Social Cognition, Vol. 34, No. 3, 2016, pp. 196–216

MATH QUESTION TYPE AND STEREOTYPE THREAT: EVIDENCE FROM EDUCATIONAL SETTINGS

Lucy C. Davies
Cancer Research UK

Mark Conner
University of Leeds

Constantine Sedikides
University of Southampton

Russell R. C. Hutter
University of Leeds

The mere effort account of task performance posits that task evaluation apprehension leads to performance concerns, motivating task-takers to do well and thus potentiating a dominant response. When applied to stereotype threat (ST), mere effort posits that ST motivates test-takers to disprove a stereotype, facilitating a prepotent solve response, augmenting solve type question (e.g., equations) performance, but decreasing comparison type question (e.g., estimations) performance. We tested the mere effort account of ST. In Experiment 1, undergraduates (women) engaged in a practice statistics exam. ST did not facilitate performance on solve questions, but it debilitated performance on comparison questions. In Experiment 2, middle and high school students (girls and boys) engaged in a math practice exam. ST augmented girls' performance on solve questions and debilitated performance on comparison questions. The manipulation that produced stereotype threat in girls lifted boys' performance. This research documents mere effort ST effects in educational settings.

The stereotype that "women are poorer at mathematics" may be threatening to performance, undermining women's mathematical ability in test situations (Nosek et al., 2009; Picho, Rodriguez, & Finnie, 2013). The stereotype threat (ST)

Address correspondence to Russell R. C. Hutter, School of Psychology, University of Leeds, Leeds LS2 9JT, UK. E-mail: r.r.c.hutter@leeds.ac.uk

effect is defined as performance decrement following exposure to a negative stereotype (Steele, 1997). The effect occurs when individuals (1) expect that performance evaluation in light of a negative stereotype will occur, and (2) experience a corresponding fear that their performance will confirm the stereotype (Brodish & Devine, 2009).

ST may be one factor among others contributing to the gender gap in math performance. For example, 2013 UK A-level (Advanced) mathematical exam results revealed that 18% of men received A* grades in comparison to 14.8% of women (Adams, 2013). The stereotype that women are poorer at mathematics or men are better at mathematics can trigger female test-takers' concern that their performance may be evaluated in line with, or conform to, the negative stereotype (Shapiro & Neuberg, 2007). This concern may in turn disrupt and undermine their mathematical performance (Schmader, Johns, & Forbes, 2008). Women's math potential may be impeded in "real-world" situations that involve standardized ability tests (e.g., scholastic examinations such as Graduate Record Examinations [GREs] or employment selection tests (Eccles, Jacobs, & Harold, 1990; Hyde, Fennema, & Lamon, 1990). Alternatively, reactive effects (i.e., performance increments) have sometimes been observed when stereotypically threatening information about an outgroup is encountered. Stereotype lift is a tangible increase in performance when participants make downward comparisons with outgroups considered stereotypically poorer at the task (Chalabaev, Stone, Sarrazin, & Croizet, 2008; Walton & Cohen, 2003). Men, for example, performed better on a math test when they were made aware of a negative math stereotype about women (Walton & Cohen, 2003).

Part of the ST literature has focused on question difficulty to understand how the math-gender stereotype affects women's math performance (O'Brien & Crandall, 2003; Steele, Spencer, & Aronson, 2002). But how do ST effects interact with type of math question? In particular, does variation in how a question can be answered impact differentially on math performance under ST? Jamieson and Harkins (2007) offered an explanation based on mere effort. We build on the mere effort account (when applied to ST) to investigate how math question type may differentially affect women's math performance in response to the threatening math-gender stereotype in real educational settings.

THE MERE EFFORT ACCOUNT

The mere effort account seeks to explain the effect of evaluation on task performance. Before delineating the relevance of mere effort to ST, we will briefly describe the account. According to Jamieson and Harkins (2007), there are parallels between the evaluation-performance literature and the ST literature in terms of mediating processes. These processes include processing interference and effort withdrawal. Task evaluation apprehension causes concern with performance. When potential for evaluation exists, participants solve a greater number of simple word triads on a remote associates task (RAT), in which they must examine a set of three words and then generate an associated fourth word. Potential for evalua-

tion, however, leads to fewer difficult triads solved (Harkins, 2001). Harkins (2006) argued that simple triads prompt evaluated participants to perform well, because the answer is closely associated with the triads. Given that the answer is remotely associated with triads for difficult triads, weak activation of each triad word is required to complete the task successfully. Evaluation apprehension motivates task takers to do well, which potentiates a prepotent (i.e., dominant or most likely) response. Prepotent responding leads to the generation of words *closely associated* with the triads on the RAT, facilitating the correct answer on simple triads, but inhibiting the correct answer on difficult triads.

The mere effort account to task performance is a derivative of drive theory (Zajonc, 1965), in which dominant responses occur as a function of increased drive resulting from the (physiologically arousing) mere presence of others. Physiological arousal facilitates dominant responses and inhibits non-dominant ones (Hull, 1943). Accordingly, facilitation of a dominant response will have more corrosive performance consequences on effortful tasks than on easy tasks (Baron, 1986; Cottrell, 1972; Zajonc, 1965), because effortful tasks provide fewer opportunities to implement dominant responses. Evaluation apprehension results from drive (Cottrell, 1972). Therefore, task effort or difficulty may moderate ST. However, effort or difficulty associated with a task does not *necessarily* result in activation of dominant responses (Jamieson & Harkins, 2007). Instead, such activation occurs when triggered evaluation apprehension motivates participants. This newfound motivation directs effort towards correcting inaccurate responses. Correction, though, can only occur if perceivers recognize their response as inaccurate, have knowledge of the correct response, and are in position to implement the response (as in antisaccade tasks; Jamieson & Harkins, 2007). Correction can be more taxing for some types of problem than others. Therefore, *problem type* is a function of problem difficulty activating dominant responses. ST, in common with the potential for evaluation, raises participants' concerns about task performance (Jamieson & Harkins, 2007). Mere effort therefore qualifies as an explanatory mechanism for ST effects.

MATH QUESTION TYPE AND STEREOTYPE THREAT

The mere effort account posits that ST motivates individuals to perform well and potentiates a prepotent response. If the prepotent response is the correct approach to solve the problem, performance will be facilitated; however, if the prepotent response is the incorrect approach to solve the problem, performance will be debilitated. Specifically, prior research has documented two problem types. Solve or conventional problems require a prepotent approach of applying learned material (e.g., formulae, algorithms), whereas comparison problems activate a less conventional, non-learning approach based in logic, estimation, or intuition (Gallagher et al., 2000; Jamieson, 2009; Quinn & Spencer, 2001) (see Appendix).

In contrast to other approaches to ST (Ganley et al., 2013; Steele et al., 2002), the mere effort account proposes that type of math problem rather than difficulty level is key to understanding ST performance effects (Jamieson & Harkins, 2007, 2009;

O'Brien & Crandall, 2003). Ganley and colleagues (2013), for example, used "fairly difficult" (authors' term) mathematics assessments, informed by the ST literature (Neuville & Croizet, 2007; Nguyen & Ryan, 2008; Spencer, Steele, & Quinn, 1999; Steele, 1997), and conducted follow-up analyses only on difficult items (i.e., those with less than 50% correct). This type of analyses does not control for potential differences in how individuals engage with type of math question (solve vs. compare), which may influence performance as a function of the prepotent tendency (Jamieson & Harkins, 2007).

