Journal of Research in Personality 63 (2016) 36–43

Contents lists available at ScienceDirect

Journal of Research in Personality

journal homepage: www.elsevier.com/locate/jrp

Do performance avoidance goals moderate the effect of different types of stereotype threat on women's math performance?

Katherine M. Finnigan [a], Katherine S. Corker [b,*]

[a] University of California, Davis, United States
[b] Kenyon College, United States

ARTICLE INFO

Article history:
Received 6 February 2016
Revised 28 April 2016
Accepted 12 May 2016
Available online 14 May 2016

Keywords:
Stereotype threat
Academic achievement
Achievement goals
Gender stereotypes
Math achievement

ABSTRACT

Stereotype threat is considered to be a robust effect that explains persistent gender gaps in math performance and scientific career trajectories. Some evidence suggests stereotype threat effects are buffered by adoption of performance avoidance goals (Chalabaev, Major, Sarrazin, & Cury, 2012). With 590 American female participants, we closely replicated Chalabaev et al. (2012). Results showed no significant main or interaction effects for stereotype threat or performance avoidance goals, despite multiple controls. We conclude that effects of stereotype threat might be smaller than typically reported and find limited evidence for moderation by avoidance achievement goals. Accordingly, stereotype threat might not be a major part of the explanation for the gender gap in math performance, consistent with recent meta-analyses (Flore & Wicherts, 2015).

1. Introduction

Despite research suggesting girls receive higher marks than boys across all subjects, including math and science courses, girls in the United States score lower than boys on the math section of the SAT (College Board, 2013; $d = -0.27$), and make up only 25% of the STEM (science, technology, engineering, and mathematics) workforce when women make up 47% of the overall workforce (National Math & Science Initiative, 2013). Given the importance of standardized tests for college admissions, the need for research to accurately identify the forces behind women's poorer performance on tests of math ability is imperative for optimal interventions and policy changes. The present research aimed to replicate the effect of stereotype threat on math performance and to determine how different types of stereotype threat impact performance. Few direct replications of stereotype threat have been published, and we aim to test this effect in a large online sample of American women. Further, we tested how threat interacts with performance avoidance goals, which are defined as aiming at avoiding poor performance relative to one's peers (Elliot & McGregor, 2001).

1.1. Different types of stereotype threat

One of the most prominent psychological theories used to explain the math-achievement gap is stereotype threat, which occurs when individuals are fearful of confirming negative stereotypes associated with their group. Steele and Aaronson (1995) posit that "in situations where the stereotype is applicable, one is at risk of confirming it as a self-characterization, both to one's self and to others who know the stereotype" (p. 808). This theory was applied to the women in mathematics stereotype by Spencer, Steele, and Quinn (1999)—a landmark paper that illustrated the negative effect of framing a test as diagnostic of math ability or gender differences on female math performance. Many studies have used cues stemming from Spencer et al., ranging from test framings to threatening test administrators (Shapiro & Neuberg, 2007). Recent research has attempted to make sense of these different types of stereotype threat and how they distinctly impact performance.

Shapiro and Neuberg's (2007) multi-threat framework classifies threats based on the threat's source and target. The three sources of stereotype threat are self, outgroup (e.g., men), and ingroup (e.g., other women in mathematics). The two targets of threat are the self and the group (e.g., an individual woman is the self, the female gender as a whole is the group). The authors theorize that these threats have different effects on individuals; for example, group-as-target threats are believed to lead to more negative comparisons between the self and ingroup and outgroup members

* Corresponding author at: Department of Psychology, Kenyon College, 203 N. College Rd., Gambier, OH 43022, United States.

E-mail address: corkerk@kenyon.edu (K.S. Corker).

http://dx.doi.org/10.1016/j.jrp.2016.05.009
0092-6566/© 2016 Elsevier Inc. All rights reserved.

than self-as-target threats do. The current study focuses on two "self-as-source" threats: *group-concept threat* and *self-concept threat*. Group-concept threat targets the stereotyped group to which one belongs and differentiates the stereotyped group from the majority (e.g., men in mathematics). Group-concept threats thus confirm, to a woman in mathematics, that her *gender as a whole* is bad at math. Self-concept threat is the fear that one's performance will confirm that *oneself* is bad at math. Unlike group-concept threat, self-concept threat targets one's inherent attributes or skills, and self-concept threat does not require identification with the group (Shapiro & Neuberg, 2007). Using two threat manipulations, we aimed to experimentally test if self-concept threat or group-concept threat differentially affect math test performance and if these threats interact with variables believed to moderate the effect of stereotype threat (e.g., gender identity and math identity).

1.2. Threat and performance avoidance goals

In the last decade, stereotype threat research has shifted in focus from investigating the existence of stereotypes to examining the mechanisms that influence test performance for women (Shapiro & Williams, 2012). Performance avoidance goals have been identified as a potential consequence of stereotype threat for women in mathematics because individuals endorse those goals to avoid performing poorly in comparison to peers (Elliot & McGregor, 2001). For example, if a student aims to prevent scoring low on the SAT in comparison to her friends and other students she knows, she adopts a performance avoidance goal orientation. Performance avoidance goals and stereotype threat are associated with anxiety, fear of failure, and negative evaluations of self-competence, and thus they typically impair performance (Smith, Sansone, & White, 2007).

However, individuals who adopt performance-avoidance goals are in a state similar to those completing tasks that are inherently threatening, and theory suggests that matching goal orientation to task characteristics boosts performance (Cesario, Higgins, & Scholer, 2008). Stereotype threat (particularly group-concept threat) emphasizes social comparisons, a feature that is an integral component of performance-avoidance goals. Grimm, Markman, Maddox, and Baldwin (2009) posit that an avoidance goal orientation matches a stereotype threat situation better than an approach goal orientation does; this fit may then lead stereotyped individuals to use cognitive strategies that help them avoid negative outcomes and perform as well as non-threatened individuals.

