

# Paradoxical Gender Gaps in Mathematics Achievement: Pressure as a key

Emily McLaughlin Lyons

Stem Research Network, TERC, 2067 Massachusetts Ave.  
Cambridge, MA 02140

Almaz Mesghina

Department of Comparative Human Development, University of Chicago, Chicago, IL 60637

Lindsey Engle Richland

School of Education, University of California, Irvine; 3200 Education; Irvine, CA 92697

## Abstract

Two studies explore gender gaps that favor girls in low-stakes learning contexts yet are not evident in high-stakes achievement measures. Study 1 ( $n = 386$ ) combined control data across multiple experiments testing student's learning from a challenging proportional reasoning lesson to explore consistent gender gaps in favor of girls. This learning gap could not be explained by the baseline mathematics, affective, motivational, or Executive Function individual differences we measured. In Study 2 ( $n = 178$ ), we experimentally manipulated pressure, raising the stakes by telling some students that their performance would determine whether or not their entire class received an incentive. Gender gaps in favor of girls remained in the absence of pressure, but when external pressure was imposed before or after learning, the female advantage disappeared. These data suggest managing feelings of pressure in learning or testing contexts may be an important step in ultimately increasing female representation in math-intensive fields.

**Keywords:** Mathematics; Gender Gaps; Learning; Reasoning; Pressure; STEM

## Introduction

Gender gaps in careers that rely upon mathematical skills (Science, Technology, Engineering, and Mathematics (STEM) careers) persist in most countries around the world (National Science Foundation, 2017), leading many to assume the solution to be improving girls' early mathematics achievement – yet gender gaps in mathematics are more paradoxical. While boys sometimes outperform girls on high-stakes mathematics assessments (Ellison & Swanson, 2018, Reardon et al., 2018), girls now often outperform boys on many measures of mathematics achievement requiring sustained effort in low-stakes settings (e.g., grades and study behaviors; see Cimpian et al., 2016). Moreover, even when boys and girls exhibit equal mathematics performance, gender differences are often evident in mathematics attitudes and ability perceptions, with girls displaying higher mathematics anxiety and lower (yet more accurate) perceptions of their mathematics ability (e.g. Cimpian et al., 2016, Devine et al., 2012).

We explore cognitive explanations for this paradox, theorizing that the role of pressure may be a key, with working memory (WM) engaged in worries (see Beilock, 2008), differentially impacting the cognitive load of

mathematics for girls and boys. We report on two studies investigating gender differences in fifth and sixth grade students' learning and engagement during mathematics instruction.

Human performance actually peaks under conditions of moderate stress and arousal (see Sapolsky, 2015; Yerkes & Dodson, 1908), yet as stress increases, outcomes plateau and eventually decline (see Sapolsky, 2015). Mathematical thinking and problem solving relies on high level cognitive resources to map correspondences across problems and contexts, to manipulate goals and calculations in mind, and to generalize, make inferences, and overall engage relational and attentional processing (see Vendetti et al., 2015). Pressure too can load these processes. If pressure is perceived as a threat, it can generate intrusive thoughts and worries that are verbally rehearsed (Ashcraft & Kirk, 2001; Eysenck et al., 2007; Schmader et al., 2008). These worries can thereby take up limited cognitive resources, like WM and other executive functions (EFs), that are necessary for task engagement and mathematics performance (Beilock, 2008; Maloney et al., 2014; Schmader et al., 2008). Thus, too much pressure can reduce mathematical learning and test performance due to the competing resource allocation to verbal worries (see Lyons et al., 2017; Maloney et al., 2014).

In regard to gender, women and girls may have higher levels of baseline pressure, even in the absence of imposed pressure (e.g. Goetz et al., 2013). This is especially true in mathematics contexts, in which females may worry that their learning and performance will be judged based on negative stereotypes about women and math (Spencer et al., 1999;). Like imposed pressure, experiencing stereotype threat prior to math tasks can induce worries that consume EFs that are necessary for mathematics performance (Schmader et al., 2008). As early as age 5, girls subscribe to negative stereotypes about women and math (Ambady et al., 2001) and by early elementary school show decrements in mathematics performance when primed to think about these stereotypes (Ambady et al., 2001; Galdi et al., 2014).