Jamieson (2009, Experiment 2) documented the relevance of problem type. He tested the math-gender ST by manipulating orthogonally math problem type (solve vs. comparison problems) and math problem difficulty (test average of 75% correct for easy vs. 50% correct for difficult problems). If question difficulty is a determinant of ST, then performance should be facilitated regardless of question type on an easy math test and should be debilitated on a hard test (O'Brien & Crandall, 2003). Jamieson obtained support for question type as a determinant of ST. Regardless of question difficulty, the experience of ST debilitated performance on comparison problems and facilitated performance on solve problems. That is, under ST, performance did not differ as a function of difficulty level, but instead depended on whether the prepotent response was correct or not. Similarly, Jamieson and Harkins (2009) tested math solve and comparison questions that had a mean overall accuracy of 50% for each type (comparison range = 38% to 60%, solve range = 42% to 63%). The questions were taken from the quantitative section of the GRE and were in multiple choice question format. ST effects still occurred as a function of question type when controlling for question difficulty. In light of these findings, further investigation is needed on the role of question type in exam performance.

MERE EFFORT AND QUESTION TYPE IN ECOLOGICAL SETTINGS

An inevitable facet of many contemporary work and education settings is performance evaluation. The mere effort account, with its foundations in evaluation-performance research, could offer a plausible explanation of ST in applied settings. This is because the account is founded on the premise that evaluation apprehension motivates people to do well, activating prepotent responding. Also, as we stated earlier, mere effort has successfully been tested as a mediator in the relationship between ST and performance (Jamieson & Harkins, 2007), but not in applied settings. Further, as Harkins (2006) suggested, an understanding of the processes that facilitate and inhibit performance in applied situations offers a stepping stone towards interventions aimed at improving performance. Offering additional justification for the relevance of the mere effort account in applied settings, Harkins (2006) pointed out that evaluation is an important component in several applied settings, including education.

Is type of question relevant in applied educational settings when a negative, self-referent, and performance-related stereotype becomes activated? If so, is the

mere effort account suitable in understanding ST in applied settings? In laboratory research, when a negative stereotype is associated with performance, individuals are motivated to perform well and actively set out to disprove the stereotype by implementing the prepotent response for the task (Harkins, 2006; Jamieson & Harkins, 2007; McFall, Jamieson, & Harkins, 2009). Therefore, performance is dependent on whether the prepotent response is correct or not. The prepotent response for women, when attempting math problems following ST, is to apply a solve approach. As such, question type rather than question difficulty may determine performance under ST in educational settings. Harkins (2006) made a case for the role of mere effort in applied settings, but this has not been tested in conjunction with math question type. The novel contribution of our research was to test the mere effort account of differential gender math performance following ST in a naturalistic environment.

OVERVIEW

We tested the hypothesis that, under ST, women underperform on comparison questions relative to solve questions (relative to controls) in educational settings. In particular, we conducted two experiments to test ST and question type effects in the field, relying on (1) undergraduate students' performance on a university mock statistics exam, and (2) middle and high school students' performance on a General Certificate of Secondary Education (GCSE) practice math exam.

EXPERIMENT 1

In Experiment 1, we tested the gender math stereotype and the differential role that question type might play on performance effects during a first year psychology undergraduate mock statistics exam at a UK University. The Research Skills 1 (RS1) practice statistics examination has been administered for the last five years as part of this course. The course is a core component of the psychology undergraduate curriculum, designed to educate students in statistical research methods and analyses. Tests presumed to be diagnostic of math ability produce ST effects (Huguet & Regner, 2007). Statistics forms a key part of the UK math curriculum, and is included in GCSE, Advanced Subsidiary (AS), and Advanced (A-level: formal academic qualifications taken by students ranging in age from 14 to 18 years) syllabi.

We hypothesized that women subject to ST would perform better on solve questions (where prepotent responses are correct) and worse on comparison questions (where prepotent responses are incorrect) compared to their non-threatened counterparts. To our knowledge, this is the first field experiment that investigates ST using question type in a statistics exam.

METHOD

Participants and Design. We recruited 210 women ranging in age between 18 and 21 years (M = 18.32, SD = .58) via an opportunity sample during their first year psychology RS1 practice exam. All participants identified as British Caucasian, with English as their first language. The number of available students enrolled on the RS1 module determined our sample size and data stoppage. We did not record the performance of women falling outside those criteria, and we did not record the performance of men. All participants had achieved at least a GCSE B grade Math, which constituted an entry requirement at their institution. We assigned participants to a 2 (condition: stereotype threat, no stereotype threat) x 2 (question type: comparison, solve) mixed design. The first factor was between subjects, the second within subjects.

Procedure. A female examiner placed a 40-item multiple choice question test on each desk before participants entered the examination room. The exam tests were distributed in random order, and the examiner had no influence at which desk the participants chose to sit. After being seated at individual desks, participants were verbally informed of the conditions applied (e.g., no talking or conferring). As was the customary practice in the exam, all students carried pocket calculators. Participants were allocated to the ST versus no ST conditions. They were instructed to turn over their tests and read the instructions carefully before they began. All participants were given 1.5 hours to complete as many of the multiple choice questions as possible. Upon completion, they were instructed to raise their hand, and were subsequently administered a manipulation check.

After participants had responded to demographic questions, we manipulated ST by stating (based on Keller & Dauenheimer, 2003): "In previous years in the RS1 exam we have found that women are less competent at statistics compared to men" (i.e., ST) or "In previous years in the RS1 exam we have found no differences in statistical ability across men and women" (i.e., no ST).

Dependent Measures. The dependent measures were correct scores for solve and comparison questions on the exam. The research skills practice exam consisted of 40-item pen and paper multiple choice questions. Each question had four possible answers, with each correct answer worth one mark. The main experimenter and two independent raters reviewed and categorized questions as solve, comparison, or non-categorizable, based on relevant definitions supplied by Jamieson (2009, p. 14). The pilot was conducted by a second experimenter, previously unassociated with this research. The instructions involved an example and definition of both a solve and a comparison question, each accompanied by a paragraph explaining why this was so. The interrater agreement between coder 1 and 2 was 87.5%, agreement between coder 1 and the experimenter was also 87.5%, and agreement between the main experimenter and coder 2 was 80.0%. This resulted in n = 18 solve questions and n = 8 comparison questions. Questions that could not be categorized (i.e., answerable with a combination of solve and comparison approaches) resulted in n = 14. We entered in the analyses the proportion of correct scores for solve and comparison questions due to the greater number of the former relative to the latter. Finally, we included a manipulation check (Jamieson & Harkins, 2007): "To what extent are there gender differences in performance on this task?" (1 = *no gender differences*, 11 = *gender differences*).