This matching effect has been tested and found in studies on stereotype threat for women in math. Chalabaev, Major, Sarrazin, and Cury (2012) operationalized stereotype threat through a "math ability" cue (self-concept threat) in one study and a "gender differences" cue (group-concept threat) in a second study, in accordance with Spencer et al.'s (1999) manipulations. Both of Chalabaev et al.'s studies demonstrated that inducing a performance-avoidance goal produces better performance under stereotype threat on a math test when compared to no goal, an effect subsequently detected by others (Deemer, Smith, Carroll, & Carpenter, 2014).

1.3. The present study

Given the theoretical links between stereotype threat and performance-avoidance goals and the evidence supporting their interaction, we sought (1) to test this interaction in a large-scale, pre-registered replication, using Chalabaev et al.'s (2012) basic design, and (2) to extend their design to directly compare two types of threat from Shapiro and Neuberg's (2007) framework. Additionally, we tested for differences in magnitude of effect between self-concept and group-concept threat with various moderators (detailed below). We predicted that we would find a stereotype threat by performance avoidance goal interaction, such that women under threat would perform better on a math test with a performance avoidance goal compared to no goal. We also predicted that the effect of group-concept threat would be stronger when compared to self-concept threat. We tested these hypotheses with an all-female sample, because the performance of males is not inhibited by stereotype threat in the domain of mathematics (Walton & Cohen, 2003).

Replications of stereotype threat studies are relatively uncommon, and we aimed to test this effect in a larger sample than the vast majority of research in this area (Stoet & Geary, 2012). Policy-makers and social psychologists point to stereotype threat as a major force behind the gender gap in STEM, although there are few tests of stereotype threat in this domain on samples of non-undergraduates (Flore & Wicherts, 2015). By conducting highly-powered test of stereotype threat in a diverse sample, this study will provide insight into the generalizability of stereotype threat, and it will test the extent to which threat can impact performance in an online setting.

2. Method

2.1. Pilot study & design

We pre-registered our design and analysis plan (see https://osf.io/kms6g). In order to measure math performance under differing conditions of stereotype threat, we used 10 released questions from the GRE quantitative reasoning section, also used in Chalabaev et al. (2012). We ran a pilot study testing math ability with three tests of different degrees of difficulty (i.e., sixth grade, tenth grade, and the GRE) to ensure that these questions were appropriately difficult for the intended audience of female Mechanical Turk workers. The GRE was best suited to enable the detection of stereotype threat, as per Spencer et al. (1999), because the other two tests were not difficult enough; see the online supplemental materials, also available at https://osf.io/jze8c/, for more detailed results of the pilot study.

In order to determine the appropriate sample size for hypothesis tests, we conducted a sensitivity analysis using R Statistical Software (Version 3.1.1), specifically the *pwr* package (Champley, 2012). The analyses revealed that with 600 participants (100 in each condition) and 80% power, assuming $\alpha = 0.05$, the study could reliably detect an effect of $d = 0.40$. With the same number of participants and 90% power assuming $\alpha = 0.05$, the study could detect an effect of $d = 0.46$.

We therefore aimed to collect 600 participants. Hypotheses were tested in a 3 (Type of Threat: Self-Concept Threat, Group-Concept Threat, No Threat) × 2 (Goal: Performance Avoidance vs. No Goal) between-subjects design.

2.2. Materials

2.2.1. Stereotype threat and achievement goal manipulations

Participants received one of six prompts, in the form of audio instructions. In the neutral (no threat), no goal condition, participants were told, "You are going to perform a problem solving test." In the math ability (self-concept threat), no goal condition, participants were told, "You are going to perform a math test." In the gender differences (group-concept threat), no goal condition, participants were told, "Previous research has sometimes shown gender differences in math ability…the test you are about to take has been shown to produce gender differences." In all three performance avoidance goal conditions, participants were additionally told that the test "is designed to help us identify people who are exceptionally weak in their mathematical reasoning abilities." See https://osf.io/ac259/ for full text of the manipulations.

2.2.2. Math performance

Each of the 10 questions was multiple-choice with five answer choices per question. This set of questions was shown to be a fairly reliable measure of math performance within this sample ($M = 4.32$, $SD = 2.22$, $\alpha = 0.63$).

2.2.3. Math and gender identity

The Math Identification Questionnaire (MIQ; Brown & Josephs, 1999) is a measure of domain identification with mathematics; individuals who are highly math-identified are thought to consider their mathematics abilities and skills to be important parts of their identity, and they highly value mathematics (Brown & Josephs, 1999). The questionnaire contained four items (e.g., "My math abilities are very important to me") rated on a Likert scale of 1 (*strongly disagree*) to 7 (*strongly agree*). For this sample, the average of the four questions produced a reliable index of math identification ($\alpha = 0.75$), with an average score of $M = 4.18$ ($SD = 1.34$). We computed a Pearson's r correlation between the MIQ score and the number of GRE questions correctly answered and found a moderate, positive correlation between MIQ scores and math test scores ($r = 0.31$).

Schmader (2002) modified four items from the Collective Self-Esteem Scale (Luhtanen & Crocker, 1992) in order to assess the "perceived importance of gender identity to self-definition" (Schmader, 2002, p. 196). These same four items (e.g., "Being a woman is an important reflection of who I am") were rated by our participants on a Likert scale of 1 (*strongly disagree*) to 7 (*strongly agree*). As in Schmader (2002), this set of questions was a reliable measure of gender identity ($\alpha = 0.84$). We averaged the four items together ($M = 5.10$, $SD = 1.34$). The correlation with math test scores was not significantly different from zero ($r = -0.02$).