In this paper, we report on findings from a series of studies investigating the role of learning context in shaping gender differences in mathematics achievement. In Study 1, we reported findings on gender gaps from several different studies conducted by our lab. In follow up analyses, we focused on student-level cognitive, motivational, and

affective factors that might explain the gender gap in mathematics learning. In Study 2, we considered the role of pressure in learning contexts. We experimentally manipulated increased pressure either before or after the mathematics lesson to test impacts on the gender gap. Students were randomly assigned within classroom to receive a pressure manipulation either before the lesson, before the immediate posttest, or not at all.

In all studies, we implemented a pretest, lesson-and-immediate-posttest, delayed-posttest design, assessing both immediate learning and retention for proportional reasoning concepts and procedures. Items on all three assessments were identical, but counterbalanced in order. Students interacted with a high-quality, cognitively demanding video-taped math lesson in their regular math classrooms - maximizing ecological validity, while allowing for controlled stimuli. Because we wanted to assess initial learning of proportional reasoning, we chose fifth and sixth grades because they possess the prerequisite fraction and division skills but have yet to receive formal instruction on proportional reasoning. Both the lesson and assessment items were challenging in that they were cognitively demanding – requiring students to hold in mind and manipulate multiple solution strategies (Begolli & Richland, 2016). Thus, the lesson and assessment items were appropriate for their grade level, but the content and presentation were challenging.

## Study 1

The primary aim of Study 1 was to characterize a large, consistent gender gap in mathematics learning from one lesson on proportional reasoning. We have used this lesson for 7 studies that were conducted between 2015 and 2017 in 16 diverse elementary schools in the greater Chicago area. For the purposes of Study 1, we analyzed the data of students in the non-experimental conditions. Students in these conditions did not receive any imposed pressure; their learning and performance contexts were quite low-stakes as students' grades were not impacted by their performance in the experiment. A secondary exploratory aim of the study was to investigate whether gender differences in students' affective, cognitive, and motivational factors may contribute to gender differences in their learning. We tested whether girls and boys differed at baseline in their EFs, prior related knowledge, learning orientations, or mathematics anxiety. We also tested whether they differed in their subjective experience of the mathematics learning opportunity by assessing their situational interest and ability perceptions after the lesson. Finally, we explored whether any differences explained the observed gender gap in learning from the proportional reasoning lesson.

## Method

### Participants

Participants were 386 diverse fifth and sixth grade students (17% White, 48% Black, 17% Latinx, all others under 1%, missing, or answered "Other.").

### Procedure

Procedures were group administered during three visits to each classroom over a two-week period. Students completed all procedures alongside their peers in their everyday math classes.

**Session 1.** Students completed a pretest assessing their initial understanding of proportional reasoning, the material to be covered in the lesson. A subset of students also completed measures of mathematics anxiety ( $n = 249$ ) and learning orientations ( $n = 252$ ).

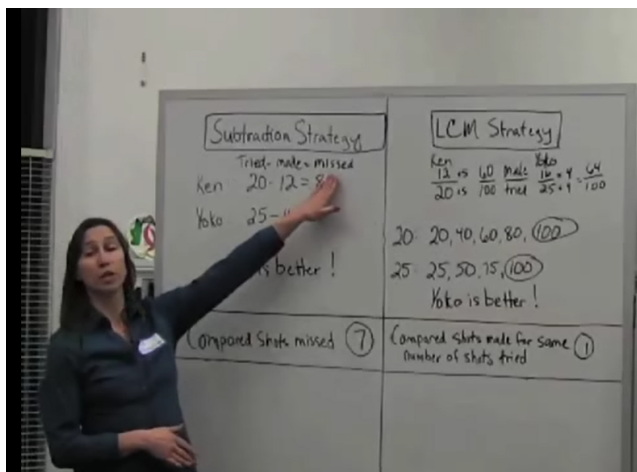
**Session 2.** Two to three days later, students viewed a previously-recorded, conceptually challenging mathematics lesson on individual computers. All students completed a post-test immediately following the lesson. A subset of students ( $n = 343$ ) also completed a self-report measure of their situational interest during the lesson. A different subset of students ( $n = 55$ ) also completed a self-report measure of their perceived ability on the posttest.