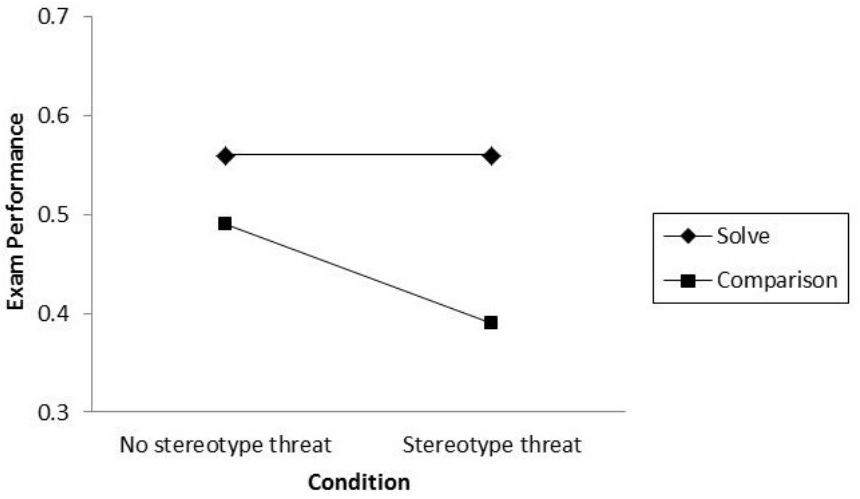

FIGURE 1. Mean proportion of problems solved on the Research Skills 1 exam as a function of condition and question type in Experiment 1.

RESULTS

Manipulation Check. We analyzed responses to the manipulation check with an independent samples t-tests. Participants in the ST condition (M = 5.45, SD = 2.76) reported that gender differences existed to a greater extent compared to those in the no ST condition (M = 3.53, SD = 2.63), $t(208) = 5.07$, $p < .001$, $d = .71$, 95% CI; 1.17, 2.67. Thus, participants in the ST condition were aware that task performance could reflect the negative group stereotype.

Exam Performance. We subjected participants' math test performance to a 2 x 2 mixed analysis of variance (ANOVA). We indexed math performance with the proportion of correct scores for solve and comparison questions (calculated by dividing participants' score for each question type by the maximum score for each question type). We obtained a significant main effect for condition. Participants performed worse in the ST conditions (95% CI; .45, .50; M = .48, SE = .01) compared to the no ST conditions (95% CI; .49, .56; M = .52, SE = .02), $F(1, 208) = 4.92$, $p = .01$, $\eta_p^2 = .02$. We also obtained a significant main effect for question type. Participants performed worse on the comparison questions (95% CI; .42, .01; M = .44, SE = .01) than on the solve questions (95% CI; .54, .58; M = .56, SE = .01), $F(1, 208) = 78.33$, $p < .001$, $\eta_p^2 = .27$.

Importantly, we also obtained a significant interaction, $F(1, 208) = 13.89$, $p < .001$, $\eta_p^2 = .06$ (Figure 1). We broke down the interaction on the basis of question type. In the case of the ST conditions, participants underperformed when answering comparison questions (M = .39, SD = .19) relative to solve questions (M = .56, SD = .16), $t(122) = 9.87$, $p < .001$, $d = 1.80$, 95% CI; .13, .19. However, a similar pattern emerged in the no ST condition: participants underperformed when answering comparison questions (M = .49, SD = .19) relative to solve questions (M = .56, SD = .18), $t(86) = 3.30$, $p = .001$, $d = .71$, 95% CI; .03, .11). We further unpacked the interaction on the basis of condition. For comparison questions, participants underperformed in the ST (M = .39, SD = .19) relative to the no ST (M = .49, SD = .19) condition, $t(208)$

= -3.66, $p < .001$, d = -.51, 95% CI; -.15, -.04). However, for solve questions, participants performed identically in the ST (M = .56, SD = .16) and no ST (M = .56, SD = .18) conditions, $t(208) = .11$, $p = .46$. Overall, these results are partially consistent with the hypothesis.

DISCUSSION

The Experiment 1 results provide initial evidence for mere effort effects (i.e., of question type following ST) under real exam conditions. ST differentially impacted women's math performance based on type of math question. In the ST condition, participants performed better on solve than comparison questions. However, unexpectedly the same pattern emerged in the no ST condition. In addition, performance on comparison questions was facilitated in the no ST relative to the ST condition, but performance for solve questions was equal across ST and no ST conditions.

The lack of facilitation for solve questions in the ST (vs. no ST) condition, when considered with performance debilitation for comparison questions in the ST (vs. no ST) condition, requires explanation. First, performance was poorer for comparison than solve questions across both ST and no ST conditions, suggesting that the comparison questions were overall harder than the solve questions. Ensuring that question types are of equal difficulty highlights a practical problem in field investigations of ST: Preexisting exams do not allow researchers to match questions for difficulty. Second, Jamieson and Harkins (2009) argued that the effect of performance debilitation on comparison problems is greater than the effect of facilitation on solve problems. This occurs because the prepotent response to solve equations (i.e., using relevant formulas and operations) is female test-takers' motivated strategy. Therefore, the solve approach is generally known and applied by female test-takers, but women experiencing ST are more motivated (relative to non-threatened women) to apply this approach for disproving the negative self-applicable stereotype. Further, we would not expect differences to exist between threatened and control participants in terms of knowledge and formulas underpinning solve questions, meaning that the effect of elevated motivation is restricted to threatened participants trying harder to apply the solve approach to as many questions as possible relative to controls. Given that time is limited in most tests, this further restricts the role of solve facilitation. Hence, for solve questions, the enhanced performance under ST is limited to women's increased effort to disprove the stereotype within the restricted examination time. In contrast, for comparison questions, performance is more strongly debilitated by ST, because solve facilitation means that applying the correct approach is unlikely for threatened women—resulting in an immediate disadvantage.

The potentiated solve approach serves to handicap test-takers' comparison mathematical ability: Threatened female test-takers must both recognize that the solve approach is wrong and adopt the correct comparison approach. In addition, the more motivated they are, the stronger the solve approach will be potentiated,

hindering comparison performance further (Jamieson & Harkins, 2009). Indeed, we obtained a significant effect of ST on exam performance, indicating that ST led female test-takers (relative to their non-threatened counterparts) to underperform on the exam. We addressed the issue of question type difficulty next by examining the interactive effects of question type and ST in a middle and high school sample.

EXPERIMENT 2

Understanding the experience of ST in a school math exam environment is a crucial component of efforts to reduce inequalities in education (Huguet & Regner, 2007; Wei, 2012; Wicherts, Dolan, & Hessen, 2005). Recent UK educational statistics indicated that, for math and additional math GCSEs in 2011 and 2012, a higher cumulative percentage of boys achieved a greater number of top grades than girls (GCSE Results, 2012). Similarly, the school in the current experiment reported gender differences between the number of A*s achieved: In 2011, 32% of boys compared to 21% of girls, and in 2012, 29% of boys compared to 14% of girls, achieved the top math grade. We asked, in Experiment 2, whether and how ST interacts with the type of math question encountered. We included not only girls, but also boys. As Stoet and Geary (2012) argued, it is useful to include a control group consisting of men (here, boys), because it allows researchers to draw clearer conclusions about how ST may contribute to gender differences in performance.

