MIND, BRAIN, AND EDUCATION

Complicated Gender Gaps in Mathematics Achievement: Elevated Stakes during Performance as One Explanation

Emily Lyons[1], Almaz Mesghina[2], and Lindsey E. Richland[3]

ABSTRACT— Gender gaps in mathematics achievement persist in many contexts and when visible, these gaps are paradoxical. Low-stakes measures of mathematics achievement such as grades and study behaviors favor girls, while gaps tend to reverse on assessments/competitions. We explore whether different impacts of raising performance stakes could be one explanation. Study 1 experimentally manipulated the stakes by imposing a performance-contingent, social-evaluative pressure either: before instruction ($n = 66$), before testing ($n = 61$), or none ($n = 54$). Pressure, particularly when experienced during instruction, reduced learning among girls. In contrast, boys trended toward enhanced learning under pressure. In the absence of pressure, girls exhibited strikingly larger gains in learning. Study 2 drew from a larger dataset ($n = 386$) to interrogate whether girls' superior learning in the no-pressure context might simply be an artifact of differences in prior knowledge, cognitive resources, or demographic variables, but the effect replicated and was not explained by these factors.

Men and women do not differ meaningfully in mathematics aptitude (see Spelke, 2005). Yet, mathematics achievement trajectories continue to be patterned by gender in complex and seemingly paradoxical ways that ultimately result in mathematics-intensive fields remaining heavily dominated by men (World Economic Forum, 2017). At the same time, gender differences in mathematics are not unidirectional achievement gaps. In elementary and middle school, girls tend to outperform boys on some measures of mathematics achievement, often those reflecting everyday mathematical knowledge and practices (e.g., grades, study behaviors; Cimpian, Lubienski, Timmer, Makowski, & Miller, 2016). The gender gap reverses on high-stakes mathematics assessments and individual and team-based competitions, such that when gaps are apparent, they favor boys (Ceci & Williams, 2010; Ellison & Swanson, 2018; OECD, 2013; though see Reardon, Fahle, Kalogrides, Podolsky, & Zarate, 2019). Gender gaps are especially pronounced at high levels of achievement (Ellison & Swanson, 2018). Strikingly, between 2008 and 2017 the female composition of the International Mathematics Olympiad teams ranged from 0% to a high of 14% (Steegh et al., 2019).

Here we explore the hypothesis that gender gaps may be shaped less by differences in mathematics ability, and more by contextual factors (also see Hyde & Mertz, 2009). Specifically, across two studies with fifth and sixth grade students (10–12-year-olds), we examine how one contextual factor, the stakes of the learning and performance context, may contribute to disparate gender gaps in mathematics achievement. We focus in particular on the role of *social-evaluative pressure* (see Beilock, Kulp, Holt, & Carr, 2004) in shaping mathematics performance through introducing a situation in which a full class reward is contingent on individual students' performance.

[1]Department of Education, Springfield College, USA
[2]Department of Psychology, Northwestern University, USA
[3]School of Education, University of California Irvine, USA

Address correspondence to Emily Lyons, 263 Alden Street, Springfield, MA, 01109; e-mail: elyons3@springfieldcollege.edu

On the one hand, making a desired reward contingent on performance could improve learning by motivating students to put forth increased effort (e.g., Bettinger, 2012). On the other hand, it could harm learning if anxious ideation and intrusive thoughts interfere with students' ability to concentrate on the lesson (e.g., Beilock, 2008; Schmader, Johns, & Forbes, 2008).

Gender and Mathematics Achievement Patterns

Most meta-analyses assessing mathematics gender gaps report negligible differences between boys' and girls' achievement in elementary and middle school (Halpern, 1986; Hyde, Fennema, & Lamon, 1990; Hyde, Lindberg, Linn, Ellis, & Williams, 2008; Lindberg, Hyde, & Peterson, 2011). At the same time, more recent work finds that, among high achievers, gender gaps in standardized mathematics test performance favoring boys start as early as Kindergarten and these gaps widen and spread throughout the achievement distribution as children advance through elementary school (Cimpian et al., 2016; Robinson & Lubienski, 2011): a gap equivalent to girls falling 2.5 months behind boys by fifth grade (Fryer & Levitt, 2010). Gender gaps favoring boys also arise in standardized national and international assessments of higher-level mathematics skill (College Board, 2015; Guiso, Monte, Sapienza, & Zingales, 2008; Lubienski, Robinson, Crane, & Ganley, 2013) and mathematics competitions (Ellison & Swanson, 2018).

Though gender gaps in test performance often favor boys, measures of mathematics achievement requiring sustained effort in low-stakes, everyday settings tend to favor girls. On average, middle and high school girls in the United States earn higher grades in mathematics classes (Duckworth & Seligman, 2006; Easton, Johnson, & Sartain, 2017; Kenney-Benson, Pomerantz, Ryan, & Patrick, 2006), invest more time toward mathematics homework (Gershenson & Holt, 2015; Guiso et al., 2008), and enroll in advanced mathematics courses at similar rates (Goldin, Katz, & Kuziemko, 2006; NSF, 2008) compared to boys. Grades and homework evaluate additional factors beyond assessments, including students' ability to self-regulate and their teachers' subjective evaluations, which often favor girls (Guez, Peyre, & Ramus, 2020). Compared to standardized assessments, grades are better predictors of success in upper-level and post-secondary courses, including mathematics (Allensworth & Clark, 2020; Maruyama, 2012). Additionally, girls' higher grades are unlikely due to teachers' bias in evaluation: In fact, holding achievement constant, teachers consistently rate girls' mathematics proficiency lower than that of boys (Cimpian et al., 2016).

In low-stakes mathematics contexts, girls typically outperform boys, but in high-stakes assessments of mathematical achievement, boys typically outperform girls, suggesting the context in which learning occurs and is evaluated may be an important factor shaping mathematics gender gaps. Below, we review literature that has manipulated the stakes of learning and performance contexts, revealing both positive and negative effects of high-stakes assessments for student outcomes. Critically, the direction of said effects depends on student gender.

Raising the Stakes Can Incentivize Performance

Research, predominantly in the field of behavioral economics, has shown the incentivizing potential of raising the stakes with performance-contingent rewards. Field experiments reveal that rewarding elementary students with money or prizes for correct responses on standardized exams boosts their performance (e.g., Bettinger, 2012), with larger effects for boys (Levitt, List, Neckerman, & Sadoff, 2016). Studies have also suggested that men prefer and, when given the option, chose higher-stakes performance contexts (particularly on tasks deemed male-typical, e.g., math; Niederle & Vesterlund, 2007, 2011).