**Session 3.** One week after the lesson, students completed a delayed posttest and a measure of EFs.

### Math lesson

We examined student learning during a single high-quality yet challenging instructional opportunity. In the thirty-minute video lesson, a teacher introduces proportional reasoning to a real class of fifth grade students. Specifically, the teacher compares a correct strategy (least common multiple) and an incorrect strategy (subtraction, a common misconception) to solve proportional reasoning problems. The teacher uses high-quality and highly recommended teaching strategies (e.g. simultaneous comparison of strategies, linking gestures) that are highly recommended but often pose a challenge to students insofar as they must manipulate a lot of information at once (see Figure 1; Begolli & Richland, 2016).

**Figure 1.** Screenshot from the video proportional reasoning lesson.





### Measures

Below, we first provide details on our measures of student mathematics achievement and learning. We then provide details on measures of student learning orientations, EFs, and ability perceptions.

**Baseline measure: Pretest mathematics score.** We obtained measures of students' misconception use and accuracy prior to the mathematics lesson. Students' pretest

misconception score was calculated as the proportion of

6) Shani and Keisha have both set up lemonade stands. Shani's lemonade recipe uses 2 cups of lemon juice and 1 cup of water. Keisha's lemonade recipe uses 3 cups of lemon juice and 2 cups of water.

Shani's Lemonade	Keisha's Lemonade
	

Whose lemonade tastes *more* "lemony"?

problems they attempted to solve using subtraction at pretest (out of a possible 5 items). Students' pretest accuracy score was calculated as the proportion of problems answered correctly (10 items) or solved using a valid strategy (5 items) at pretest (15 points total). For full test properties, see Begolli and Richland (2016). See Figure 2 for a sample assessment item.

**Figure 2.** Sample item from the proportional reasoning assessment.

**Mathematics gains.** We obtained four primary measures of mathematics gains following the lesson: immediate gains in accuracy, immediate gains in misconception use, sustained gains in accuracy, and sustained gains in misconception use. Immediate and sustained gains in accuracy were calculated by subtracting the proportion of problems students answered correctly or solved using a valid strategy (out of 15 possible points) at pretest from their scores on the immediate and delayed posttests. Immediate and sustained changes in misconception use were calculated by subtracting the proportion of problems students had attempted to solve using subtraction at the pretest from the proportion of problems they used subtraction on at the immediate and delayed posttests (out of 5 possible problems).

**Baseline measure: Attentional Control Measure of EF.** Student EFs were assessed using the d2 Test of Attention, a measure of sustained and selective attention and inhibitory control (Brickenkamp & Zillmer, 1998). The task requires participants to search for target characters ("d"s with two dashes surrounding it) from among perceptually similar distractors (e.g., "d"s with one dash, "p"s with two dashes) under a time pressure and was group administered to each class. The focal outcome score analyzed in this study was the total number of items processed minus errors (TN-E), which yielded a range of 133 to 539.

**Baseline measure: Learning orientations.** Students' learning orientations were assessed using the Mastery Goal Orientation, Performance-Approach Goal Orientation, and Performance-Avoid Goal Orientation subscales from the Patterns of Adaptive Learning Survey instrument (Midgley et al., 2000).

**Baseline measure: Mathematics anxiety.** Due to time constraints during data collection, mathematics anxiety was assessed using a single item measurement. Students used a 5-point Likert scale (1: *Not at all* to 5: *Very much*) to report the extent to which they agreed with the statement "Math makes me nervous." This measure of mathematics anxiety was modeled after other single-item mathematics anxiety measures (Gogol et al., 2014; Núñez-Peña et al., 2014) that

retain high validity, test-retest reliability, and correlation with other full-scale math anxiety measures.

**Subjective experience: Situational interest.** Situational interest was assessed using an abbreviated version of the Situational Interest Survey, an instrument designed to measure five components of situational interest in a task for middle school students (Chen et al., 2001). In our abbreviated version, students completed two items each in which they reported on the extent to which they found the content in the mathematics lesson enjoyable (instant enjoyment) and would like to learn more about the content (exploration intention).

**Subjective experience: Ability perceptions.** After the immediate post-test, a subset of students responded to the following question on a 5-point Likert scale (1: *Very bad* to 5: *Very good*): "How well do you think you did on the test?" This item, modelled after the mathematics subscale of the Self-Description Questionnaire (Ganley & Lubienski, 2016; e.g. "I get good grades in math"), was modified to assess students' task-specific perception of their mathematics ability.