We relied on a GCSE target age sample (14–16 years of age). Therefore, we intended for the math test to consist of both solve and comparison questions (assessed within subjects) that are typical of a GCSE exam. Also, in contrast to Experiment 1, we intended for the number of solve and comparison questions to be equal. To that effect, we conducted a pilot study. We administered a combined math test consisting of 18 questions (9 solve and 9 comparison), all set at the GCSE (higher tier) level and taken from non-calculator exams selected from an online academic source (www.aqa.org.uk). Each question was worth between three and six marks (depending on difficulty and reflecting scoring in GCSE exams). Thirty young women between 18 and 23 years of age ($M = 19.67$, $SD = 1.67$) were allotted 35 minutes to complete the test. This was analogous to the time per question allocated in GCSE examinations. We controlled for mathematical ability by ensuring that all participants had a GCSE grade of C or above. All of the young women identified as British Caucasian and spoke English as their first language. To create a test comprised of two differing question types (i.e., solve vs. comparison), while of equal difficulty, we selected five questions (from the original 18), worth three marks, across each question type that elicited similar overall scores. We did so using criteria specified by Jamieson (2009). The overall total scores of the questions selected for the solve ($M = 44.40$, $SD = 13.01$) versus comparison ($M = 42.2$, $SD = 7.26$) questions did not differ in difficulty, $t(8) = .33$, $p = .75$. Therefore, the finalized versions of the new test consisted of 10 (five solve, five comparison) three-mark questions balanced for difficulty.

We hypothesized that threatened girls would perform more poorly on comparison problems and better on solve problems than non-threatened girls. According to Jamieson and Harkins (2009), comparison problem debilitation has a greater effect than solve problem facilitation. As such, we anticipated for ST to harm performance on comparison questions more than it improved performance on solve questions. Specifically, we hypothesized that the math performance of threatened girls would be worse than the performance of non-threatened girls. In regard to boys' math performance, we did not expect it to differ as a function of ST, although we thought it might result in a stereotype lift effect (Chalabaev et al., 2008; Walton & Cohen, 2003) manifested in improved performance irrespective of question type.

METHOD

Participants and Design. We tested 191 UK middle and high school students (94 girls) ranging in age between 14 and 16 years ($M = 14.79$, $SD = .56$). All had identified as British Caucasian, with English as their first language. The number of available students attending the school determined our sample size and data stoppage. We did not record students who fell outside those criteria. We used a 2 (condition: stereotype threat, no stereotype threat) x 2 (gender: boy, girl) x 2 (question type: comparison, solve) mixed design. The first two factors were between subjects, and the last factor was within subjects.

Procedure. We conducted the experiment in the school examination hall during a time-tabled math lesson period. The female examiner, accompanied by three school math teachers (all men), set the examination hall with a math test on each individual desk. The ST and no ST exam tests were distributed in random order on desks. As is typical in GCSE examination procedure, participants were invited into the exam hall and required to sit at a desk. (The experimenter had no influence on which desk the participants chose to sit.) Examination conditions were enforced verbally (i.e., no talking or conferring, only stationary permitted on desks, no calculators). Participants were instructed to read carefully the information on the front of the math test. They were allowed 35 minutes to complete as many of the 10 questions as possible. After the allotted time, participants were each given the one item pen-and-paper manipulation check to complete. The examiner then collected the tests and manipulation check, before debriefing participants as a group. Following responses to demographics items, participants in the ST condition were informed that, "Previous research has shown gender differences on this test," whereas those in the no ST condition were informed that, "Previous research has shown no gender differences on this test."

Dependent Measures. As mentioned above, the math test consisted of 10 GCSE math questions (five solve, five comparison) evenly matched for difficulty (see pilot study described above), with three marks allocated for each correct answer. Similar to Experiment 1, the main experimenter and two independent raters reviewed and categorized questions as solve or comparison. The procedure was identical to Experiment 1, with the exception that only 10 questions were reviewed. Agreement between coder 1 and 2 was 70.0%, agreement between coder 1 and the experimenter was 80.0%, and agreement between the main experimenter

and coder 2 was 70.0%. This resulted in $n = 5$ solve questions, $n = 5$ comparison questions, and $n = 0$ non-categorizable questions.

The dependent measure was math score (out of 30). The test closely resembled GCSE exam formatting, including the test cover and formula sheet. We used the same manipulation check as in Experiment 1.

RESULTS

Manipulation Check. Participants in the ST condition ($M = 5.45$, $SD = 2.94$) reported that gender differences existed on the test to a greater extent than those in the no ST condition ($M = 4.50$, $SD = 2.51$), $t(189) = 2.41$, $p = .00085$, $d = .35$, 95% CI [.17, 1.73]. We conclude that participants in the ST condition were aware of the negative stereotype.

Exam Performance.[1] We subjected math test performance to a 2 x 2 x 2 mixed ANOVA. The gender main effect was not significant, $F(1, 187) = .52$, $p = .47$. The condition main effect was marginal, $F(1, 187) = 3.13$, $p = .08$, $\eta_p^2 = .02$: Performance tended to be better in the ST condition (95% CI; 6.36, 7.60; $M = 6.98$, $SE = .31$) than in the no ST condition (95% CI; 5.67, 6.80; $M = 6.23$, $SE = .29$). The question type main effect was significant, $F(1, 187) = 5.62$, $p = .0095$, $\eta_p^2 = .03$: Participants performed better on solve questions (95% CI; 6.38, 7.42; $M = 6.90$, $SE = .26$) than on comparison questions (95% CI; 5.85, 6.80; $M = 6.30$, $SE = .23$). The results for condition and question type replicated those of Experiment 1.

1. Proportion correct. We also analyzed the proportion of problems correct and obtained a similar pattern of results as for the main analyses. To calculate proportion of problems correct, we divided participants' score for each question type by the maximum score for each question type. We subjected the data to a 2 (condition) x 2 (gender) x 2 (question type) mixed ANOVA, with repeated measures on the third factor. The gender main effect was not significant, $F(1, 187) = .52$, $p = .47$. The condition main effect was marginal, $F(1, 187) = 3.13$, $p = .08$, $\eta_p^2 = .02$: Participants tended to answer a greater proportion of questions in the ST (95% CI; 0.47, 0.51; $M = .47$, $SE = .02$) compared to the no ST (95% CI; 0.38, 0.42; $M = 0.42$, $SE = .02$) condition. The question type main effect was significant, $F(1, 187) = 5.62$, $p = .00095$, $\eta_p^2 = .03$: Participants answered a greater proportion of solve (95% CI; 0.43, 0.50; $M = 0.46$, $SE = .02$) than comparison (95% CI; 0.39, 0.45; $M = 0.42$, $SE = .15$) questions. The results for condition and question type replicated those of Experiment 1.

We also obtained a significant Question Type x Gender interaction, $F(1, 187) = 4.76$, $p = .03$, $\eta_p^2 = .03$. Girls answered a greater number of questions for solve ($M = 0.48$, $SD = 0.24$) than comparison ($M = 0.43$, $SD = 0.19$) questions, $t(93) = 2.36$, $p = .02$, $d = .51$, 95% CI; 0.009, 0.11; however, for boys, the proportion of math questions answered was unaffected by question type, $t(96) = .17$, $p = .43$. We found no significant gender differences emerged for answers to comparison questions, $t(189) = .30$, $p = .38$, but girls tended to perform better ($M = 0.48$, $SD = 0.24$) than boys ($M = 0.43$, $SD = 0.25$) on solve questions, $t(189) = -1.34$, $p = .09$, $d = .51$, 95% CI; -0.12, 0.02.