Raising the Stakes Can Threaten Performance

While the possibility of being rewarded (or sanctioned) for performance can be incentivizing, it can also be distracting. Raising the stakes can in some cases harm performance by generating intrusive thoughts and worries that consume limited cognitive resources that are necessary for math engagement and performance (Beilock, 2008; Schmader et al., 2008). In everyday mathematics learning and performance contexts, the stakes may already be heightened for girls, as they contend with negative stereotypes about women and math (Ambady, Shih, Kim, & Pittinsky, 2001; Galdi, Cadinu, & Tomasetto, 2014) and experience higher math anxiety (Devine, Fawcett, Szucs, & Dowker, 2012; Hill et al., 2016) and lower confidence (Goetz, Bieg, Lüdtke, Pekrun, & Hall, 2013) than boys, despite equivalent ability. If girls are already concerned about their performance, potentially experiencing the everyday mathematics classroom as high-stakes, then raising the stakes further may be experienced as threatening. This could harm girls' learning and performance by depleting cognitive resources necessary to succeed.

Impacts of Raising the Stakes: Gender Matters

Critically, research in other male-typical fields (here, science) that has experimentally heightened the stakes while learning reveal differential effects by gender. For example, Souchal et al. (2014) found that girls saw larger learning gains from a science lesson than boys when they were told they would not be evaluated on their performance (low-stakes). The gender gap reversed when students were informed

that their performance would be graded (high-stakes). Cotner and Ballen (2017) similarly found that undergraduate women underperformed relative to men on biology exams, but not on other, lower-stakes graded activities (see also Ballen, Salehi, & Cotner, 2017; Guez et al., 2020).

Present Studies

The gendered patterns described above suggest that the stakes of the classroom learning and performance context should be more meaningfully considered when analyzing gender gaps in mathematics. Most of the studies to date examine impacts of pressure experienced solely while testing—yet relatively less is known about effects of pressure when experienced during *learning*—particularly mathematics learning. Impacts of raising the stakes prior to learning could compound over time and contribute to subsequent patterns of student engagement and retention.

We report on findings from two studies examining fifth and sixth grade students' mathematics achievement across high and low-stakes learning and performance contexts, where the instruction itself is held constant through video-based delivery. With this design, we can examine gender gaps in learning itself, versus solely final performance measures, allowing us to examine how any observed gaps may have unfolded.

Students in these studies were drawn from 21 different private, charter, and traditional public schools in the greater Chicago area. As is increasingly reflective of the nation's student body, the majority of students who participated in these studies were non-white. Although much prior research on gender gaps in mathematics has focused on majority white populations (for discussion, see Leyva, 2017), understanding the role of gender in shaping mathematics trajectories among diverse populations is of critical importance given issues of intersectionality (see Velez & Spencer, 2018) and the potential for gender and race in particular to intersect in complex ways to shape children's mathematics learning experiences (e.g., Brown & Leaper, 2010; McGee & Bentley, 2017; Young, Young, & Capraro, 2017).

In the present studies, we assess students' learning from one single high-quality lesson on ratio. In Study 1, we test how raising the stakes with pressure influences boys' and girls' learning from the video lesson, showing gender gaps in mathematics learning favoring girls only in the low-stakes, unpressured learning context. We extend this finding in Study 2, where we pooled data from seven prior studies in which our research team has used this same ratio video lesson. We meta-analyzed the data from all students who were assigned to a low-stakes, nonexperimental condition. We again found a gender gap favoring girls' learning from the same lesson, replicating our findings from the no pressure condition in Study 1.

STUDY 1

In Study 1, we experimentally manipulated whether students' learning and performance contexts were high versus low-stakes by imposing a performance-contingent, social-evaluative pressure (Beilock et al., 2004) prior to or immediately following students viewing the lesson.

How does raising the stakes impact boys' and girls' learning? On the one hand, we predicted that introducing a performance-contingent, social-evaluative pressure could boost motivation and effort, resulting in improved learning and performance. On the other hand, raising the stakes could also result in anxious ideation and intrusive thoughts that interfere with learning and performance. We tested whether average impacts of the pressure manipulation on learning and performance differed between boys and girls. As prior research suggests a larger performance boost in response to incentives (Levitt et al., 2016) and high-stakes testing contexts (Schlosser, Neeman, & Attali, 2019) among boys and men, we believed manipulating the stakes with pressure could partially explain gender gaps in mathematics achievement. Moreover, we further consider how raising the stakes shapes boys' and girls' mathematics engagement and learning, as opposed to just performance.

Method

Participants

A total of 205 fifth grade students from five schools around Chicago participated. Twenty-seven students who were absent on one or more study days were excluded due to missing data, leaving 178 students. Demographic information is provided in Table 1.

Procedure

Procedures were group administered during three visits to each classroom over a 2-week period, as outlined in Figure 1. At session 1, students completed a pretest assessing their initial understanding of ratio. Two to three days later, at session 2, students viewed a previously recorded, conceptually challenging mathematics lesson on individual laptops, immediately followed by a posttest. A week later, at session 3, students completed a final posttest and a demographics questionnaire. Students completed all tasks alongside their peers in their everyday mathematics classes.

Pressure Manipulation

Prior to session 2, students were randomly assigned within each classroom to receive a performance-contingent, social-evaluative pressure manipulation prior to learning (LP condition), prior to testing (TP condition), or not at all (no pressure [NP] condition). To enable within-classroom condition assignment, the social-evaluative pressure

Table 1
Study 1 Participants' Self-Reported Demographics by School

	School 1 traditional public		*School 2 catholic*		*School 3 catholic*		*School 4 charter*		*School 5 private*		
	Girl	*Boy*	*Girl*	*Boy*	*Girl*	*Boy*	*Girl*	*Boy*	*Girl*	*Boy*	*Total*
White	1	0	5	7	0	1	3	2	16	9	44
Black	15	20	0	0	2	6	0	0	5	5	53
Latinx	0	2	0	0	4	1	17	21	0	0	45
Biracial	5	1	0	1	2	2	7	6	8	4	36
Total	21	23	5	8	8	10	27	29	29	18	178