## Results

### Analytic Plan

We first describe pretest performance and report on the emergence of overall gender gaps in learning that are evident both immediately following the lesson and at a one-week delay. We next test whether these gender gaps are robust or might be explained by demographic variables or differences in baseline cognitive factors like prior knowledge and student EFs. We then report on exploratory analyses in which we examine in a subset of participants whether boys and girls differed in baseline motivational factors, affective factors, or subjective experience of the lesson.

### Pretest Performance and Overall Gender Gaps in Learning

Pretest achievement did not differ between boys and girls on either the accuracy or misconception measure. However, a one-way analysis of variance (ANOVA) indicated that girls had larger gains in correct content both immediately following the lesson ( $F(1,383) = 9.59, p = .002$ ) and at a one-week delay ( $F(1,383) = 17.06, p < .001$ ), and also had greater sustained declines in misconception use ( $F(1,383) = 4.73, p = .03$ ). Girls on average declined in their use of the misconception strategy from pre-test to immediate and delayed post-test, whereas boys on average increased in misconception use at both timepoints.

We next conducted regression analyses to test whether these gender differences were robust or might simply be artifacts of demographic variables, differences in pretest performance, or EFs. Additionally, we tested whether gender may interact with our covariates of interest, suggesting prior knowledge, EFs, or demographic factors may differentially benefit boys and girls. When adding controls, gender remained an important predictor of learning, with girls showing larger immediate ( $B = 1.29; SE B = 0.40; p = .001$ ) and sustained ( $B = 1.75, SE B = 0.39, p < .001$ ) gains in mathematics content, as well as greater sustained declines in misconception use ( $B = -0.40, SE B = 0.19, p = .04$ ). With

regards to Model 3, gender did not interact with either race or EFs to predict any changes in students' accuracy or misconception use (see Table 1 for full regression model for students' sustained gains in accuracy).

A next set of analyses examined whether this gender gap could be explained by learning orientation, math anxiety, or situational interest, but they could not (all  $ps > .05$ ).

Lastly, a one-way ANOVA revealed a significant gender difference in ability perceptions: ( $F(1, 53) = 3.82, p = .05$ ). Girls reported having lower confidence in their performance than boys. Notably, this is despite girls demonstrating greater learning gains than boys.

## Discussion

In sum, we found a consistent, large gender gap favoring girls in learning from a conceptually-rich mathematics lesson on proportional reasoning. This gender gap existed across multiple studies, schools, and classrooms. Though they had similar pretest performance, girls reliably learned more than boys from the lesson and retained these gains after a week's delay. Girls also had lower rates of misconception use at both immediate and delayed posttests, again indicative of greater learning from the lesson, whereas boys increased in their use of the misconception. These findings are consistent with a growing body of literature (e.g. Easton et al., 2017; Souchal et al., 2014) demonstrating that in low-stakes learning and performance contexts, girls outperform boys. What remains to be addressed, however, is why girls outperform boys in these settings.

Greater learning gains among girls in this low-stakes setting could not be explained by individual differences in students' mathematics anxiety, learning orientations, situational interest, or EFs. Boys and girls in our sample did not differ on any of these variables. However, consistent with prior work (e.g. Else-Quest et al., 2010; Ganley & Lubienski, 2016), we did find gender differences in students' ability perceptions, as boys reported greater confidence in their performance than girls, despite girls outperforming boys.

**Table 1.** Regression analyses showing relations between gender and sustained gains in accuracy. Controls include student EFs, pretest score, and race.

	Sustained Gains in Accuracy				
	$R^2$	$B$	$SE B$	$t$	$p$
<i>Model 1: No controls</i>	0.04				
Gender		1.66	0.4	4.13	<.001
<i>Model 2: With controls</i>	0.15				
Gender		1.75	0.39	4.49	<.001
Race		0	0.15	-0.1	0.93
Pretest		-0.3	0.06	-4.8	<.001
EF		0.02	0.003	5.36	<.001

<i>Model 3: Testing interactions</i>	0.16				
Gender		0.85	2.06	0.41	0.68
Race		0.13	0.21	0.63	0.53
Pretest		-0.2	0.09	-2.2	0.03
EF		0.01	0.005	2.6	0.01
Gender* Race		-0.3	0.3	-0.9	0.35
Gender*Pretest		-0.2	0.12	-1.9	0.06
Gender*EF		0.01	0.006	1.13	0.26

## Study 2

In Study 2, we turn our attention to external factors. We focus in particular on the role of learning context in shaping gender gaps, specifically the extent to which learning is higher-stakes or pressured. How does heightened pressure impact children's learning and does this differ for boys and girls?