In addition, we obtained a significant Question Type x Condition interaction, $F(1, 187) = 5.28$, $p = .02$, $\eta_p^2 = .03$. In the ST condition, participants answered more solve ($M = 0.50$, $SD = 0.23$), than comparison ($M = 0.43$, $SD = 0.24$) questions, $t(86) = 2.46$, $p = .02$, $d = .53$, 95% CI; 0.01, 0.13, but, in the no ST condition, the proportion of questions answered did not differ across question type, $t(103) = .001$, $p = 0.99$, 95% CI; -.0.04, 0.04. Alternatively, participants answered more solve questions in the ST ($M = 0.50$, $SD = 0.23$) relative to the no ST ($M = 0.42$, $SD = 0.25$) condition, $t(189) = 2.39$, $p = .009$, $d = .51$, 95% CI; 0.01, 0.15, reflecting a pattern in the main analyses, but not in Experiment 1. However the proportion of comparison questions did not differ as a function of ST condition $t(189) = .45$, $p = .33$. Also, as with the main analyses, the Gender x Condition interaction was not significant, $F(1, 187) = .62$, $p = .43$.

These effects were qualified by the Question Type x Gender interaction, $F(1, 187) = 4.76$, $p = .03$, $\eta_p^2 = .03$. The performance of boys was unaffected by question type, $t(96) = .17$, $p = .43$. However, consistent with their general preference for conventionally structured questions (Gallagher et al., 2000), girls performed better on solve ($M = 7.18$, $SD = 3.65$) than on comparison ($M = 6.27$, $SD = 2.84$) questions, $t(93) = 2.36$, $p = .01$, $d = .51$, 95% CI; 0.15, 1.68. Further, the performance of girls and boys did not differ on comparison questions, $t(189) = .30$, $p = .38$, but it was marginally different on solve questions, $t(189) = -1.34$, $p = .09$, $d = .51$, 95% CI; -1.78, .34, with girls ($M = 7.18$, $SD = 3.65$) tending to perform better than boys ($M = 6.46$, $SD = 3.73$).

In accord with Experiment 1, we obtained a significant Question Type x Condition interaction, $F(1, 187) = 5.28$, $p = .02$, $\eta_p^2 = .03$. In the ST condition, participants performed worse on the comparison questions ($M = 6.45$, $SD = 3.41$) than on the solve questions ($M = 7.51$, $SD = 3.45$), $t(86) = 2.46$, $d = .53$, $p = .008$, 95% CI; .20, 1.91. In contrast, in the no ST condition, participants' performance did not differ across question type, $t(103) = .001$, $p = .990$. When viewed from the standpoint of question type, performance on the solve questions was better under the ST ($M = 7.51$, $SD = 3.45$) than the no ST ($M = 6.24$, SD 3.45) condition, $t(189) = 2.39$, $p = .009$, $d = .51$, 95% CI; .22, 2.31. The increase in solve question performance in the ST condition differs from Experiment 1, but is in line with our hypothesis. By contrast, performance on the comparison questions did not differ as a function of condition, $t(189) = .45$, $p = .33$. The Gender x Condition interaction was not significant, $F(1, 187) = .62$, $p = .43$.

The crucial triple interaction was significant, $F(1, 187) = 16.64$, $p < .001$, $\eta_p^2 = .08$. We broke it down on the basis of Condition x Question Type across gender. We first

1. (continued). The three-way interaction was significant as before, $F(1, 187) = 16.64$, $p < .001$, $\eta_p^2 = .08$. We broke it down on the basis of Condition x Question Type across gender. We examined girls' exam performance data first. A significant question type main effect showed that girls solved more questions for solve (95% CI; 0.44, 0.54; $M = 0.49$, $SE = 0.02$) than comparison (95% CI; 0.37, 0.45; $M = 0.41$, $SE = 0.02$) questions, $F(1, 92) = 10.56$, $p = .002$, $\eta_p^2 = .10$. The condition main effect was not significant, $F(1, 92) = 0.55$, $p = .46$. The Condition x Question Type interaction was significant, $F(1, 92) = 20.71$, $p < .001$, $\eta_p^2 = .18$. In the ST condition, girls solved fewer comparison ($M = 0.37$, $SD = 0.19$) than solve questions ($M = 0.56$, $SD = 0.25$), $t(39) = 4.41$, $p < .001$, $d = 1.41$, 95% CI; 0.10, 0.27. As expected, this pattern did not emerge in the no ST condition, with no significant differences for the proportion of problems solved for comparison ($M = 0.43$, $SD = 0.18$) relative to solve ($M = 0.42$, $SD = 0.23$) questions, $t(53) = -1.17$, $p = 0.25$. When viewed across question type, as expected, for comparison questions, girls underperformed on questions in the ST ($M = 0.37$, $SD = 0.19$), relative to the no ST ($M = 0.45$, $SD = 0.18$) condition, $t(92) = -2.06$, $p = .021$, $d = -.43$, 95% CI; -0.16, -0.002; however, for solve questions, girls answered more questions in the ST ($M = 0.56$, $SD = 0.25$) relative to the no ST ($M = 0.42$, $SD = 0.23$) condition, $t(92) = 2.77$, $p = .0035$, d = .58, 95% CI; -.04, 0.23. The facilitation for solve performance in the ST versus the no ST condition was in line with the main results, but contrasted to the non-significant effect in Experiment 1.

We next analyzed boys' performance. We found a significant main effect for condition, $F(1, 95) = 2.93$, $p = .045$, $\eta_p^2 = .03$, in accord with ST lift effects (Chalabaev et al., 2008; Walton & Cohen, 2003). Boys answered a greater number of questions in the ST (95% CI; 0.41, 0.53; $M = 0.47$, $SE = 0.03$) relative to the no ST (95% CI; 0.34, 0.45; $M = 0.39$, $SE = 0.03$) condition. The question type main effect, $F(1, 95) = .02$, $p = .90$, and the Condition x Question Type interaction, $F(1, 95) = 1.56$, $p = .21$, were not significant. Therefore, as with the main analyses, boys answered correctly a greater proportion of questions under ST conditions, but their answers were not affected by math question type.

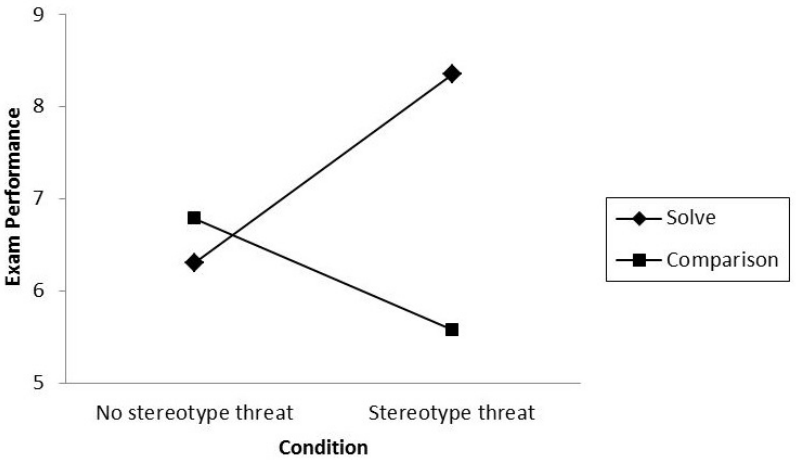

FIGURE 2. Mean girls' math score on the GCSE math exam as a function of condition and question type in Experiment 2.