To test the stereotype threat hypothesis, it is required that participants be identified with the domain in question (math) and with their gender (female; Steele, 1997). We compared the scale means in the current sample to values obtained in other published studies. When necessary, we transformed these published values to a seven-point scale by dividing means and standard deviations by the number of scale points used and then multiplying by 7. We then meta-analyzed the means using the *metafor* package with random effects in R (Viechtbauer, 2010). We located four studies that reported the gender identification questionnaire, but two of these studies failed to include standard deviations. The remaining two studies, when meta-analytically combined, had a mean of 5.20 [95% CI: 4.74, 5.67], which placed the current mean of 5.10 inside the confidence interval. By contrast, we identified six studies using the math identification questionnaire, which had a meta-analytic mean of 4.73 [95% CI: 4.29, 5.16]. Our mean of 4.18 was lower than the lower bound of the confidence interval, indicating that our sample was somewhat less math identified than previous samples. However, one study (Gilbert, O'Brien, Garcia, & Marx, 2015) had a mean close to the value currently observed. Furthermore, the mean was above the mid-point of the seven-point scale, giving some evidence that math identification was not excessively low in the current sample.

2.3. Participants

The full sample consisted of $N = 606$ individuals recruited from Mechanical Turk (Buhrmester, Kwang, & Gosling, 2011). We planned exclusions for analyses: participants had to be female, from the United States, and between the ages of 18–60 years ($N = 10$ participants did not meet these criteria). We also excluded participants who skipped more than five questions ($N = 6$) on the math test to ensure that the results were not impacted by participants who appeared to exert minimal effort in the study. The final sample contained $N = 590$ individuals, with an age range of 18–60 ($M = 32.80$, $SD = 9.99$). Of those participants who indicated their race, 78.1% identified as White, 7.8% identified as Asian/Asian-American, 7.3% identified as Black/African-American, 4.2% indicated that they were a different race, and 1.0% indicated that they were Native American. Of these individuals, 6.9% identified as Hispanic or Latino. A small number of participants (1.5%) did not indicate their race.

Of those individuals who indicated their education level, 38.6% received a high school diploma/GED, 49.5% received a 2-/4-year college degree, and 11.7% received an advanced degree (e.g., M. D., J.D., PhD.). One participant did not indicate education level. Many participants indicated that they had previously taken standardized tests, specifically the SAT (48.5%), the ACT (31.9%), and the GRE (10.7%). Some participants (16.3%) were enrolled in an undergraduate or graduate program when they completed the study. In order to control for math experience, participants indicated how many years of math classes they had taken since their freshman year of high school; answers ranged from 0 to 20 years ($M = 4.93$, $SD = 2.41$). Usable SAT, ACT, and GRE tests scores were not reported for the majority of participants, so we were unable to consider achievement test scores as a control variable.

2.4. Procedure

The study was advertised through Mechanical Turk as a study of cognitive tasks (see https://osf.io/wj26u/ for full posting). Participants who completed the pilot study were ineligible. Participants were compensated $1.00 for taking part in the study. Users were redirected to an online survey website after giving their consent to participate.

To ensure that participants were able to hear the audio manipulation, participants first listened to a female voice state a key word ("television") and were asked to choose the correct word from a list. If they did not select the correct word, participants were redirected to a page explaining the necessity for having a computer that could play content from YouTube.com; participants who selected the correct answer were directed to a page that contained the main audio manipulation.

The audio manipulations were recorded in a female voice, using the same scripts from the tape-recorded instructions in Chalabaev et al. (2012). Participants randomly received one of six audio instructions for the test, in accordance with the 3 (Math Ability, Gender Differences, No Threat) × 2 (Performance Avoidance Goal, No Goal) design.

The instructions began with the stereotype threat/goal audio manipulations. After hearing the audio manipulation, participants read additional specific directions, which indicated that they could use a pencil and scratch paper and that they had one minute to complete each question, with four seconds in between each question (as in Chalabaev et al., 2012). Next, participants received the 10 GRE questions in random order. If participants did not answer a question after one minute, they were immediately advanced to the four second buffer screen and then the next question. Incomplete questions were scored as incorrect (however, results were similar even if the test was scored for accuracy, i.e., the number of questions correct out of the number of questions answered).

Manipulation checks were administered immediately following GRE questions. Participants answered an overt, open-ended item: "What type of ability does this test measure?" Participants then rated two items on a Likert scale from 1 (*strongly disagree*) to 7 (*strongly agree*): "The purpose of this task is to identify students who are exceptionally weak" (performance-avoidance goal) and "The purpose of this task is to identify students who are exceptionally strong" (performance-approach goal). These questions were identical to the manipulation checks used in Chalabaev et al. (2012).

After the manipulation checks, participants completed the Math Identity Questionnaire and the Gender Identity Questionnaire in a counterbalanced order; items within each of the questionnaires were also randomized. Following these two questionnaires, participants completed a demographic questionnaire.

2.5. Summary of major differences from Chalabaev et al. (2012)

The current study used a larger sample than Chalabaev et al.'s (2012) studies: $N = 590$ vs. $N = 86$ and $N = 58$. Chalabaev et al. used a 2 (threat present vs. threat absent) × 3 (goal type: performance avoidance, performance approach, and no goal) design in both studies. As detailed above, we compared performance avoidance goals to the no goal condition under one of two types of threat (group-concept vs. self-concept) or no threat.

Participants completed the basic procedure from Chalabaev et al.'s (2012) study before completing the math identity and gender identity measures; these measures were not collected in the original study. As noted above, the observed means for gender identification are consistent with other studies of stereotype threat. However, the current sample scored slightly lower on identification with math than other stereotype threat samples.