We predicted that, on the one hand, heightened pressure could boost motivation and effort, resulting in improved learning and performance. But, on the other hand, pressure could also result in anxious ideation and intrusive thoughts that interfere with learning and performance.

We examine whether average impacts of pressure on learning and performance differ between boys and girls. We believe pressure may play a role in shaping gender gaps in mathematics achievement, as prior research suggests a larger performance boost in response to incentives (Levitt et al., 2016) and high-stakes testing contexts (Attali et al., 2011) among males compared to females. Moreover, we expand upon the literature on boys' and girls' mathematics achievement under pressure to include learning, as opposed to just test performance. Prior work suggests that manipulating the framing of an upcoming assessment can create gender differences in high school students' STEM learning (Souchal et al., 2014). Such a finding suggests that gender differences in average impacts of pressure on learning and performance may play a role in explaining some of the seemingly paradoxical patterns of children's mathematics achievement noted above.

## Method

### Participants

Study participants were fifth grade students drawn from five schools in the Chicago area. Participating schools included a traditional public school, two Catholic schools, a charter school, and a private school. A total of 205 students participated. 27 students who were absent on one or more study days were excluded due to missing data, leaving 178 students (90 girls; 25% Hispanic, 30% African American, 25% White, 20% Biracial).

### Design and Procedures

Study measures and procedures were identical to Study 1 with a couple exceptions. First, at the end of Session 2, all students were provided a non-required math puzzle activity

to complete if desired. Second, we did not assess ability perceptions in Study 2.

Lastly, and most critically, we added a pressure manipulation in Study 2. Prior to session 2, students were randomly assigned within each classroom to experience heightened pressure during learning (LP condition), during testing (TP condition), or not at all (NP condition). To enable within-classroom condition assignment, the pressure manipulation was delivered via computer, either at the start (LP) or end (TP) of the lesson.

We modeled our pressure manipulation following Beilock and colleagues' (2004), which effectively induced feelings of pressure and social-evaluative threat by informing participants that their performance would determine not only whether or not they would receive a reward, but also whether or not a partner would receive a reward (Beilock et al., 2004).

Either before learning (LP condition) or before testing (TP condition), students in the pressure conditions were told that they would be taking a test, and if they scored at least 80%, their class would be given a pizza party, but if they failed to earn 80% or higher, their class would lose the pizza party. In contrast, students in the NP condition were told the aim of the study was to better understand how students learn math and were told that after the lesson they would be asked to solve some problems. All prompts were made visible and narrated on the laptop screen.

## Results

### Analytical Plan

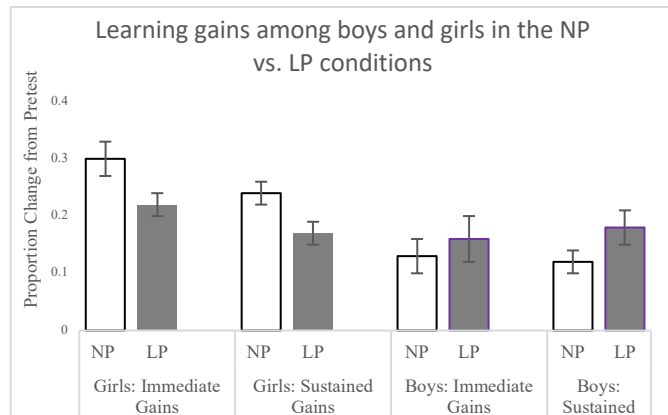
We first describe pretest performance between the three conditions and confirm success of random assignment. In the description of pretest performance, we also report on student-level factors that predicted pretest performance. We next describe gender differences in student learning gains across the no pressure (NP) and high pressure (LP and TP) experimental conditions, and test whether impacts of pressure during learning or testing differed for boys and girls. Finally, we describe gender differences in student engagement in the high pressure vs. no pressure experimental conditions, and test whether the role of pressure in shaping engagement differed for boys and girls. In learning and engagement analyses, we narrow our focus on the role of pressure while learning (LP) in particular.