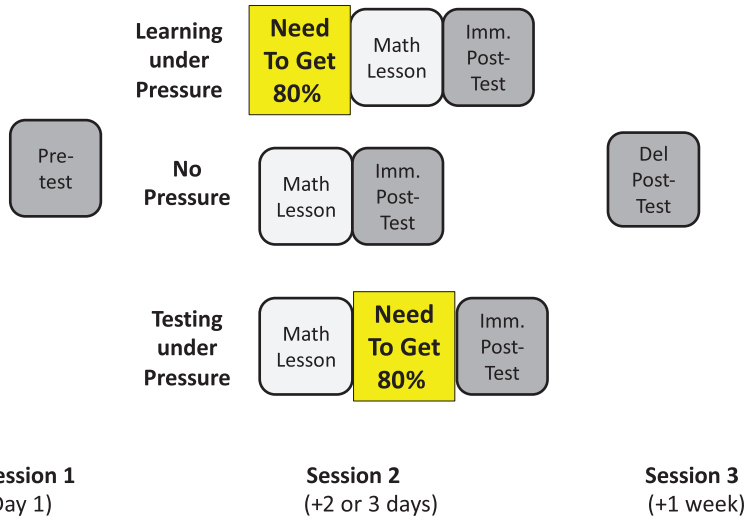

Fig. 1. Overview of experimental procedure, time course, and conditions.

manipulation was delivered via laptop, either at the start (LP) or end (TP) of the lesson.

We modeled our pressure manipulation after Beilock et al. (2004), who 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. In our study, students in the high-stakes, pressured conditions (LP or TP) 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. Conversely, students in the low-stakes, no added pressure (NP) condition were told the aim of the study was to better understand how students learn mathematics and that they would be asked to solve some problems after the lesson. All prompts were made visible and narrated on the laptop. See Appendix S1 for full condition manipulations.

Mathematics Lesson

Students watched a 30-min recording of classroom instruction during which a teacher introduces ratio to a class of fifth grade students. The teacher compares a correct strategy (least common multiple) and an incorrect strategy (subtraction, a common misconception) to solve ratio problems, both generated first and described by students in the videotaped classroom. This is a recommended teaching strategy (Common Core State Standards Initiative, 2010), but requires considerable higher-order thinking on the part of the learners (Richland et al., 2016), making this lesson a challenging learning opportunity. The video lesson was segmented with interactive questions for participants to complete during learning. All students received the same instruction while remaining seated in their typical mathematics classrooms, affording much experimental control.

Assessment

The same assessment was administered at pretest and both posttests with counterbalanced item orders. The assessment was designed to assess students' conceptual understanding of ratio in addition to their procedural knowledge and included a mix of open-ended response items and multiple-choice questions (Begolli & Richland, 2016; see

Appendix S1 for sample items). Importantly, the assessments included items that not only measured students' ability to solve problem types introduced during the video instruction, but also their ability to transfer their understanding of ratio to untaught problem types and contexts. We obtained two scores for each test: accuracy, which was the proportion of problems answered correctly or solved using a valid strategy (out of 15), and misconception use, which was the proportion of problems students attempted to solve using subtraction (out of 5). Our four outcomes of interest were immediate and sustained gains in accuracy and misconception use, which were calculated by subtracting the pretest score from each posttest score.

Measures

Student Cognitive Resources. Student cognitive resources 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.

Student Engagement. We assessed student mathematics motivation and engagement using both a behavioral (optional math puzzles) and self-report measure (situational interest and exploration intention).

Optional Math Activity: After completing the immediate posttest, students were given the option to engage with four math puzzles. The math puzzles were selected because they were designed to be fun and engaging: They involved detecting patterns and completing sequences. Students were informed they could choose whether or not to complete the puzzles. The students' other option was to independently read quietly.

Math Enjoyment and Exploration: Using a 5-point Likert scale, students rated the extent to which they found the lesson enjoyable (two items) and the extent to which they were motivated to learn more (two items) (Chen, Darst, & Pangrazi, 2001).

Results

Analytical Plan

We first describe pretest performance between the three conditions and confirm successful random assignment.

In our focal analyses, we first examine *predictors of student learning in the high (TP and LP) versus low-stakes (NP) experimental conditions* using a series of regression analyses. In the analyses, pretest performance and students' school were entered first as control variables (Step 1). Next, student gender, race and ethnicity (largest group, African American, as reference), and cognitive resources (range: 132–444) were added to the model (Step 2). Mirroring broad-scale patterns of achievement, we found that girls had larger learning gains in the low-stakes leaning context, while the gender gap disappeared, and showed possible trends toward reversing, in the high-stakes conditions.

To examine this effect of gender and learning context more precisely, we next tested interactions between gender and each of the high-stakes social-evaluative pressure manipulations (LP and TP). Analyses revealed a significant interaction between gender and pressure condition, thus we subsequently examined impacts of pressure separately for boys and girls.

Finally, we describe gender differences in student engagement in the high-stakes (LP and TP) versus low-stakes (NP) experimental conditions.

We use regression analyses to test all focal analyses and report standardized beta coefficients and standard errors.

Pretest Performance

Pretest performance differs neither between the three conditions nor between boys and girls (all $ps > .24$; see Table 2 for means). Pretest performance was not predicted by student cognitive resources or race and ethnicity, but did differ between schools. An indicator variable for school (largest school, School 4, as reference category), along with pretest performance were included as controls in all analyses of learning.

Predictors of Student Learning in the High Versus Low-Stakes Conditions

Low-Stakes Condition. Among the 54 students (28 Girls) in the NP condition, girls exhibited significantly larger

Table 2
Mean Pretest Performance, Learning Gains, Among Boys and Girls Assigned to the No Pressure (NP), Learning Pressure (LP), and Testing Pressure (TP) Conditions in Study 1

	Pretest score (SE)	*Immediate gains (SE)*	*Sustained gains (SE)*
NP			
Boys ($n = 26$)	0.25 (0.06)	0.13 (0.04)	0.12 (0.04)
Girls ($n = 28$)	0.24 (0.05)	0.30 (0.07)	0.24 (0.06)
LP			
Boys ($n = 33$)	0.21 (0.05)	0.16 (0.04)	0.18 (0.05)
Girls ($n = 30$)	0.21 (0.04)	0.22 (0.04)	0.18 (0.04)
TP			
Boys ($n = 30$)	0.20 (0.05)	0.23 (0.06)	0.22 (0.06)
Girls ($n = 31$)	0.28 (0.05)	0.24 (0.05)	0.20 (0.05)

Table 3
Predictors of Student Learning in the No Pressure (NP) Condition

	Immediate gains		*Sustained gains*	
Predictor	ΔR^2	β	ΔR^2	β
Step 1	0.25*		0.17*	
Pretest		−0.45**		−0.38**
School 1		−0.20		−0.18
School 2		−0.20		−0.20
School 3		−0.25^		−0.11
School 5		0.05		0.04
Step 2	0.16*		0.18*	
Girl		0.29*		0.21^
Cognitive Resources		0.10		0.21
White		0.42		0.41
Latinx		−0.08		0.03
Biracial		−0.06		−0.02
Total R^2	0.41		0.40	

^$p < .10$.
*$p < .05$.
**$p < .01$.

immediate mathematics gains ($\beta = 0.29$, $p = .03$) and trended toward greater sustained gains ($\beta = 0.21$, $p = .09$). Gender was the only student characteristic that predicted either immediate or sustained mathematics gains among NP students. Neither student cognitive resources nor race and ethnicity predicted gains following the lesson, although it is possible the model may have been underpowered to detect these relations. Results of the full regression analysis are shown in Table 3.