Participants were allowed to use scratch paper and a calculator, which was not allowed in the original. We used adult women participants and collected data online, whereas Chalabaev et al. (2012) collected data from exclusively undergraduate students in person at the University of California, Santa Barbara. Our participants were paid $1 for their participation, whereas Chalabaev et al.'s participants were given course credit or $10 for their participation. Chalabaev et al. additionally manipulated performance approach goals, but did not find a significant effect, and to simplify the design, we eliminated this factor from the design. We believe the replication to be fairly exact in all other ways.

3. Results

All statistical analyses were completed in R Statistical Software (Version 3.1.1), and the script for the analyses is available at https://osf.io/xfs9d/, and the data file itself is at https://osf.io/pabns/. 95% confidence intervals for effect sizes were calculated using the *compute.es* package in R (Del Re, 2014). Participants were randomly assigned to one out of six possible conditions: Math Ability/Performance Avoidance goal ($N = 103$), Math Ability/No goal ($N = 92$), Gender Differences/Performance Avoidance goal ($N = 102$), Gender Differences/No goal ($N = 90$), Neutral/Performance Avoidance goal ($N = 111$), Neutral/No goal ($N = 92$). See Fig. 1 for the means and 95% confidence intervals for math test performance in each condition.

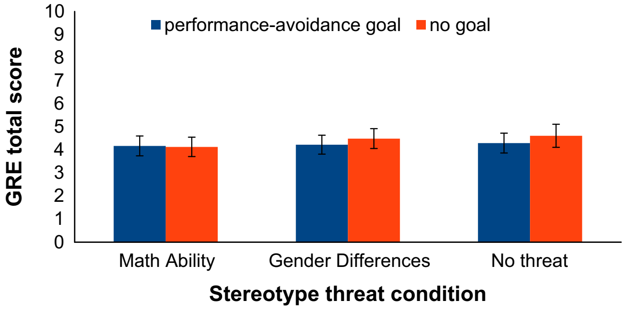

Fig. 1. Means and 95% confidence intervals for the total score on the math test by condition.

3.1. Manipulation check

The first manipulation check was meant to ensure that the performance-avoidance goal manipulation was successful. Participants who received the manipulation should have agreed with the statement "The purpose of this test was to identify students who are exceptionally weak." An independent samples *t*-test revealed that participants who were in the performance-avoidance goal condition ($M = 5.97$, $SD = 1.43$) agreed significantly more with the item than those who did not receive the manipulation ($M = 3.23$, $SD = 1.53$; Welch's $t(561.47) = 22.24$, $p < 0.001$, $d = 1.84$, 95% CI = [1.64, 2.03]).

For the overt stereotype threat manipulation check item, two raters scored the responses, and the first author's coding was used to score the items as math-related or non-math-related (Kappa = 0.98, 95% CI = [0.97, 1.00]). We completed a Pearson's chi-square test that indicated that those in the Math Ability condition were significantly more likely to indicate that the test measured math ability than those in the other two conditions, $\chi^2(1, N = 590) = 9.05$, $p = 0.003$, Odds Ratio = 1.57, 95% CI = [1.17, 2.11]. Conversely, those who did not receive any stereotype threat manipulation were significantly more likely to indicate that the test was a measure of problem-solving ability than those in the threat conditions, $\chi^2(1, N = 590) = 97.92$, $p < 0.001$, Odds Ratio = 5.04, 95% CI = [3.65, 6.95].

3.2. Task performance

We hypothesized that there would be a crossover interaction, such that those in the performance-avoidance condition would perform better than those in the no goal condition when under stereotype threat, but would perform worse than those in the no goal condition in the neutral condition under no threat. We also hypothesized that the effect of threat would be stronger when threat was operationalized as group-concept threat (i.e., gender differences threat) compared to self-concept threat (i.e., math ability threat).

3.2.1. Predicted model

To test the 3 (Math Ability, Gender Differences, Neutral) × 2 (Performance Avoidance Goal, No Goal) model, we constructed a linear regression with one contrast coded variable that compared the two stereotype threat manipulations to the neutral (control) condition (−2 = Neutral, 1 = Gender Differences, 1 = Math Ability), one contrast coded variable that compared each of the stereotype threat conditions to each other (−1 = Gender Differences, 0 = Neutral, 1 = Math Ability), and a dummy coded variable for the goal manipulation (1 = performance-avoidance goal; 0 = no goal). This method was chosen to enable a test of the main effect of stereotype threat across the two threat conditions, as well as a comparison of the two threat types.

When entered into a linear model, the model revealed no significant main or interaction effects (see Table 2; see Fig. 2 for effect size comparison between current study and Chalabaev et al., 2012, for threatened participants' performance under performance avoidance goals vs. no goals).

The performance boost for those induced with performance-avoidance goals under math ability threat was far weaker in our study ($d = 0.08$, 95% CI = [−0.20, 0.37]) than the effect size ($d = 0.72$, 95% CI = [−0.02, 1.45]) found in Chalabaev et al. (2012). The effect size for the performance-avoidance goal/no goal comparison under gender differences threat

Table 1
Studies in mini meta-analysis of GIQ and MIQ scores.

Author(s)/year	Scale	*M*	*SD*	Scale	Item	*N*	Trans. *M*	Trans. *SD*
Cundiff, Vescio, Loken, and Lo (2013)	GIQ	1.22	0.82	−2.5 to 2.5	4	876	5.51	0.96
Eriksson and Lindholm (2007)	GIQ	3.45	NA	1–5	4	112	4.83	NA
Schmader (2002)	GIQ	3.82	NA	1–5	4	32	5.35	NA
Wout, Danso, Jackson, and Spencer (2008)	GIQ	4.83	1.17	1–7	4	59	4.83	1.17
Brown and Josephs (1999)	MIQ	5.19	1.70	1–7[a]	5	33	5.19	1.70
Brown and Pinel (2003)	MIQ	4.90	1.08	1–7	5	49	4.90	1.08
Gilbert, O'Brien, Garcia, and Marx (2015)	MIQ	4.07	1.21	1–7	11	198	4.07	1.21
Inzlicht and Ben-Zeev (2003)	MIQ	5.60	1.15	1–9	13	115	4.36	0.89
Inzlicht and Kang (2010)	MIQ	5.37	1.40	1–7	5	34	5.37	1.40

Note. "GIQ" = Gender Identification Questionnaire (Luhtanen & Crocker, 1992; Schmader, 2002); "MIQ" = Math Identification Questionnaire; Trans. *M*/*SD* = Transformed Mean/Standard Deviation to a 7-point scale for comparison to our scale.