### Pretest Performance

Pretest performance did not differ between the three conditions or between boys and girls (all  $ps > .24$ ). Pretest performance was not predicted by student EFs or race, but did differ between schools. A dummy variable for student school, along with pretest performance, were included as controls in all analyses of learning outcomes.

### Gender Gaps in Learning in the High vs. No Pressure Experimental Conditions

We first conducted a series of regressions to examine predictors of student learning in the absence of pressure, among the 54 students (28 girls) assigned to the NP condition. The only student characteristic that predicted either immediate or sustained learning gains among students in the



NP condition, after controlling for school and pretest performance, was gender, mirroring findings from Study 1. Girls exhibited significantly larger immediate learning gains ( $\beta_{\text{standardized}} = 0.29, p = .03$ ), and trended towards greater sustained learning gains ( $\beta_{\text{standardized}} = 0.21, p = .09$ ).

We then repeated the same analysis to examine predictors of student learning in the pressure conditions; gender did not predict learning gains among students in either the LP or TP conditions (all  $ps > .45$ ). Mirroring broader-scale patterns of achievement, girls had larger learning gains in the lower pressure context, while the gender gap disappeared, and showed possible trends towards reversing, in the pressure conditions. Neither student EFs nor race predicted learning outcomes, although it is possible that the model may have been too underpowered to detect these relations.

### Gender Differences in Impacts of Learning Pressure on Learning Outcomes

To examine whether gender differences in the role of pressure in shaping learning might help explain the differences in gender gaps in math across pressure vs. no pressure contexts, we next examined main effects and interactions of the Learning Pressure (LP) manipulation and gender.

Pretest score, along with a dummy variable for school, were first entered into the regression (Step 1), as control variables. Main effects (student gender, LP) were added at Step 2. Finally, to test the possibility that the role of heightened pressure in shaping learning differed for boys and girls in this study, an LP-by-gender interaction term was added to the analysis (Step 3). The analysis indicated that gender interacted with the LP manipulation to predict sustained learning gains ( $\beta_{\text{standardized}} = -0.32, p = .05$ ) and may have also interacted to predict immediate learning gains ( $\beta_{\text{standardized}} = -0.27, p = .07$ ).

To better understand these interactions, we next examined effects of the LP study manipulation among boys and girls separately. The analyses indicate that, among girls, heightened pressure during instruction predicted smaller learning gains, suggesting that pressure acted more as a distracting threat than as a motivating incentive (see Figure 3). Girls assigned to the LP condition had smaller immediate ( $\beta_{\text{standardized}} = -0.26, p = .04$ ) and sustained ( $\beta_{\text{standardized}} = -0.29, p = .003$ ) learning gains than NP girls. In contrast, the LP manipulation did not harm boys' learning. Instead, boys who were assigned to the LP condition actually had numerically



larger learning gains as compared to boys assigned to NP, although these differences were not statistically significant (Figure 3).

**Figure 3.** Difference in boys' and girls' learning gains between the No Pressure (NP) and Learning Pressure (LP) conditions. Error bars are  $\pm 1$  standard error.

### Gender Differences in Impacts of Learning Pressure on Engagement during Learning

We next conducted a series of single linear regressions to test for gender differences in engagement (enjoyment, exploration intention, and likelihood of completing a non-required math activity) across LP vs. NP conditions. Mirroring findings for learning outcomes, in the NP condition, girls exhibited higher engagement as compared to their male counterparts. NP girls attempted more optional math puzzles ( $B = 1.63$ ,  $SE B = 0.69$ ,  $p = .02$ ) and completed a greater number of these puzzles successfully ( $B = 1.96$ ,  $SE B = 0.61$ ,  $p = .002$ ). Additionally, with regards to situational interest in the lesson, girls in the NP condition trended towards reporting greater exploration intention ( $B = 0.47$ ,  $SE B = 0.29$ ,  $p = .09$ ) and numerically reported greater enjoyment than boys.

We next examined whether gender predicted these same outcomes among students experiencing heightened pressure during learning (LP). Among students in LP, boys completed more optional math puzzles successfully ( $B = -1.11$ ,  $SE B = 0.55$ ,  $p = .045$ ), and reported numerically greater enjoyment and exploration intention than girls. In sum, under no pressure, girls had greater engagement than boys, but after applying pressure during learning, these patterns reversed.