examined girls' exam performance. The question type main effect was significant: Girls' performance was better on the solve (95% CI; 6.60, 8.06; M = 7.33, SE = .37) than comparison (95% CI; 5.60, 6.80; M = 6.18, SE = .29) questions, $F(1, 92) = 10.56$, $p = .001$, $\eta_p^2 = .10$. The condition main effect was not significant, in contrast to Experiment 1, $F(1, 92) = .55$, $p = .46$. Importantly, as in Experiment 1, the Condition x Question Type interaction was significant, $F(1, 92) = 20.71$, $p < .001$, $\eta_p^2 = .18$: In the ST condition, girls performed worse on comparison questions (M = 5.58, SD = 2.91) than on solve questions (M = 8.35, SD = 3.68), $t(39) = 4.41$, $p < .001$, $d = 1.41$, 95% CI; 1.50, 4.05] (Figure 2). As expected but in contrast to Experiment 1, this pattern did not emerge in the no ST condition: girls' performance on comparison (M = 6.78, SD = 2.71) relative to solve (M = 6.31, SD = 3.40) questions did not differ significantly, $t(53) = -1.17$, $p = .25$. Further, as hypothesized and as in Experiment 1, for comparison questions, girls underperformed in the ST (M = 5.58, SD = 2.91) relative to the no ST (M = 6.78, SD = 2.71) condition, $t(92) = -2.06$, $p = .021$, $d = -.43$, 95% CI; -2.36, -.04. In contrast, for solve questions, girls' performance was augmented in the ST (M = 8.35, SD = 3.68) relative to the no ST (M = 6.31, SD = 3.40) condition, $t(92) = 2.77$, $p = .0035$, $d = .58$, 95% CI; .58, 3.49. The increase in solve performance in the ST versus the no ST condition was in contrast to the non-significant effect in Experiment 1. The results are consistent with the hypothesis.

Next, we examined boys' exam performance. The condition main effect was significant, $F(1, 95) = 2.93$, $p = .045$, $\eta_p^2 = .03$. In line with stereotype lift research (Chalabaev et al., 2008; Walton & Cohen, 2003), boys' math performance was augmented in the ST condition (95% CI; 6.09, 7.89; M = 6.99, SE = .45) relative to the no ST condition (95% CI; 5.04, 6.78; M = 5.91, SE = .44). Neither the question type main effect, $F(1, 95) = .02$, $p = .90$, nor the Condition x Question Type interaction, $F(1, 95) = 1.56$, $p = .21$, was significant (in the ST condition, boys' performance was M = 7.19, SE = 0.46 on comparison questions and M = 6.79, SE = 0.53 on solve questions; in the no ST condition, boys' performance was M = 5.66, SE = 0.44 on comparison questions and M = 6.16, SE = 0.51 on solve questions.). Therefore,

boys' math performance was lifted by the math-gender ST, but was unaffected by the type of math question (solve vs. comparison) encountered.

DISCUSSION

Experiencing ST in a school GCSE practice exam resulted in differential performance depending on question type for girls. Consistent with previous mere effort account laboratory findings (Jamieson & Harkins, 2009, Experiment 1), we obtained a significant interaction among condition, question type, and gender. As hypothesized, girls' math performance under ST was dependent on type of math question encountered. Specifically, threatened girls performed better on solve questions (i.e., prepotent response correct) and worse on comparison questions (i.e., prepotent response incorrect) than non-threatened girls. However, ST did not reduce girls' math performance overall.

Experiment 2 demonstrated, for the first time, that the effect of ST is contingent upon the type of question encountered in a real exam setting. In addition, boys' math performance improved under ST, irrespective of question type. This fits with the stereotype lift literature showing tangible increases in boys' math performance when they are made aware of the negative math stereotype about women (Walton & Cohen, 2003). Boys' performance was lifted, possibly because they were able to draw comparisons with women's and girls' stereotypically poorer performance. However, we did not measure directly downward comparisons. The stereotype lift effect found here was not observed by Jamieson and Harkins (2009, Experiment 1). Real exam scenarios could intensify test diagnosticity when downward comparisons are possible, but this possibility requires empirical scrutiny. In all, performance debilitation for girls and performance facilitation for boys, following ST, simultaneously exacerbated math gender differences.

GENERAL DISCUSSION

For the first time, we observed an interactive effect of question type and ST in ecological (i.e., educational) test settings. The literature indicated that the mere effort account can explain a variety of performance/evaluation effects, but the account had not been used before under conditions of ST in applied settings. In Experiment 1, psychology undergraduate students under examination conditions completed the mock statistics exam. Similarly, in Experiment 2, middle and high school students under examination conditions completed the GCSE math test. The pattern of findings was generally consistent with the mere effort account of math-gender ST (Jamieson & Harkins, 2009, 2011). This pattern indicated that ST differentially impacts women's math performance depending on the type of math question on which they focus. Given the importance of extending the interactive effects of ST and math question type in real exam or school settings (Huguet & Regner, 2007;

Wei, 2012; Wicherts et al., 2005), the present findings make a key addition to understanding ST in education.

CONTRIBUTION

Our research demonstrates that ST harms women's math performance on comparison questions in real educational test settings. In both the undergraduate and middle/high school samples, women's and girls' performance on comparison (vs. solve) questions decreased under ST (relative to no ST). ST facilitated solve question type performance in Experiment 2 (but not in Experiment 1). The observation of mere effort effects in an applied setting activated by ST is relevant to educational practice. Our results justify Harkins' (2006) contention about the explanatory utility of the mere effort account in educational settings: We obtained performance effects in accord with that account. In particular, our research was the first to report performance effects consistent with the mere effort account, following exposure to a negative self-referent stereotype in applied settings. In the original formulation of mere effort, it is *evaluation apprehension* that motivates participants and leads to prepotent responding. When mere effort is applied to ST, the *threatening performance stereotype* raises participants concerns about task performance leading to the activation of prepotent (solve) responding (Jamieson & Harkins, 2007); therefore, it is problem type rather than difficulty that debilitates performance. This manifests in performance reduction because, when faced with comparison questions, the activated solve response is inappropriate. In all, we have shown that motivation facilitates solve question responses, but inhibits comparison question responses, following exposure to a threatening task-related stereotype in applied educational settings.

Stoet and Geary (2012) have criticized ST research for using adjusted scores (which we did not use). In their meta-analysis, only 55% of reviewed articles replicated the original Spencer and colleagues' (1999) women's math ST effects, and half of those involved adjusted scores. Further, ST debilitated women's performance on comparison questions during real test-based situations, but without the real additional threatening consequences (i.e., test performance did not count towards real academic grades). The current findings are in keeping with the idea that ST effects in educational environments are real (Keller & Dauenheimer, 2003) and that the mere effort account can help explain them.

Whereas Experiment 1 results suggested that ST negatively impacted on women's math performance overall, Experiment 2 did not find ST effects for overall math performance. That is, in Experiment 2, threatened female test-takers' solve performance was facilitated but did not protect their overall math performance, as comparison performance was reduced. The combination of solve facilitation and comparison inhibition suggests that the differential effects of math question types can cancel out or mask ST effects. Studies that do not control for question type may ignore genuine ST effects.