[a] Brown and Josephs (1999) did not state the number of scale points, but in later surveys (Brown & Pinel, 2003) used a 7-point scale. Given the *M*s and *SD*s in this paper, we conclude that the authors likely used a 7-point scale for their analyses.

Table 2
Regression model predicting GRE score with no controls.

Variable	*b*	*SE b*	*t*	*p*
(Intercept)	4.41	0.13	32.83	<0.001
Contrast 1 – ST vs. No Threat (A)	−0.12	0.09	−1.26	0.21
Performance Avoidance Goals (B)	−0.17	0.18	−0.92	0.36
Contrast 2 – SC Threat vs. GC Threat (C)	−0.16	0.16	−0.99	0.33
A × B	0.08	0.13	0.58	0.56
B × C	0.14	0.23	0.61	0.55

Note. Model *df* = 584. ST = stereotype threat; SC Threat = self-concept threat; GC Threat = group-concept threat. R^2 = 0.01 (R^2 adjusted = 0.00).

(d = −0.13, 95% CI = [−0.42, 0.15]) was in the opposite direction as, and much smaller than, the effect size (d = 0.78, 95% CI = [−0.13, 1.69]) found by Chalabaev et al.

The main effect of stereotype threat was similarly weak in our sample; the math ability main effect size was small (d = −0.13, 95% CI = [−0.32, 0.07]), as was the gender differences main effect size (d = −0.02, 95% CI = [−0.19, 0.16]). The observed effect size in our sample is much smaller than in some other studies on stereotype threat, but Flore and Wicherts (2015) estimated that the average effect size of stereotype threat is equal to −0.22 for girls under age 18. Although our sample was composed of adult women, our data converges with Flore and Wicherts' estimate, showing that the effect size of stereotype threat probably smaller than previously supposed.

3.2.2. Controls

We ran the predicted regression model several times with different controls for math experience and other variables that may have moderated stereotype threat effects. Indeed, the students in our sample were significantly more math identified than non-students (t(147.81) = 2.345, p = 0.02). If math identification is important for observing stereotype threat effects, we expect student status, or math identification directly, to moderate the size of effects. Additionally, age could serve as another proxy for math identification (however, neither gender identification (r = 0.09) nor math identification (r = −0.02) correlated with age in our sample). Regardless, we tested these variables (math identification, student status, and age) for interaction effects.

None of these controls had significant interaction effects with the stereotype threat or achievement goal manipulations, including education level (see Supplementary Tables). None of the main effects for these variables were significant except for education level (see Table 3; more educated students performed better on the math test than less educated students). Controlling math identity and gender identity also failed to reveal any significant effects, although math identification was the best predictor of math performance (b = 0.508, t(588) = 7.776, p < 0.001) on its own. There was a marginally significant three-way interaction between education level, stereotype threat, and achievement goal manipulations (p = 0.07). However, as depicted in the supplemental materials (Supplemental Fig. 2), the interpretation of the interaction was unclear, with results showing a mix of small beneficial and detrimental effects of various combinations of threat and avoidance goals. There was also a marginal three-way interaction with student status, such that *non-students* showed some evidence of threat effects, but only in the absence of a performance avoidance goal (see Supplemental Table 3). At the risk of over interpreting noise, and in the absence of a clear pattern of results, we did not probe these marginal interactions further. The supplemental materials and the OSF page (https://osf.io/jze8c/) contain regression tables for math and gender identity, as well as other supplemental tables. See Table 4 for inter-correlations between the different control variables.

We considered the possibility of floor effects on the math test in exploratory analyses, but the distribution of math performance did not suggest a floor effect. Further we repeated analyses while removing participants who scored 0 out of 10 (N = 18), and we found little evidence of changes in the results (Johnson, Cheung, & Donnellan, 2014). The results do not appear to be impacted by floor effects.

4. Discussion

The current study aimed to assess the impact of different types of stereotype threat in a large sample of adult women and to determine if inducing performance-avoidance goals improves performance under threat. Our results indicate a failure to replicate Chalabaev et al. (2012), with no evidence suggesting the presence of significant stereotype threat main effects, nor any moderation by performance avoidance goals, in spite of the fact that the current replication study had a much larger sample size than the original study.

Although performance-avoidance goals were theorized to bolster performance under stereotype threat, no significant main or interaction effects of performance-avoidance goals emerged. Several controls for math experience, math identification, gender identification, education level, and age were also considered, none of which revealed any significant main or interaction effects. We did observe a lower average level of math identification in the current sample than other studies of stereotype threat, which may have inhibited our ability to observe the effect. However, we did not observe any moderation of the main effect by math identification or other similar variables (e.g., age). Further, Nguyen and Ryan (2008) suggest that moderately math-identified women are affected more strongly by stereotype threat than strongly

