109*, 405–424.

Harackiewicz, J. M., Smith, J. L., & Priniski, S. J. (2016). Interest matters: The importance of promoting interest in education. *Policy Insights from Behavioral and Brain Sciences*, *3*(2), 220–227.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research*, *58*(1), 47.

Jamieson, J. P., Mendes, W. B., Blackstock, E., & Schmader, T. (2010). Turning the knots in your stomach into bows: Reappraising arousal improves performance on the GRE. *Journal of Experimental and Social Psychology*, *46*(1), 208–212.

Jamieson, J. P., Mendes, W. B., & Nock, M. K. (2013). Improving acute stress responses. *Current Directions in Psychological Science*, *22*(1), 51–56.

Jamieson, J. P., Nock, M. K., & Mendes, W. B. (2012). Mind over matter: Reappraising arousal improves cardiovascular and cognitive responses to stress. *Journal of Experimental Psychology*, *141*(3), 417–422.

Jamieson, J. P., Peters, B. J., Greenwood, E. J., & Altose, A. J. (2016). Reappraising stress arousal improves performance and reduces evaluation anxiety in classroom exam situations. *Social Psychological and Personality Science*, *7*(6), 579–587.

Jordt, H., Eddy, S. L., Brazil, R., Lau, I., Mann, C., Brownell, S. E., … & Freeman, S. (2017). Values affirmation intervention reduces achievement gap between underrepresented minority and white students in introductory biology classes. *CBE—Life Sciences Education*, *16*(3), ar41.

Kapitanoff, S., & Pandey, C. (2017). Stereotype threat, anxiety, instructor gender, and underperformance in women. *Active Learning in Higher Education*, *18*(3), 213–229.

Klein, K., & Boals, A. (2001). Expressive writing can increase working memory capacity. *Journal of Experimental Psychology*, *130*, 520–533.

Kreutzer, K., & Boudreaux, A. (2012). Preliminary investigation of instructor effects on gender gap in introductory physics. *Physics Education Research*, *8*, 0101020–1.

Leggett, A. J. (1985). Leggett responds. *Physical Review Letters*, *54*(3), 246.

Madsen, A., McKagan, S. B., & Sayre, E. C. (2013). Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap? *Physical Review Special Topics—Physics Education Research*, *9*(2), 020121.

Malcom, S. M. (1996). Science and diversity: A compelling national interest. *Science*, *271*(5257), 1817–1819.

Matz, R. L., Koester, B. P., Fiorini, S., Grom, G., Shepard, L., Stangor, C. G., … & McKay, T. A. (2017). Patterns of gendered performance differences in large introductory courses at five research universities. *AERA Open*, *3*(4), 1–12.

Mavranezouli, I., Mayo-Wilson, E., Dias, S., Kew, K., Clark, D. M., Ades, A. E., & Pilling, S. (2015). The cost effectiveness of psychological and pharmacological interventions for social anxiety disorder: A model-based economic analysis. *PLoS ONE*, *10*(10), e0140704.

McCollough, L. (2004). Gender, context, and physics assessment. *Journal of International Women's Studies*, *5*(4), 20–30.

McLean, C. P., Asnaani, A., Litz, B. T., & Hofmann, S. G. (2011). Gender differences in anxiety disorders: Prevalence, course of illness, comorbidity and burden of illness. *Journal of Psychiatric Research*, *45*(8), 1027–1035.

McPadden, D., & Brewe, E. (2015). The impacts of instructor and student gender on student performance in introductory modeling instruction courses. In *2014 Physics Education Research Conference Proceedings held July 30–31, 2014, in Minneapolis, MN*.

National Science Foundation. (2017). *Women, minorities, and persons with disabilities in science and engineering: 2017* (Special Report NSF 17-310). Arlington, VA.

Núñez-Peña, M. I., Suárez-Pellicioni, M., & Bono, R. (2016). Gender differences in test anxiety and their impact on higher education students' academic achievement. *Procedia—Social and Behavioral Sciences*, *228*, 154–160.