High-Stakes Conditions. All gender gaps in performance disappeared in the high-stakes learning contexts. Gender did not predict learning gains among students in either LP or TP (all ps > .45).

Gender Differences in Impacts of Pressure Manipulations

Pressure before Learning (LP). We next examine the relation of student gender and the high-stakes LP and TP manipulations separately, this time testing the interaction between gender and condition. Gender marginally interacted with LP to predict sustained learning gains ($\beta = -0.32$, $p = .05$) and to a lesser extent, immediate learning gains ($\beta = -0.27$, $p = .07$). See Table 4 for full regression analysis results. To understand these possible interactions, we separately examined the main effect of LP among boys and girls. For girls, LP predicted smaller immediate ($\beta = -0.26$, $p = .04$) and sustained ($\beta = -0.29$, $p = .03$) learning gains (Table 5, Figure 2), suggesting that pressure before instruction served more as a distracting threat than as a motivating incentive. Conversely, LP did not harm boys' learning. Instead, LP boys had numerically larger learning gains as compared to NP boys, although these differences were not statistically significant (Table 5, Figure 2).

Table 4
Regression Analysis Showing Main Effects and Interactions of the Learning Pressure (LP; Top) and Testing Pressure (TP; Bottom) Manipulations and Student Gender in Predicting Immediate and Sustained Learning Gains

	Immediate gains		*Sustained gains*	
Predictor	ΔR^2	β	ΔR^2	β
Learning pressure (LP)				
Step 1	0.21***		0.14*	
Pretest		−0.42***		−0.35***
School 1		−0.21*		−0.16
School 2		−0.17^		−0.11
School 3		−0.23*		−0.13
School 5		0.11		0.13
Step 2	0.03		0.01	
Girl		0.16^		0.07
LP		−0.07		−0.03
Step 3	0.02^		0.03*	
LP × Girl		−0.27^		−0.32*
Total R^2	0.27		0.17	
Testing pressure (TP)				
Step 1	0.21***		0.16**	
Pretest		−0.44**		−0.38***
School 1		−0.23*		−0.16
School 2		−0.09		−0.03
School 3		−0.19		−0.15
School 5		0.05		0.10
Step 2	0.03		0.01	
Girl		0.17*		0.22^
TP		0.04		0.16
Step 3	0.12		0.01	
TP × Girl		−0.22		−0.20
Total R^2	0.25		0.19	

^$p < .10$.
*$p < .05$.
**$p < .01$.
***$p < .001$.

Pressure before Testing (TP). We next examined main effects and interactions of TP and gender. In contrast to the LP findings, the analysis indicated no main effect of TP or interactions with gender (Table 4). Notably, however, TP girls had numerically smaller learning gains than did NP girls, while among boys the reverse pattern was observed (Figure 3).

Gender Gaps in Engagement in the High Versus Low-Stakes Conditions

Lastly, we examined students' mathematics motivation and engagement across the three conditions. Results of the full regression are shown in Table 6. Mirroring findings for learning, in the NP condition, girls exhibited higher engagement as compared to boys. Specifically, NP girls attempted more optional mathematics puzzles ($\beta = 1.63$, $SE = 0.69$, $p = .02$) and completed a greater number of these puzzles successfully ($\beta = 1.96$, $SE = 0.61$, $p = .002$). Additionally, NP girls

Table 5
Regression Analysis Showing Impacts of Learning Pressure (LP) Among Girls and Boys Separately

	Immediate Gains				*Sustained Gains*			
	Girls		*Boys*		*Girls*		*Boys*	
Predictor	ΔR^2	β	ΔR^2	β	ΔR^2	β	ΔR^2	β
Step 1	0.28**		0.19*		0.20*		0.15	
Pretest		−0.45**		−0.39		−0.33*		−0.34*
School 1		−0.34*		−0.05		−0.28^		−0.02
School 2		−0.1		−0.15		−0.14		−0.06
School 3		−0.17		−0.28*		0.01		−0.27
School 5		0.05		0.16		0.14		0.1
Step 2	0.06*		0.01		0.07*		0.02	
LP		−0.26*		0.07		−0.29*		0.13
Total R^2	0.34		0.2		0.28		0.17	

^$p < .10$.
*$p < .05$.
**$p < .01$.