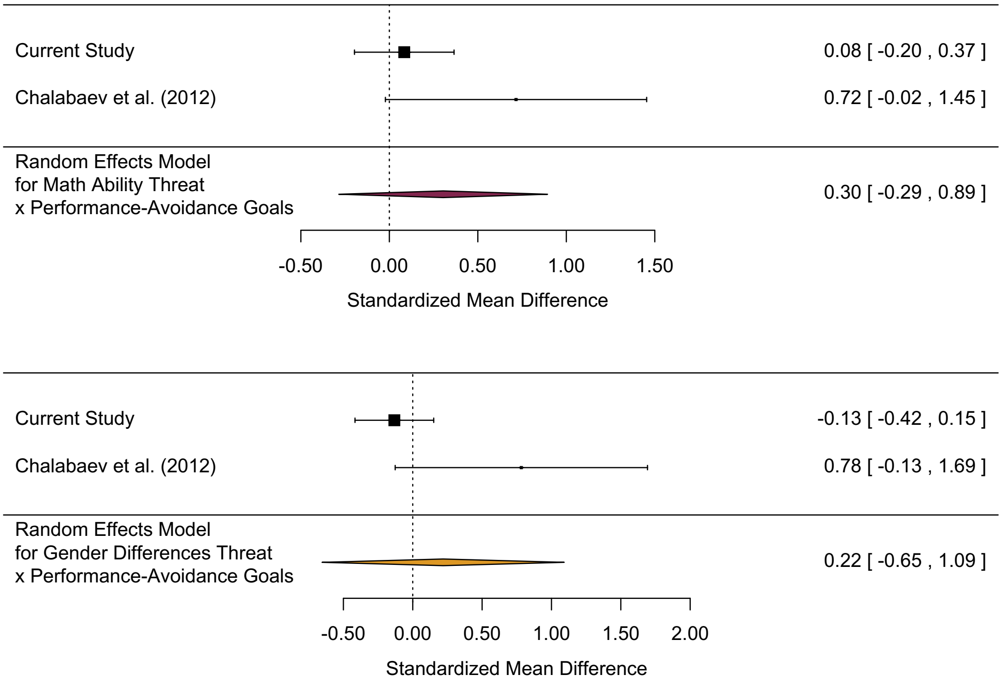

Fig. 2. Effect sizes for the current study and Chalabaev et al. (2012). Effect sizes compare performance for threatened participants under performance avoidance goal vs. no goal for self-concept threat (top panel) and group-concept threat (bottom panel). Values for Chalabaev et al. (2012) calculated based on means and standard deviations reported in their Table 1.

Table 3
Regression model predicting GRE score with education level as control.

Variable	*b*	*SE b*	*t*	*p*
(Intercept)	4.02	0.22	18.52	<0.001
Contrast 1 – ST vs. No Threat (A)	−0.29	0.16	−1.78	0.08
Performance Avoidance Goal (B)	−0.21	0.29	−0.71	0.48
Education level				
HS Degree versus College (C)	0.58	0.29	2.01	0.04
HS Degree versus > College (D)	1.06	0.47	2.25	0.02
Contrast 2 – SC Threat vs. GC Threat (E)	−0.06	0.25	−0.22	0.82
A × B	0.39	0.21	1.83	0.07
A × C	0.23	0.20	1.12	0.26
A × D	0.51	0.35	1.48	0.14
B × C	−0.07	0.39	−0.19	0.85
B × D	0.37	0.62	0.60	0.55
B × E	−0.27	0.35	−0.77	0.44
C × E	−0.17	0.35	−0.50	0.62
D × E	0.16	0.55	0.28	0.78
A × B × C	−0.51	0.28	−1.84	0.07
A × B × D	−0.64	0.44	−1.44	0.15
A × C × E	0.55	0.48	1.15	0.25
B × D × E	0.63	0.74	0.86	0.39

Note. Model *df* = 571. HS = high school; ST = stereotype threat; SC Threat = self-concept threat; GC Threat = group-concept threat. R^2 = 0.06 (R^2 adjusted = 0.03).

Table 4
Inter-correlations between control variables.

	1	2	3	4	5	6	7
1. # of Standardized tests	1.00						
2. Years of math	0.19	1.00					
3. Education level	0.33	0.24	1.00				
4. Math identity	0.14	0.10	0.18	1.00			
5. Age	−0.17	−0.06	0.07	−0.03	1.00		
6. Gender identity	0.14	0.00	0.15	0.15	0.10	1.00	
7. Student status	−0.13	−0.03	0.13	−0.09	0.35	0.02	1.00

math-identified women, indicating that our sample would be adequately identified with math to observe the effect. We cannot rule out low math identification as the mechanism behind our null effects, but the lack of consensus in the literature about the role of domain identification in stereotype threat suggests future research in this area may be warranted.

The current findings are not the first to cast doubt on the magnitude of the stereotype threat effect. Stoet and Geary (2012) found that 55% of 21 studies that attempted to replicate Spencer et al. (1999) produced a significant interaction effect between gender and stereotype threat. Half of the successful effects used SAT scores as a covariate, but only 30% of studies that did not use SAT scores as a covariate replicated the effect. It should be noted that Spencer et al. did not use SAT score as a control, and they still found an effect of stereotype threat. We used years of math education as a control for math ability, but found no moderating effects of this variable on our results. The inclusion of various controls for math ability is not uncontroversial (Stoet & Geary, 2012), but nonetheless, with or without control variables, we found no evidence for the basic effect.

Publication bias has been cited as the cause of discrepancies in results between small sample studies and larger replications (such as the current study) in the domain of stereotype threat. Flore and Wicherts (2015) estimated the effect size of stereotype threat on the performance of adolescent girls in stereotyped domains to be $d = -0.22$. They suggest that publication bias has overinflated the effect size of stereotype threat (Flore & Wicherts, 2015). Publication bias refers to the fact that studies with null findings are often neither written up nor accepted for publication by many top journals (Begg, 1994). They recommended a large-scale replication study to test the stereotype threat hypothesis, and the current research fulfills that request. Incorporating the present study into the meta-analysis would serve to further shrink this effect size estimate (see Fig. 2).