O'Dea, R. E., Lagisz, M., Jennions, M. D., & Nakagawa, S. (2018). Gender differences in individual variation in academic grades fail to fit expected patterns for STEM. *Nature Communications*, *9*, 3777.

Park, D., Ramirez, G., & Beilock, S. L. (2014). The role of expressive writing in math anxiety. *Journal of Experimental Psychology: Applied*, *20*(2), 103–111.

Paunesku, D., Walton, G. M., Romero, C., Smith, E. N., Yeager, D. S., & Dweck, C. S. (2015). Mind-set interventions are a scalable treatment for academic underachievement. *Psychological Science*, *26*(6), 784–793.

Pelch, M. (2018). Gendered differences in academic emotions and their implications for student success in STEM. *International Journal of STEM Education*, *5*, 33.

President's Council of Advisors on Science and Technology. (2012). *Engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering and mathematics*. Washington, DC: U.S. Government Office of Science and Technology.

Ramirez, G., & Beilock, S. L. (2011). Writing about testing worries boosts exam performance in the classroom. *Science*, *331*(6014), 211–213.

Roberts, T. A. (1991). Gender and the influence of evaluations on self-assessments in achievement settings. *Psychological Bulletin*, *109*(2), 297–308.

Shadish, W. R., Cook, T. D., & Campbell, T. D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Sisk, V. F., Burgoyne, A. P., Sun, J., Butler, J. L., & Macnamara, B. N. (2018). To what extent and under which circumstances are growth mind-sets important to academic achievement? Two meta-analyses. *Psychological Science*, *29*(4), 549–571.

Steele, C. M. (1997). A threat in the air. How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*(6), 613–629.

Stoet, G., Bailey, D. H., Moore, A. M., & Geary, D. C. (2016). Countries with higher levels of gender equality show larger national sex differences in mathematics anxiety and relatively lower parental mathematics valuation for girls. *PLoS ONE*, *11*(4), e0153857.

Stoet, G., & Geary, D. C. (2018). The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychological Science*, *29*(4), 581–593.

Wäschle, K., Allgaier, A., Lachner, A., Fink, S., & Nückles, M. (2014). Procrastination and self-efficacy: Tracing vicious and virtuous circles in self-regulated learning. *Learning and Instruction*, *29*, 103–114.

Weaver, A. J., & Raptis, H. (2001). Gender differences in introductory atmospheric and oceanic science exams: Multiple choice versus constructed response questions. *Journal of Science Education and Technology*, *10*(2), 115–126.

Williams, A. S. (2010). Statistics anxiety and instructor immediacy. *Journal of Statistics Education*, *18*(2). DOI: 10.1080/10691898.2010.11889495

Wright, C. D., Eddy, S. L., Wenderoth, M. P., Abshire, E., Blankenbiller, M., & Brownell, S. E. (2016). Cognitive difficulty and format of exams predicts gender and socioeconomic gaps in exam performance of students in introductory biology courses. *CBE—Life Sciences Education*, *15*(2), ar23.

Yeager, D. S., Henderson, M. D., Paunesku, D., Walton, G. M., D'Mello, S., Spitzer, B. J., & Duckworth, A. L. (2014). Boring but important: A self-transcendent purpose for learning fosters academic self-regulation. *Journal of Personality and Social Psychology*, *107*(4), 559–580.

Yeager, D. S., Romero, C., Paunesku, D., Hulleman, C. S., Schneider, B., Hinojosa, C., … & Dweck, C. S. (2016). Using design thinking to improve psychological interventions: The case of the growth mindset during the transition to high school. *Journal of Educational Psychology*, *108*(3), 374–391.

Yeager, D. S., & Walton, G. M. (2011). Social-psychological interventions in education: They're not magic. *Review of Educational Research*, *81*(2), 267–301.

Zeidner, M., & Schleyer, E. J. (1999). The big-fish–little-pond effect for academic self-concept, test anxiety, and school grades in gifted children. *Contemporary Educational Psychology*, *24*(4), 305–329.