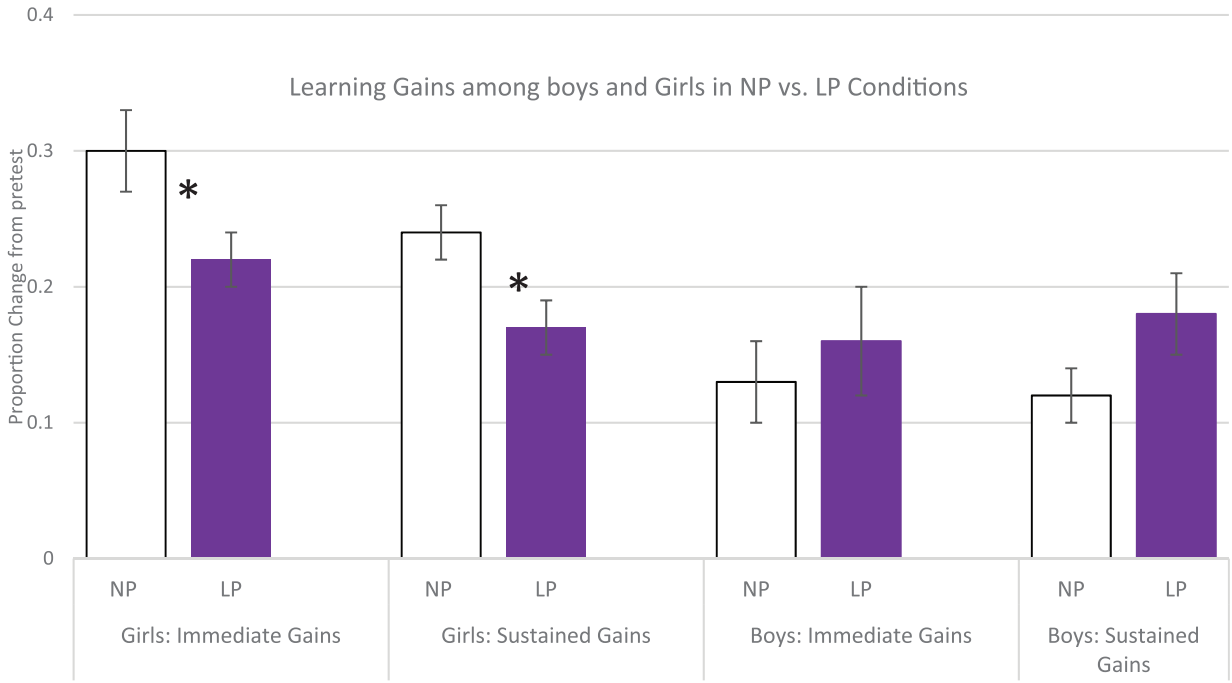

Fig. 2. Differences in boys' and girls' learning gains between the no pressure (NP) and learning pressure (LP) conditions. Error bars are ±1 *SE*. *$p < .05$.

trended toward reporting greater exploration intention and numerically reported greater enjoyment.

Conversely, in the LP and TP conditions, the gender gap in engagement not only disappeared, but in some cases reversed, mirroring findings on learning outcomes. LP boys completed more optional mathematics puzzles successfully ($\beta = -1.11$, $SE = 0.55$, $p = .04$) and reported numerically greater enjoyment and exploration intention than LP girls. TP boys reported higher enjoyment than TP girls ($\beta = -0.58$, $SE = 0.27$, $p = .03$).

Discussion

Our findings lend evidence to suggest that gender differences in responses to high and low-stakes performance contexts may explain the seemingly paradoxical ways in which mathematics achievement remains patterned by gender. Among students in the low-stakes, NP condition, girls not only learned more, but also exhibited higher engagement outcomes. In contrast, gender gaps in learning disappeared in the high-stakes conditions, with boys and girls showing similar learning gains when pressure was imposed either before

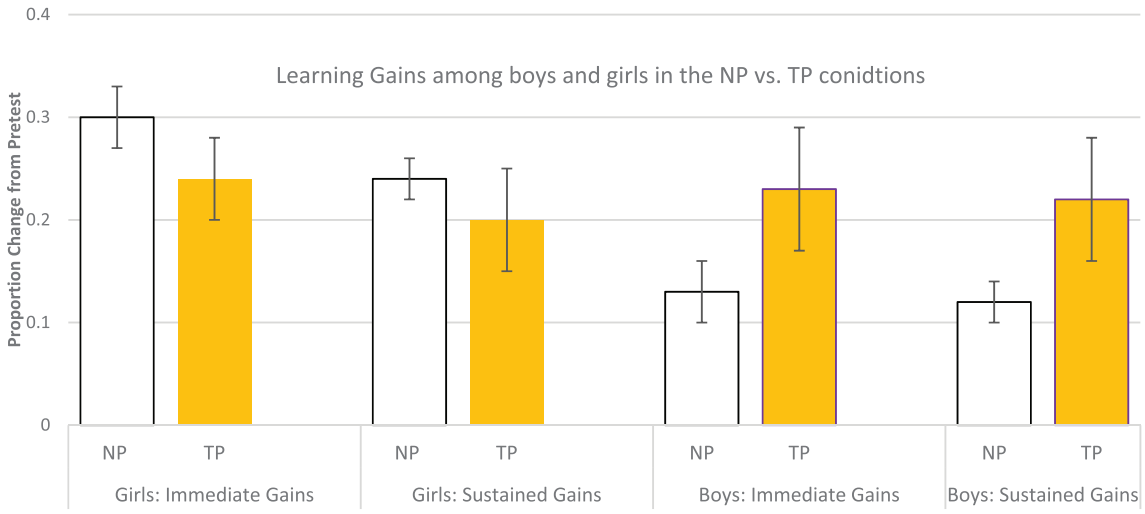

Fig. 3. Differences in boys' and girls' learning gains between the no pressure (NP) and testing pressure (TP) conditions. Error bars are ±1 *SE*. $^{*}p < .05$.

Table 6
Single Regression Analysis Showing Relations Between Gender and Engagement Outcomes in the No Pressure (NP; Top), Learning Pressure (LP; Middle), and Testing Pressure (TP; Bottom) Conditions

	Exploration intention			*Enjoyment*			*Puzzles attempted*			*Puzzles correct*		
	R^2	*β* (SE)	t	R^2	*β* (SE)	t	R^2	*β* (SE)	t	R^2	*β* (SE)	t
NP Gender	0.05	0.47^ (0.29)	1.66	0.02	0.30 (0.27)	1.10	0.10	1.63* (0.69)	2.38	0.17	1.96** (0.61)	3.21
LP Gender	0.001	−0.08 (0.28)	−0.27	0.01	−0.27 (0.29)	−0.92	0.04	−1.05 (0.69)	−1.52	0.06	−1.11* (0.55)	−2.03
TP Gender	0.04	−0.46^ (0.28)	−0.21	0.07	−0.58* (0.27)	−2.17	0.007	0.46 (0.73)	0.73	0.07	1.30 (0.61)	2.12

^$p < .10$.
*$p < .05$.
**$p < .01$.

learning or before testing. With the introduction of pressure, gender gaps in engagement actually reversed, where boys now exhibited higher engagement outcomes.

The disappearance of gender gaps across the high versus low-stakes contexts raises the question as to whether this was due to pressure facilitating boys' mathematics learning and engagement, or whether it harmed girls. The clearest answer from this experiment is that raising the stakes with pressure during learning was harmful for girls on average. Compared to girls in the NP condition, girls who experienced pressure while learning had significantly smaller learning gains immediately following the lesson and these differences persisted 1 week later, even after the threat of pressure was removed. Corroborating these learning differences, girls who experienced pressure while testing were less likely to attempt and complete optional mathematics activities than girls in the NP condition. However, data patterns suggest that the disappearance or reversal of the gender gap in the high-stakes contexts may also be partially due to pressure boosting boys' learning and engagement outcomes.