Although the current study consists of a large sample replication, some may attribute the null results to the fact that our participants were not undergraduate students, which may make them less vulnerable to stereotype threat. Many stereotype threat studies use samples composed of undergraduate students, a fact that may lead researchers to overgeneralize stereotype threat to populations of non-students. Although students ($N = 97$) performed better on the GRE test questions compared to non-students and identified more with math, none of the predicted effects were significant for students (conversely, in the no-goal condition we found some limited evidence for a threat effect in non-students). Additionally, because our sample was comprised of Mechanical Turk workers, others may assume that participants did not care about the test, which would make the effect of stereotype threat more difficult to detect (Steele & Aaronson, 1995). There is no evidence, given our manipulation checks and our participants' test results, that our sample cared less about the math test than a sample of undergraduates participating in a study.

Others will likely question the ability of the stereotype threat manipulations to generalize to an online context, especially given that participants in our study completed the math questions outside of the lab without the threat of immediate evaluation by the experimenter, which could potentially dampen threat effects. Despite this doubt, standardized tests are increasingly being administered in computerized formats, so it is important to examine the effect of threat manipulations under non-laboratory conditions (Noyes & Garland, 2008). At the very least, the current study provides evidence to question the generalizability of stereotype threat effects beyond female undergraduate samples tested in laboratory settings.

All told, the current results question not only the ability of performance-avoidance goals to improve performance under threat, but also the generalizability of stereotype threat effects for women in mathematics. Other researchers have suggested that social psychologists have not scrutinized stereotype threat as heavily as other theories, partly because of its usefulness in explaining the STEM gender gap (Flore & Wicherts, 2015; Stoet & Geary, 2012). Stoet and Geary (2012) call for a more critical reading of the stereotype threat literature to question the belief that a causal relationship exists between stereotype threat and worsened math performance.

The present study provides substantial evidence to question the definition and operationalization of stereotype threat and the validity of potential moderators of stereotype threat. Policymakers rely on social psychologists to provide insight into the nature of issues like the gender gap in STEM fields in order to ensure that interventions aimed at closing this gap are effective (Stoet & Geary, 2012). Should policymakers continue to view stereotype threat as the *principal* cause of the STEM gender gap, there may be an opportunity cost as potentially less effective interventions are pursued instead of more beneficial measures (Nguyen & Ryan, 2008; Stoet & Geary, 2012).

Multiple researchers have cited the need for quality replications of published results, especially in social psychology, in large cross-cultural samples (Flore & Wicherts, 2015; Roediger, 2012). We do not conclude that stereotype threat has no effect on performance in every context, but instead we suggest that the effect of existing threat cues may not have as strong an impact on performance as previously thought.

In order to more effectively close the gender gap in STEM fields, we suggest that researchers investigate other explanations beyond stereotype threat that can explain discrepancies in men's and women's math performance, such as the inability to fulfill communal goals and the environment of workplaces in STEM that may discriminate against women who choose to have a family and desire more flexible hours (Diekman, Clark, Johnston, Brown, & Steinberg, 2011; Heilbronner, 2012). Although stereotype threat is an appealing answer to the question of what causes the gender gap in STEM, we advise that researchers not rely solely on literature that may be skewed by inflated effect sizes and small samples.

Author contributions

K. M. F. developed the study concept and design and collected the data, under the supervision of K. S. C. K. M. F. conducted data analysis and interpretation; K. S. C independently checked all analyses. K. M. F. drafted the manuscript, and the authors revised the draft together. Both authors approved the final version of the manuscript for submission.

Conflict of interest

The authors declared no potential conflicts of interest and received no financial support for the research, authorship, and/or publication of this article.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jrp.2016.05.009.

References

Begg, C. B. (1994). Publication bias. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 400–408). New York: Russell Sage Foundation.

Brown, R. P., & Josephs, R. A. (1999). A burden of proof: Stereotype relevance and gender differences in math performance. *Journal of Personality and Social Psychology, 76*(2), 246–257. http://dx.doi.org/10.1037/0022-3514.76.2.246.

Brown, R. P., & Pinel, E. C. (2003). Stigma on my mind: Individual differences in the experience of stereotype threat. *Journal of Experimental Social Psychology, 39*(6), 626–633. http://dx.doi.org/10.1016/S0022-1031(03)00039-8.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical turk: A new source of inexpensive, yet high quality, data? *Perspectives on Psychological Science, 6*(1), 3–5. http://dx.doi.org/10.1177/1745691610393980.

Cesario, J., Higgins, E. T., & Scholer, A. A. (2008). Regulatory fit and persuasion: Basic principles and remaining questions. *Social and Personality Psychology Compass, 2*(1), 444–463. http://dx.doi.org/10.1111/j.1751-9004.2007.00055.x.

Chalabaev, A., Major, B., Sarrazin, P., & Cury, F. (2012). When avoiding failure improves performance: Stereotype threat and the impact of performance goals. *Motivation and Emotion, 36*, 130–142. http://dx.doi.org/10.1007/s11031-011-9241-x.

Champley, S. (2012). pwr: Basic functions for power analysis. R package version 1.1.1http://CRAN.R-project.org/package=pwr.

College Board (2013). 2013 College-bound seniors total group profile report Retrieved September 24, 2014, fromhttp://media.collegeboard.com/digitalServices/pdf/research/2013/TotalGroup-2013.pdf.

Cundiff, J. L., Vescio, T. K., Loken, E., & Lo, L. (2013). Do gender–science stereotypes predict science identification and science career aspirations among undergraduate science majors? *Social Psychology of Education, 16*(4), 541–554.