### Discussion

The findings reported here provide support for the possibility that gender differences in the role of pressure in shaping mathematics learning and engagement may help to explain the seemingly paradoxical ways in which mathematics achievement remains patterned by gender. Girls learned more and exhibited higher engagement outcomes under no pressure. In contrast, gender gaps in learning disappeared under pressure, with boys and girls showing similar learning gains, and boys showing greater engagement, when pressure was experienced either before or after learning.

The reversal of gender gaps across the no pressure versus heightened pressure experimental conditions raises the question as to whether this is due to heightened pressure facilitating boys' mathematics learning and engagement or harming girls' mathematics learning and engagement. Answering this has important implications for practice because better identifying when pressure helps versus harms learning could help to support learning for all students by allowing educators to leverage the potential for pressure to act as a motivator while also minimizing its potential to act as a distracting threat. The clearest answer from this experiment is that experiencing pressure during learning was harmful for girls (on average). Compared to girls in the no pressure condition, girls who experienced pressure while

learning had significantly smaller learning gains immediately following the lesson and these differences persisted one week later, even when pressure was no longer heightened. Compounding these direct effects on learning, girls who experienced pressure while testing were less likely to attempt and complete optional math activities than girls in the no pressure condition. However, trends in the data suggest that the disappearance or reversal of the gender gap in the heightened pressure experimental conditions may also be partially due to pressure boosting boys' learning and engagement outcomes.

### General Discussion

In summary, across two studies, we explored the role of pressure in the learning context as one possible explanation for the often observed, disparate patterns of gender gaps in mathematics performance. Consistent with literature finding a male advantage in higher-stakes mathematics contexts, and a female advantage in lower-stakes mathematics contexts, we found that girls outperformed boys only in the absence of imposed pressure. The gender gap disappeared when pressure, particularly pressure prior to learning, was applied in Study 2. Patterns in student self-reported engagement and motivation largely mirrored gender differences in impacts of pressure on learning, suggesting that pressure may harm girls' engagement and subsequent learning from a lesson, whereas pressure may boost engagement and learning for boys.

This work has serious implications for educators and policy makers. Importantly, the findings from this study and similar work (e.g. Attali et al., 2011; Souchal et al., 2014) suggest that high-stakes assessments might not accurately represent students' actual content knowledge. Rather, the framing of STEM assessments likely determines the extent to which one is able to demonstrate their ability, and consequently, pursue careers and other opportunities in STEM fields. Therefore, one way to reduce gender gaps in STEM may be to change how we assess STEM preparedness. For example, Souchal and colleagues (2014) find that if high-stakes science assessments are presented to students as a learning opportunity, rather than a performance-focused test of ability, gender gaps in science performance scores are substantially reduced. Importantly, both boys and girls show highest performance under a learning opportunity framing, suggesting this framing can help all students succeed in STEM (Souchal et al., 2014).

Taken together, these data provide evidence to suggest that pressure in the learning context is an important contributor to gender differences in mathematics performance, setting the foreground for differences in STEM career engagement.

### References

- Ambady, N., Shih, M., Kim, A., & Pittinsky, T. L. (2001). Stereotype susceptibility in children: Effects of identity activation on quantitative performance. *Psychological Science*, 12, 385–390.
- Ashcraft, M. H., & Kirk, E. P. (2001). The relationships