For girls, raising the stakes—particularly while learning—predicted lower learning, performance, interest, and motivation. Perhaps girls were already engaging effortfully with the lesson and assessments when the stakes were low. Raising the stakes with pressure may have generated intrusive thoughts and worries that interfered with cognitive engagement (see Beilock, 2008) while having limited incentivizing effects as girls were already engaged. This may have been the case particularly if girls were more math anxious or if an increased emphasis on performance invoked concerns that they would be judged stereotypically (e.g., Duckworth & Seligman, 2006; Ellison & Swanson, 2018). In contrast, raising the stakes appears to have facilitated boys' performance, interest, and motivation up to girls' baseline in the low-stakes context. Raising the stakes may have motivated boys, who otherwise may have been more likely to disengage during low-stakes instruction or testing (e.g., Easton et al., 2017; Gershenson & Holt, 2015), to exert more effort. Future research that probes precisely what aspects of the pressure context interfered with girls' learning and engagement while supporting those of boys could elucidate

Table 7
Descriptive Statistics *M* (*SD*) and Regression Analyses for Boys' and Girls' Learning Measures in Study 2

	*Girls (*n = *201)*	*Boys (*n = *185)*	*β (*SE*)*	t	*Sig.*
Pretest score					
Accuracy	2.40 (3.12)	2.42 (3.23)	0.01 (0.10)	0.11	.92
Misconception use	1.40 (1.56)	1.37 (1.54)	0.07 (0.10)	0.73	.47
Immediate gains					
Accuracy	3.82 (4.36)	2.52 (3.82)	0.36 (0.10)	3.64	<.001
Misconception use	−0.17 (2.01)	0.15 (1.83)	−0.19 (0.10)	−1.91	.06
Sustained gains					
Accuracy	3.17 (4.36)	1.51 (3.41)	0.49 (0.10)	4.81	<.001
Misconception use	−0.18 (1.89)	0.22 (1.72)	−0.24 (0.10)	−2.32	.02

Note. School fixed effects are included in regression analyses.

methods to raise mathematics achievement for all students across high and low-stakes settings.

STUDY 2

Study 2 aimed to address the relatively low power we held to examine the gender gaps in our data, and to explore whether the girls' higher achievement identified in our studies could be attributed to explanatory variables we were not able to include in our Study 1 analysis. We have taken advantage of the fact that we have tested this lesson in seven additional studies (four of which were previously published: Lyons, Simms, Begolli & Richland, 2017; Mesghina & Richland, 2020; Begolli, Richland, Jaeggi, Lyons, Klostermann & Matlen, 2018; Mouzaoir, Mesghina & Richland, 2020) conducted in 16 schools in the Chicago area. Here, we collapsed the data from all students in the control conditions of these studies. Students in these conditions learned and tested under low-stakes conditions—they were informed that their performance would not be reported to anyone. Students were told to try their best on the lesson and test. These allowed us to test and characterize the presence of the gender gap, and include additional tests of potentially explanatory variables to determine whether other affective, cognitive, or motivational factors could explain the identified relative larger learning gains by girls.

Method

Participants

Fifth and sixth grade classrooms at 16 private, public, and charter schools around Chicago were invited to participate in the studies analyzed here. Students who were absent on one or more study days were excluded from analyses, yielding 386 participants (52% girls). Where applicable, we used pairwise deletion to handle other sources of missing data (e.g., a tardy student). Students' self-reported race and ethnicity represented the Chicago area (48% black, 18% Latinx, 21% white, 13% biracial/other).

Because each study assessed different affective, cognitive, and motivational elements of mathematics learning, we compiled these measures in a supplemental analysis to test determinants of the gender gap: None of these factors accounted for the gendered learning gap (see Appendix S1 for analyses and details for each measure).

Design and Procedures

Study procedures were identical to Study 1, with two exceptions. First, pressure was not manipulated. We included only students who were assigned to the nonexperimental, control conditions for their respective studies, meaning that all students learned in a context equivalent to the NP condition of Study 1. Second, students in Study 2 completed additional measures assessing affective, cognitive, and motivational elements of mathematics learning, which are described more fully in Appendix S1.

Results

See Table 7 for descriptive statistics. We use regression analyses to test gender differences in pretest performance and learning gains. We include school fixed effects (largest school as reference) as schools differed in their pretest performance. Standardized beta coefficients are reported. Boys and girls did not differ in misconception use ($\beta = 0.01$) or accuracy ($\beta = 0.01$) at pretest. However, girls had larger gains in accuracy immediately following the lesson ($\beta = 0.36$), which widened at the final posttest ($\beta = 0.49$). Girls also had greater declines in misconception use at both posttests (Immediate: $\beta = -0.19$; Sustained: $\beta = -0.24$), whereas boys on average increased in their misconception use at both time points. Student race and ethnicity did not moderate gender differences in learning, nor did any motivational, affective, or cognitive factors account for the gender gap, though we did find evidence to suggest that boys might have had a less accurate—and often inflated—perception of their performance. See Appendix S1 for more details.

Discussion

In our meta-analysis of our studies assessing students' learning from the same video ratio lesson as described in Study 1, we found a large gender gap favoring girls' learning from a conceptually rich mathematics lesson on ratio in a low-stakes context. The students included in the meta-analysis were in the nonexperimental control conditions, meaning their learning was low-stakes. Notably, this gender gap emerged from data across multiple research studies, schools, and classrooms. Though they had similar pretest performance, girls immediately learned more than boys from the lesson and sustained these gains 1 week later. Our supplemental analyses suggest no cognitive, affective, or motivational elements of the learning context could explain the gap in boys' and girls' learning. These data are consistent with a growing body of literature (e.g., Souchal et al., 2014) demonstrating that girls outperform boys in low-stakes performance contexts, and provide new evidence from a learning context.

GENERAL DISCUSSION

Across two studies, we explored gender differences in mathematics performance, testing the role of high versus low-stakes performance contexts as one possible explanation for the often-disparate patterns of gender gaps in mathematics achievement. All students watched the same video lesson on ratio and completed the same assessments. In Study 1, we conducted an experiment wherein we manipulated high and low-stakes by imposing a social-evaluative pressure either before the lesson or before the test. We found that when the stakes were raised, boys enjoyed the activity more than girls, and there were no differences in learning between genders. Yet, when learning was low-stakes, girls learned significantly more from and were more engaged during the lesson than boys. We further explored the role of gender in the low-stakes learning context in Study 2: An analysis of the 386 students assigned to the nonexperimental control conditions across multiple data collections revealed a large, consistent gender gap favoring girls across a racially representative sample.