Despite the general consistency between laboratory findings and our field results, in Experiment 2 the ST performance facilitation of the math GCSE solve

questions was stronger in the field. This finding conflicts with Experiment 1 results, where the negative effects of comparison questions undermined performance. The finding is also at odds with Jamieson and Harkins' (2009) argument that the negative performance effects of addressing comparison questions impacts to a greater extent than the positive effects of responding on solve questions. Solve facilitation and comparison debilitation performance effects highlight the complex interplay between ST and math performance in real-world exam environments, and suggest that the mere effort account alone may not be able to explain overall math ST effects. It is possible that the younger participants in Experiment 2 were able to motivate themselves more towards improved performance than the older ones in Experiment 1. Ability is another possibility. In Experiment 1 all participants had obtained a GCSE of at least B grade, whereas in Experiment 2 participants had yet to take a GCSE math exam. Therefore, participants in Experiment 2 were likely to have been drawn from a wider ability range. Future research should aim to address whether age and ability affect motivation independently or interactively, and why so.

In Experiment 2, the math performance of secondary school boys improved under ST, regardless of math question type. This finding suggests that, in response to the math-gender stereotype, boys' math performance was lifted by the social comparison they might have made with the negative female math stereotype (Walton & Cohen, 2003). Thus, in contrast to female math performance, ST influenced male math performance positively regardless of math question type. Differences in how male and female math performance is affected by the negative math stereotype about women potentially widen the math-gender performance gap. In addition, findings from Experiments 1 and 2 converged on the notion that it is motivation to disprove the negative stereotype that causes individuals to rely on the prepotent response. As female test-takers are the social group negatively stigmatized by the math-gender stereotype and males are not, only women's prepotent response is activated by the ST. Hence, only women's math performance is influenced by whether the question can be answered using the solve response.

IMPLICATIONS

Explicit ST Manipulation. Our explicit ST manipulation might limit the applicability of ST in educational settings, given that students are not typically exposed to such a manipulation. Previous research has contested the validity of explicit ST measures. For example, Huguet and Regner (2007) criticized Keller and Dauenheimer (2003) for informing participants that the math test produced (or did not produce) gender differences. Huguet and Regner (2007) used quasi-ordinary classroom circumstances to manipulate ST by altering the gender composition of the groups of test-takers. Other research has addressed the potential influence of coed versus single sex learning environments as an ST manipulation (Kessels & Hannover, 2008; Picho & Stephens, 2012). However, in a recent meta-analysis, Picho and colleagues (2013) reported that ST was not moderated by nature of testing environment or participant sex composition: Women's performance was unaffected by test settings that were homogeneous or where they formed the majority. As

such, the implementation of an explicit ST in our research manipulation enabled a clearer picture of how people respond to ST, because the manipulation is often experienced at the individual level. For example, some of our test-takers may have read or been made aware of differences in gender math performance via the media prior to an exam. Those that were aware could have been adversely affected on performance. Our manipulation simply made all ST participants aware. Future empirical efforts should further address the issue of prior exposure to negative math stereotypes about women. The issue is not perhaps whether explicit manipulations are ecologically valid, but rather who is aware of these manipulations in the real world.

Implications for Education. The observed ST effects were dependent on the type of question answered. In Experiments 1 and 2, ST negatively affected comparison question performance, whereas in Experiment 2, ST facilitated performance on solve questions. The results collectively suggest that strategies to improve how women approach comparison questions should form the basis for interventions aimed at bolstering performance. However, given the nature of comparison questions (which are less reliant on recalled practiced formulae, for example), application of such strategies is perhaps difficult to implement. One approach may be to raise awareness of positive female math role models (Blanton, Crocker, & Miller, 2000), linking awareness of them to undertaking comparison questions. Our applied research vindicates the application of the mere effort account to ST. This replicates laboratory research (Harkins, 2006; Jamieson & Harkins, 2007; McFall et al., 2009). When negative, self-referent, threatening, and performance-related stereotypes become activated in educational settings, mere effort is an important performance moderator. That is, mere effort has real-world implications. Those experiencing ST may experience performance benefits when tests comprise proportionally more solve type than comparison type questions. Awareness of question type when constructing tests requires careful consideration, if egalitarian educational outcomes are to be met.

CONCLUSION

Our findings demonstrate the interactive effects of math question and ST, apparently driven by the motivation to disprove the negative stereotype, for the first time in field settings, that is, a practice statistics exam and a secondary school GCSE math mock exam. Further, female test-takers augmented performance on solve questions (e.g., more conventional problems including formulae and algorithms) following ST, implicating the role of heightened motivation to disprove the threatening negative female math stereotype. Also, the math-gender ST lifted boys' overall math performance. However, ST impacted negatively on performance pertaining to comparison questions performance. Therefore, ST is most damaging when test-takers encounter comparison questions in field settings. The findings are broadly supportive of mere effort underpinning ST. Interventions to improve women's and girls' math test performance might include training in the identification of comparison type questions.

APPENDIX

Jamieson (2009, Appendix A, p. 103) defines problem type using the following examples:

> "Examples of the problem types found on the quantitative GRE test. These problems appeared in the math tests used in this research.

SOLVE TYPE

If the total surface area of a cube is 24, what is the volume of the cube?
a. 8
b. 24
c. 64
d. $48\sqrt{6}$
e. 216

For this problem, the individual must apply the formula for the volume of a cube, which is: length x width x height (all of which are the same value for a cube). To get the length of a side, the individual divides 24 by 6 (there are 6 faces on a cube) to obtain the area of one face, 4. The length of one side is 2 (area = length x width). To compute volume, the test-taker then cubes 2 to get the answer, 8. Thus, solve problems involve the application and computation of equations.

COMPARISON TYPE

$n = (7)(193)$	
Column A	Column B
The number of distinct positive factors of n	10

a. The quantity in Column A is greater
b. The quantity in Column B is greater
c. The two quantities are equal
d. The relationship cannot be determined from the information given

This problem can be solved by using intuition. First, the test-taker must realize that each number presented (7, 19, 3) is a prime number. Thus, the test-taker can logically deduce that the factors of the end product can only be multiples of 7 and 19. Thus, the factors of the final product are: 7, 7*19, 7*192, 19, 192, 193, plus the final product itself (7*193) and 1. Because the goal of the problem is not to compute the value of n, but simply to determine whether the number of positive factors of n is greater than, less than, or equal to 10, all the test-taker now needs to do is to add up the number of distinct positive factors (8) to find that Column B is greater than Column A. Thus, the correct answer choice is "b," and only intuition and logic were used. No calculations were necessary."

REFERENCES

Adams, R. (2013, August 15). A-level results 2013: fewer students get top grades for second year running. *The Guardian*. Retrieved from http://www.theguardian.com/education/2013/aug/15/a-level-results-2013-students-top-grades.

Baron, R. S. (1986). Distraction-conflict theory: Progress and problems. *Advances in Experimental Social Psychology, 19*, 1-39. doi:10.1016/s0065-2601(08)60211-7

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182. doi:10.1037/0022-3514.51.6.1173

Beilock, S. L., Rydell, R. J., & McConnell, A. R. (2007). Stereotype threat and working memory: Mechanisms, alleviation, and spillover. *Journal of Experimental Psychology–General, 136*, 256-276. doi:10.1037/0096-3445.136.2.256

Blanton, H., Crocker, J., & Miller, D. T. (2000). The effects of in-group versus out-group social comparison on self-esteem in the context of a negative stereotype. *Journal of Experimental Social Psychology, 36*, 519-530. doi:10.1006/jesp.2000.1425

Brodish, A. B., & Devine, P. G. (2009). The role of performance-avoidance goals and worry in mediating the relationship between stereotype threat and performance. *Journal of Experimental Social Psychology, 45*, 180-185. doi:10.1016/j.jesp.2008.08.005

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

Chalabaev, A., Stone, J., Sarrazin, P., & Croizet, J. C. (2008). Investigating physiological and self-reported mediators of stereotype lift effects on a motor task. *Basic and Applied Social Psychology, 30*, 18-26. doi:10.1080/01973530701665256

Cottrell, N. B. (1972). Social facilitation. In C. McClintock (Ed.), *Experimental social psychology* (pp. 185-236). New York, NY: Holt, Rinehart & Winston.