- among working memory, math anxiety, and performance. *Journal of Experimental Psychology: General*, 130, 224–237.
- Attali, Y., Neeman, Zvika., & Schlosser, A. (2011). Rise to the Challenge or Not Give a Damn: Differential Performance in High vs. Low Stakes Tests. *IZA Discussion Paper No. 5693*.
- Begolli, K.N. & Richland, L.E. (2016). Teaching mathematics by comparison: Analog visibility as a double-edged sword. *Educational Psychology*, 107, 194-213.
- Beilock, S. L. (2008). Math performance in stressful situations. *Current Directions in Psychological Science*, 17, 339–343.
- Beilock, S. L., Kulp, C. A., Holt, L. E., & Carr, T. H. (2004). More on the fragility of performance: Choking under pressure in mathematical problem solving. *Journal of Experimental Psychology: General*, 133, 584-600.
- Brickenkamp, R. & Zillmer, E. (1998). The d2 test of attention. Hogrefe & Huber.
- Chen, A., Darst, P. W., & Pangrazi, R. P. (2001). An examination of situational interest and its sources. *British Journal of Educational Psychology*, 71(3), 383-400.
- Cimpian, J. R., Lubienski, S. T., Timmer, J. D., Makowski, M. B., & Miller, E. K. (2016). Have gender gaps in math closed? Achievement, teacher perceptions, and learning behaviors across two ECLS-K cohorts. *AERA Open*, 2(4).
- Devine, A., Fawcett, K., Szucs, D., & Dowker, A. (2012). Gender differences in mathematics anxiety and the relation to mathematics performance while controlling for test anxiety. *Behavioral and Brain Functions*, 8(33), 1-9.
- Easton, J. Q., Johnson, E., & Sartain, L. (2017). *The predictive power of ninth-grade GPA*. Chicago, IL: The University of Chicago Consortium on School Research.
- Ellison, G. & Swanson, A. (2018). Dynamics of the Gender Gap in High Math Achievement. The National Bureau of Economic Research: NBER Working Paper No. 24910.
- Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, 136, 103–127.
- Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion*, 7(2), 336-353.
- Galdi, S., Cadinu, M., & Tomasetto, C. (2014). The roots of stereotype threat: When automatic associations disrupt girls' math performance. *Child Development*, 85(1), 250-263.
- Ganley, C. M. & Lubienski, S. T. (2016). Mathematics confidence, interest, and performance: Examining gender patterns and reciprocal relations. *Learning and Individual Differences*, 47, 182-193.
- 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.
- Gogol, K., Brunner, M., Goetz, T., Martin, R., Ugen, S., Keller, U., Fischbach, A., & Preckel, F. (2014). "My questionnaire is too long!" The assessments of motivational-affective constructs with three-item and single-item measures. *Contemporary Educational Psychology*, 39(3), 188-205.
- Levitt, S. D., List, J. A., Neckerman S., & Sadoff, S. (2016). The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance. *AEJ: Policy*, 8(4), 183 – 219.
- Lyons, E. M., Simms, N., Begolli, K. N., & Richland, L. E. (2017). Stereotype threat effects on learning from a cognitively demanding mathematics lesson. *Cognitive Science*, 42(2), 678-690.
- Maloney, E. A., Sattizahn, J., & Beilock, S. L. (2014). Anxiety and cognition. *WIREs Cognitive Science*, 5, 403–411.
- Midgley, C., Maehr, M.L., Hruda, L., Anderman, E.M., Anderman, L., Freeman, K.E., Gheen, M., Kaplan, A., Kumar, R., Middleton, M.J., Nelson, J., Roeser, R., & Urdan, T. (2000). *Manual for the Patterns of Adaptive Learning Scales*. University of Michigan.
- National Science Foundation (NSF) (2017). Women, minorities, and persons with disabilities in science and engineering: 2017 (Special Report NSF 17-310).
- Núñez-Peña M. I., Guilera, G., & Suarez-Pellicioni, M. (2014). The single-item math anxiety scale: AN alternative way of measuring mathematical anxiety. *Journal of Psychoeducational Assessment*, 32(4), 306-317.
- Reardon, S. F., Fahle, E. M., Kalogrides, D., Podolsky, A., & Zarate, R. C. (2019). Gender Achievement Gaps in U.S. School Districts. *American Educational Research Journal*, 56(6), 2474-2508.
- Sapolsky, R. M. (2015). Stress and the brain: Individual variability and the inverted-U. *Nature Neuroscience*, 18, 1344-1346.
- Schmader, T. & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology*, 85(3), 440-452.
- Schmader, T., Johns, M., & Forbes, C. (2008). An integrated process model of stereotype threat effects on performance. *Psychological Review*, 115(2), 336-356.
- Souchal, C., Toczek, M., Darnon, C., Smeding, A., Butera, F., & Martinot, D. (2014). Assessing does not mean threatening: The purpose of assessment as a key determinant of girls' and boys' performance in a science class. *British Journal of Educational Psychology*, 84, 125-136.
- Spencer, S., Steele, C., & Quinn, D (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, 351, 4-28.
- Vendetti, M., Matlan, B., Richland, L., Bunge, S. (2015). Analogical reasoning in the classroom: insights from cognitive science. *Mind, Brain, and Education*, 9(2), 100–106.
- Yerkes, R. M. & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit formation. *Journal of Comparative Neurology and Psychology*, 18, 458–482.