Deemer, E. D., Smith, J. L., Carroll, A. N., & Carpenter, J. P. (2014). Academic procrastination in STEM: Interactive effects of stereotype threat and achievement goals. *The Career Development Quarterly, 62*(2), 143–155. http://dx.doi.org/10.1002/j.2161-0045.2014.00076.x.

Del Re, A. C. (2014). compute.es: Compute effect sizes. R package version 0.4http://cran.r-project.org/web/packages/compute.es/compute.es.pdf.

Diekman, A. B., Clark, E. K., Johnston, A. M., Brown, E. R., & Steinberg, M. (2011). Malleability in communal goals and beliefs influences attraction to stem careers: Evidence for a goal congruity perspective. *Journal of Personality and Social Psychology, 101*(5), 902–918. http://dx.doi.org/10.1037/a0025199.

Elliot, A. J., & McGregor, H. A. (2001). A 2 × 2 achievement goal framework. *Journal of Personality and Social Psychology, 80*(3), 501–519. http://dx.doi.org/10.1037/0022-3514.80.3.501.

Eriksson, K., & Lindholm, T. (2007). Making gender matter: The role of gender-based expectancies and gender identification on women's and men's math performance in Sweden. *Scandinavian Journal of Psychology, 48*(4), 329–338. http://dx.doi.org/10.1111/j.1467-9450.2007.00588.x.

Flore, P. C., & Wicherts, J. M. (2015). Does stereotype threat influence performance of girls in stereotyped domains? A meta-analysis. *Journal of School Psychology, 53*(1), 25–44. http://dx.doi.org/10.1016/j.jsp.2014.10.002.

Gilbert, P. N., O'Brien, L. T., Garcia, D. M., & Marx, D. M. (2015). Not the sum of its parts: Decomposing implicit academic stereotypes to understand sense of fit in math and English. *Sex Roles, 72*, 25–39. http://dx.doi.org/10.1007/s11199-014-0428-y.

Grimm, L. R., Markman, A. B., Maddox, W. T., & Baldwin, G. C. (2009). Stereotype threat reinterpreted as a regulatory mismatch. *Journal of Personality and Social Psychology, 96*(2), 288–304. http://dx.doi.org/10.1037/a0013463.

Heilbronner, N. N. (2012). The STEM pathway for women: What has changed? *Gifted Child Quarterly, 57*(1), 39–55. http://dx.doi.org/10.1177/0016986212460085.

Inzlicht, M., & Ben-Zeev, T. (2003). Do high-achieving female students underperform in private? The implications of threatening environments on intellectual processing. *Journal of Educational Psychology, 95*(4), 796–805. http://dx.doi.org/10.1037/0022-0663.95.4.796.

Inzlicht, M., & Kang, S. K. (2010). Stereotype threat spillover: How coping with threats to social identity affects aggression, eating, decision making, and attention. *Journal of Personality and Social Psychology, 99*(3), 467–481. http://dx.doi.org/10.1037/a0018951.

Johnson, D. J., Cheung, F., & Donnellan, M. B. (2014). Hunting for artifacts: The perils of dismissing inconsistent replication results. *Social Psychology, 45*(4), 318–320.

Luhtanen, R., & Crocker, J. (1992). A collective self-esteem scale: Self-evaluation of one's social identity. *Personality and Social Psychology Bulletin, 18*(3), 302–318. http://dx.doi.org/10.1177/0146167292183006.

National Math and Science Initiative (2013). Closing the gender gap in STEM [Weblog post] Retrieved fromhttps://nms.org/Blog/TabId/58/PostId/83/closing-the-gender-gap-in-stem.aspx.

Nguyen, 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*(6), 1314–1334. http://dx.doi.org/10.1037/a0012702.

Noyes, J. M., & Garland, K. J. (2008). Computer- vs. paper-based tasks: are they equivalent? *Ergonomics, 51*(9), 1352–1375. http://dx.doi.org/10.1080/00140130802170387.

Roediger, H. L. (2012). Psychology's woes and a partial cure: The value of replication. *APS Observer, 25*(2), 9. Retrieved fromhttp://dev.psychologicalscience.org/index.php/publications/observer/2012/february-12/psychologys-woes-and-a-partial-cure-the-value-of-replication.html/comment-page-1#.VpxhQ5MrJPs.

Schmader, T. (2002). Gender identification moderates stereotype threat effects on women's math performance. *Journal of Experimental Social Psychology, 38*(2), 194–201. http://dx.doi.org/10.1006/jesp.2001.1500.

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*(2), 107–130. http://dx.doi.org/10.1177/1088868306294790.

Shapiro, J. R., & Williams, A. M. (2012). The role of stereotype threats in undermining girls' and women's performance and interest in STEM fields. *Sex Roles, 66*, 175–183. http://dx.doi.org/10.1007/s11199-011-0051-0.

Smith, J. L., Sansone, C., & White, P. H. (2007). The stereotyped task engagement process: The role of interest and achievement motivation. *Journal of Educational Psychology, 99*(1), 99–114. http://dx.doi.org/10.1037/0022-0663.99.1.99.

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. http://dx.doi.org/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. http://dx.doi.org/10.1037/0003-066X.52.6.613.

Steele, C. M., & Aaronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*(5), 797–811. http://dx.doi.org/10.1037/0022-3514.69.5.797.

Stoet, G., & Geary, D. C. (2012). Can stereotype threat explain the gender gap in mathematics performance and achievement? *Review of General Psychology, 16*(1), 93–102. http://dx.doi.org/10.1037/a0026617.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48http://www.jstatsoft.org/v36/i03/.

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

Wout, D., Danso, H., Jackson, J., & Spencer, S. (2008). The many faces of stereotype threat: Group- and self-threat. *Journal of Experimental Social Psychology, 44*(3), 792–799. http://dx.doi.org/10.1016/j.jesp.2007.07.005.