Across both studies, and consistent with a growing literature, we found that girls outperformed boys only in the lower-stakes, NP contexts. The stakes of the performance context may be an important contributor to gender differences in mathematics performance and feelings of interest and engagement, which contribute to longer-term likelihood of a STEM career. The detrimental impacts of pressure on girls' mathematics motivation and engagement are especially important to consider in light of the female advantage in reading (see Breda & Napp, 2019). As females with high math achievement tend to have even higher reading achievement, the opportunity cost of avoiding high stakes math contexts if found unenjoyable and pursuing a non-STEM career may seem especially low (see Breda & Napp, 2019).

Raising the stakes has the potential to both incentivize and threaten mathematics performance, yet little work has considered this duality on STEM performance (though see Cotner & Ballen, 2017; Souchal et al., 2014), nor specifically in explaining gender gaps in children's mathematics learning.

We assessed students' mathematics learning across 1 week and from one brief video lesson. However, given that mathematics concepts build upon themselves, initial gender gaps in learning may compound over time, underscoring the need for intervention. Importantly, this work and broader current achievement data patterns suggest that conducting research to improve classroom teaching in mathematics, while important, may not be the key lever to improving girls' participation in STEM pipelines, since much data suggests they are already learning as much, if not more than, boys in low-stakes contexts. Instead, we suggest that these more malleable contextual factors could be powerful leverage points impacting not only gateway high-stakes performance measures but also engagement and affective relationships to mathematics.

Acknowledgments—This work was supported by the Spencer Foundation (201800012); the NAEd/Spencer Dissertation Fellowship and NSF Fellowship award (DGE-1144082) to Emily M. Lyons; and the Institute of Education Sciences, U.S. Department of Education (R305A170488 and R305B140048) to the University of Chicago. The opinions expressed are those of the authors and do not necessarily represent views of the funders. Earlier versions of this work were previously presented at the annual conferences of the Cognitive Science Society on July 29, 2020 and at the American Education Research Association on April 6, 2019.

Conflict of Interest

None.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix S1: Supporting information.

REFERENCES

Allensworth, E. A., & Clark, K. (2020). High school GPAs and ACT scores as predictors of college completion: Examining assumptions about consistency across high schools. *Educational Researcher*, *49*(3), 198–211. https://doi.org/10.3102/0013189X20902110

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. https://doi.org/10.1111/1467-9280.00371

Ballen, C. J., Salehi, S., & Cotner, S. (2017). Exams disadvantage women in introductory biology. *PLoS One*, *12*(10), e0186419. https://doi.org/10.1371/journal.pone.0186419

Begolli, K. N., Richland, L. E., Jaeggi, S. M., Lyons, E. M., Klostermann, E. C., & Matlen, B. J. (2018). Executive Function in Learning Mathematics by Comparison: Incorporating Everyday Classrooms into the Science of Learning. *Thinking & Reasoning*. https://doi.org/10.1080/13546783.2018.1429306

Begolli, K. N. & Richland, L. E. (2016). Teaching Mathematics by Comparison: Analog visibility as a double-edged sword. *Journal of Educational Psychology*, *108*(2), 194–213. https://doi.org/10.1037/edu0000056

Beilock, S. L. (2008). Math performance in stressful situations. *Current Directions in Psychological Science*, *17*, 339–343. https://doi.org/10.1111/j.1467-8721.2008.00602.x

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. https://doi.org/10.1037/0096-3445.133.4.584

Bettinger, E. P. (2012). Paying to learn: The effect of financial incentives on elementary school test scores. *Review of Economics and Statistics*, *94*(3), 686–698. https://doi.org/10.1162/rest_a_00217

Breda, T., & Napp, C. (2019). Girls' comparative advantage in reading can largely explain the gender gap in math-related fields. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(31), 15435–15440. https://doi.org/10.1073/pnas.1905779116

Brickenkamp, R., & Zillmer, E. (1998) *The d2 test of attention*. Seattle, Washington: Hogrefe & Huber.

Brown, C., & Leaper, C. (2010). Latina and European American girls' experiences with academic sexism and their self-concepts in mathematics and science during adolescence. *Sex Roles*, *63*, 860–870. https://doi.org/10.1007/s11199-010-9856-5

Ceci, S. J., & Williams, W. M. (2010). Sex differences in math-intensive fields. *Current Directions in Psychological Science*, *19*(5), 275–279. https://doi.org/10.1177/0963721410383241

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. https://doi.org/10.1348/000709901158578

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), 233285841667361. https://doi.org/10.1177/2332858416673617

College Board (2015). *SAT percentile ranks for males, females, and total group*. New York, NY: College Board.

Common Core State Standards Initiative (2010) *Common core state standards for mathematics*. Washington DC: National Governors Association Center for Best Practices and the Council of Chief State School Office.

Cotner, S., & Ballen, C. J. (2017). Can mixed assessment methods make biology classes more equitable? *PLoS One*, *12*(12), e0189610. https://doi.org/10.1371/journal.pone.0189610

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. https://doi.org/10.1186/1744-9081-8-33

Duckworth, A. L., & Seligman, M. E. P. (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. *Journal of Educational Psychology*, *98*(1), 198–208. https://doi.org/10.1037/0022-0663.98.1.198

Easton, J. Q., Johnson, E., & Sartain, L. (2017). *The predictive power of ninth-grade GPA*. Chicago, II: The University of Chicago Consortium on School Research.

Ellison, G., & Swanson, A. (2018) *Dynamics of the gender gap in high math achievement* (NBER Working Paper No. 24910). Cambridge, MA: The National Bureau of Economic Research. https://doi.org/10.3386/w24910

Fryer, R. G., & Levitt, S. D. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics*, *2*(2), 210–240. https://doi.org/10.1257/app.2.2.210

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. https://doi.org/10.1111/cdev.12128

Gershenson, S., & Holt, S. B. (2015). Gender gaps in high school students' homework time. *Educational Researcher*, *44*(8), 432–441. https://doi.org/10.3102/0013189x15616123

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. https://doi.org/10.1177/0956797613486989

Goldin, C., Katz, L. F., & Kuziemko, I. (2006). The homecoming of American college women: The reversal of the college gender gap. *Journal of Economic Perspectives*, *20*(4), 133–156. https://doi.org/10.1257/jep.20.4.133

Guez, A., Peyre, H., & Ramus, F. (2020). Sex differences in academic achievement are modulated by evaluation type. *Learning and Individual Differences*, *83–84*, 101935. https://doi.org/10.1016/j.lindif.2020.101935

Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Culture, gender, and math. *Science*, *320*(5880), 1164–1165. https://doi.org/10.1126/science.1154094

Halpern, D. F. (1986) *Sex differences in cognitive abilities*. Hillsdale, NJ: Erlbaum.