Eccles, J. S., Jacobs, J. E., & Harold, R. D. (1990). Gender-role stereotypes, expectancy effects, and parents socialization of gender differences. *Journal of Social Issues, 46*, 183-201. doi:10.1111/j.1540-4560.1990.tb01929.x

Gallagher, A. M., De Lisi, R., Holst, P. C., McGillicuddy-De Lisi, A. V., Morely, M., & Cahalan, C. (2000). Gender differences in advanced mathematical problem solving. *Journal of Experimental Child Psychology, 75*, 165-190. doi:10.1006/jecp.1999.2532

Ganley, C. M., Mingle, L. A., Ryan, A. M., Ryan, K., Vasilyeva, M., & Perry, M. (2013). An examination of stereotype threat effects on girls' mathematics performance. *Developmental Psychology, 49*, 1886-1897. doi:10.1037/a0031412

Harkins, S. (2001). The role of task complexity, and sources and criteria of evaluation in motivating task performance. In. S. Harkins (Ed.), *Multiple perspectives on the effects of evaluation on performance: Toward an integration* (pp. 99-131). Norwell, MA: Kluwer Academic. doi:10.1007/978-1-4615-0801-4_5

Harkins, S. G. (2006). Mere effort as the mediator of the evaluation-performance relationship. *Journal of Personality and Social Psychology, 91,* 436-455. doi:10.1037/0022-3514.91.3.436

Huguet, P., & Regner, I. (2007). Stereotype threat among schoolgirls in quasi-ordinary classroom circumstances. *Journal of Educational Psychology, 99,* 545-560. doi:10.1037/0022-0663.99.3.545

Hull, C. (1943). *Principles of behavior*. New York, NY: Appleton-Century-Crofts.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107*, 139-155. doi:10.1037/0033-2909.107.2.139

Jamieson, J. P. (2009). *The role of motivation in blatant stereotype threat, subtle stereotype threat, and stereotype priming*. Doctoral dissertation, Northeastern University. Retrieved from iris.lib.neu.edu/cgi/viewcontent.cgi?article=1009&context=psych_diss&sei-redir=1&referer=http%3A%2F%2F

Jamieson, J. P., & Harkins, S. G. (2007). Mere effort and stereotype threat performance effects. *Journal of Personality and Social Psychology, 93*, 544-564. doi:10.1037/0022-3514.93.4.544

Jamieson, J. P., & Harkins, S. G. (2009). The effect of stereotype threat on the solving of quantitative GRE problems: A mere effort interpretation. *Personality and Social Psychology Bulletin, 35*, 1301-1314. doi:10.1177/0146167209335165

Jamieson, J. P., & Harkins, S. G. (2011). Distinguishing between the effects of stereotype priming and stereotype threat on math performance. *Group Processes & Intergroup Relations, 15*, 291-304. doi:10.1177/1368430211417833

Keller, J. (2007). Stereotype threat in classroom settings: The interactive effect of domain identification, task difficulty and stereotype threat on women students' maths performance. *British Journal of Educational Psychology, 77*, 323-338. doi:10.1348/000709906X113662

Keller, J., & Dauenheimer, D. (2003). Stereotype threat in the classroom: Dejection mediates the disrupting threat effect on women's math performance. *Personality and Social Psychology Bulletin, 29*, 371-381. doi:10.1177/0146167202250218

Kessels, U., & Hannover, B. (2008). When being a girl matters less: Accessibility of gender-related self-knowledge in single-sex and coeducational classes and its impact on students' physics-related self-concept of ability. *British Journal of Educational Psychology, 78*, 273-289. doi:10.1348/000709907X215938

McFall, S. R., Jamieson, J. P., & Harkins, S. G. (2009). Testing the mere effort account of the evaluation-performance relationship. *Journal of Personality and Social Psychology, 96,* 135-154. doi:10.1037/a0012878

Neuville, E., & Croizet, J. C. (2007). Can salience of gender identity impair math performance among 7–8 year old girls? The moderating role of task difficulty. *European Journal of Psychology of Education, 22*, 307-316. doi:10.1007/BF03173428

Nguyen, H. H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology, 93,* 1314-1334.

Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., … Greenwald, A. G. (2009). National differences in gender-science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences of the United States of America, 106*, 10593-10597. doi:10.1073/pnas.0809921106

O'Brien, L. T., & Crandall, C. S. (2003). Stereotype threat and arousal: Effects on women's math performance. *Personality and Social Psychology Bulletin, 29*, 782-789. doi:10.1177/0146167203029006010

Picho, K., Rodriguez, A., & Finnie, L. (2013). Exploring the moderating role of context on the mathematics performance of women under stereotype threat: A meta-analysis. *Journal of Social Psychology, 153*, 299-333. doi:10.1080/00224545.2012.737380

Picho, K., & Stephens, J. M. (2012). Culture, context and stereotype threat: A comparative analysis of young Ugandan women in coed and single-sex schools. *Journal of Educational Research, 105*, 52-63. doi:10.1080/00220671.2010.517576

Quinn, D. M., & Spencer, S. J. (2001). The interference of stereotype threat with women's generation of mathematical problem-solving strategies. *Journal of Social Issues, 57,* 55-71.

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

Shapiro, J. R., & Neuberg, S. L. (2007). From stereotype threat to stereotype threats: Implications of a multi-threat framework for causes, moderators, mediators, consequences, and interventions. *Personality and Social Psychology Review, 11,* 107-130. doi:10.1177/1088868306294790

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35*, 4-28. doi:10.1006/jesp.1998.1373

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52*, 613-629. doi:10.1037/0003-066X.52.6.613

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test-performance of African Americans. *Journal of Personality and Social Psychology, 69*, 797-811. doi:10.1037/0022-3514.69.5.797

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image:

The psychology of stereotype and social identity threat. *Advances in Experimental Social Psychology, 34,* 379-440. doi:10.1016/S0065-2601(02)80009-0

Stoet, G., & Geary, D. C. (2012). Can stereotype threat explain the gender gap in mathematics performance and achievement? *Review of General Psychology, 16*, 93-102. doi:10.1037/a0026617

Walton, G. M., & Cohen, G. L. (2003). Stereotype lift. *Journal of Experimental Social Psychology, 39*, 456-467. doi:10.1016/S0022-1031(03)00019-2

Wei, T. E. (2012). Sticks, stones, words, and broken bones: New field and lab evidence on stereotype threat. *Educational Evaluation and Policy Analysis, 34*, 465-488. doi:10.3102/0162373712452629

Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology, 89*, 696-716. doi:10.1037/0022-3514.89.5.696

Zajonc, R. B. (1965). Social facilitation. *Science, 149*, 269-274. doi:10.1126/science.149.3681.269