Hill, F., Mammarella, I. C., Devine, A., Caviola, S., Passolunghi, M. C., & Szucs, D. (2016). Maths anxiety in primary and secondary school students: Gender differences, developmental changes, and anxiety specificity. *Learning and Individual Differences*, *48*, 45–53. https://doi.org/10.1016/j.lindif.2016.02.006

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological*

Bulletin, *107*(2), 139–155. https://doi.org/10.1037/0033-2909.107.2.139

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A., & Williams, C. (2008). Gender similarities characterize math performance. *Science*, *321*(5888), 494–495. https://doi.org/10.1126/science.1160364

Hyde, J. S., & Mertz, J. E. (2009). Gender, culture, and mathematics performance. *Proceedings of the National Academy of Sciences*, *106*(22), 8801–8807. https://doi.org/10.1073/pnas.0901265106

Kenney-Benson, G. A., Pomerantz, E., Ryan, A., & Patrick, H. (2006). Sex differences in math performance: The role of children's approach to schoolwork. *Developmental Psychology*, *42*(1), 11–26. https://doi.org/10.1037/0012-1649.42.1.11

Levitt, S. D., List, J. A., Neckerman, S., & Sadoff, S. (2016). The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. *American Economic Journal: Economic Policy*, *8*(4), 183–219. https://doi.org/10.1257/pol.20130358

Leyva, L. (2017). Unpacking the male superiority myth and masculinization of mathematics at the intersections: A review of research on gender in mathematics education. *Journal for Research in Mathematics Education*, *48*(4), 397–433. https://doi.org/10.5951/jresematheduc.48.4.0397

Lindberg, S. M., Hyde, J. S., & Peterson, J. L. (2011). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin*, *136*(6), 1123–1135. https://doi.org/10.1037/a0021276

Lubienski, S. T., Robinson, J. P., Crane, C. C., & Ganley, C. M. (2013). Girls' and boys' mathematics achievement, affect and experiences: Findings from ECLS-K. *Journal for Research in Mathematics Education*, *44*(4), 634–645. https://doi.org/10.5951/jresematheduc.44.4.0634

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: A Multidisciplinary Journal*, *42*(2), 678–690. https://doi.org/10.1111/cogs.12558

Maruyama, G. (2012). Assessing college readiness: Should we be satisfied with ACT or other threshold scores? *Educational Researcher*, *41*(7), 252–261. https://doi.org/10.3102/0013189x12455095

McGee, E. O., & Bentley, L. (2017). The troubled success of black women in STEM. *Cognition and Instruction*, *35*(4), 265–289. https://doi.org/10.1080/07370008.2017.1355211

Mesghina, A., & Richland, L. E. (2020). Impacts of expressive writing on children's anxiety and mathematics learning: Developmental and gender variability. *Contemporary Educational Psychology*, *63*. https://doi.org/10.1016/j.cedpsych.2020.101926

Mouzaoir, S., Mesghina, A., & Richland, L. E. (2020). Academic Buoyancy in Conceptually Difficult Math Learning: Exploring Students' Self-Disclosed Emotions, Beliefs, and Perception, Special Issue, Northwestern Undergraduate Research Journal. https://doi.org/10.21985/n2-skjv-z318

National Science Foundation (2008). *Science and engineering indicators*. Alexandria, VA: National Science Foundation.

Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *Quarterly Journal of Economics*, *122*(3), 1067–1101. https://doi.org/10.1162/qjec.122.3.1067

Niederle, M., & Vesterlund, L. (2011). Gender and competition. *Annual Review of Economics*, *3*(1), 601–630. https://doi.org/10.1146/annurev-economics-111809-125122

OECD (2013) *PISA 2012 results: Ready to learn: Students' engagement, drive and self-beliefs*. (Vol. *III*). Paris, France: OECD Publishing.

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. https://doi.org/10.3102/0002831219843824

Richland, L. E., Begolli, K. N., Simms, N., Frausel, R. R., Lyons, E. A. (2016). Supporting mathematical discussions: The roles of comparison and cognitive load. *Educational Psychology Review*, *29*(1), 41–53. https://doi.org/10.1007/s10648-016-9382-2

Robinson, J. P., & Lubienski, S. T. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school. *American Educational Research Journal*, *48*(2), 268–302. https://doi.org/10.3102/0002831210372249

Schlosser, A., Neeman, Z., & Attali, Y. (2019). Differential performance in high vs. low stakes tests: Evidence from the GRE test. *The Economic Journal*, *129*(623), 2916–2948. https://doi.org/10.1093/ej/uez015

Schmader, T., Johns, M., & Forbes, C. (2008). An integrated process model of stereotype threat effects on performance. *Psychological Review*, *115*(2), 336–356. https://doi.org/10.1037/0033-295x.115.2.336

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. https://doi.org/10.1111/bjep.12012

Spelke, E. S. (2005). Sex differences in intrinsic aptitude for mathematics and science?: A critical review. *American Psychologist*, *60*(9), 950–958. https://doi.org/10.1037/0003-066x.60.9.950

Steegh, A. M., Höffler, T. N., Keller, M. M., & Parchmann, I. (2019). Gender differences in mathematics and science competitions: A systematic review. *Journal of Research in Science Teaching*, *56*(10), 1431–1460. https://doi.org/10.1002/tea.21580

Velez, G., & Spencer, M. B. (2018). Phenomenology and intersectionality: Using PVEST as a frame for adolescent identity formation amid intersecting ecological systems of inequality. *New Directions for Child and Adolescent Development*, *161*, 75–90. https://doi.org/10.1002/cad.20247

World Economic Forum (2017) *The global gender gap report*. Geneva, Switzerland: World Economic Forum.

Young, J. L., Young, J. R., & Capraro, M. M. (2017). Black girls' achievement in middle grades mathematics: How can socializing agents help. *Clearing House*, *90*(3), 70–76. https://doi.org/10.1080/00098655.2016.1270657