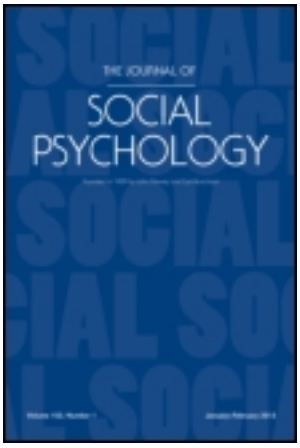

The Journal of Social Psychology

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/vsoc20

Exploring the Moderating Role of Context on the Mathematics Performance of Females Under Stereotype Threat: A Meta-Analysis

Katherine Picho [a], Ariel Rodriguez [b] & Lauren Finnie [b]

[a] Connecticut Institute for Clinical and Translational Science, University of Connecticut Health Center
[b] University of Hartford

Accepted author version posted online: 01 Nov 2012.Version of record first published: 20 Mar 2013.

To cite this article: Katherine Picho , Ariel Rodriguez & Lauren Finnie (2013): Exploring the Moderating Role of Context on the Mathematics Performance of Females Under Stereotype Threat: A Meta-Analysis, The Journal of Social Psychology, 153:3, 299-333

To link to this article: http://dx.doi.org/10.1080/00224545.2012.737380

The Journal of Social Psychology, 2013, *153*(3), 299–333

Exploring the Moderating Role of Context on the Mathematics Performance of Females Under Stereotype Threat: A Meta-Analysis

KATHERINE PICHO
Connecticut Institute for Clinical and Translational Science, University of Connecticut Health Center

ARIEL RODRIGUEZ
LAUREN FINNIE
University of Hartford

ABSTRACT. The current meta-analysis synthesized 17 years of research on stereotype threat (ST). Specifically, it examined the moderating effect of contextual factors on ST. Findings revealed that, on average, females in ST conditions performed less well on mathematics tests than their control counterparts ($d = |0.24|$). Results also showed that females did not benefit more from female-only testing situations, or testing contexts where they formed the majority. Nevertheless, the trend in ST effects differed by broader contextual factors like geography and level of education, with females in countries with small gender-gaps showing better performance under ST conditions, and ST effects being greater for students in middle and high school compared to college students.

Keywords: females, mathematics, meta-analysis, stereotype threat

THE NUMBER OF WOMEN IN SCIENCE, technology, engineering and mathematics (STEM) related fields has risen dramatically over the course of recent decades (Burelli, Arena, Shettle, & Fort, 1996). However, females remain underrepresented in these domains. According to the United States (U.S.) Department of Commerce women held less than 25% of STEM jobs in 2011 (Beede, Julian, Langdon, McKittrick, Khan, & Doms, 2011). The leaking pipeline has been attributed to gender gaps in math performance that start as early as middle school (Halpern, Benbow, Geary, Gur, Hyde, & Gernsbacher, 2007). Sex differences in

Address correspondence to Katherine Picho, Connecticut Institute for Clinical and Translational Science, University of Connecticut Health Center, Department of Research Design, Epidemiology & Biostatistics, 263 Farmington Ave., MC 6233, Farmington, CT 06030, USA; edpsychresearch@gmail.com (e-mail).

mathematics performance have been linked to innate differences in spatial ability (Baenninger & Newcombe, 1995; Levine, Huttenlocher, Taylor, & Langrock, 1999; Terlecki, Newcombe & Little, 2007), brain development, and hormonal differences (Ardila, Rosselli, Matute, & Inozemtseva, 2011; Halpern, 1992; Wilder & Powell 1989), as well as to socio-cultural factors like stereotype threat (Steele, 1997).

Stereotype threat (ST) is a social psychological phenomenon that inhibits the performance of members of stereotyped groups on difficult tasks in contexts where negative stereotypes about the ability of their group are highlighted (Steele, 1997). ST impacts individuals with a moderate to strong identification to the domain (Steele, 1997). For these individuals, simply an awareness of (and not necessarily a belief in) the negative stereotype about one's group (e.g., females are bad at math) is necessary for ST to occur (Steele, 1997). ST has been reported to negatively impact the performance of females on quantitative tests (Ambady et al., 2001; Huguet & Regner, 2007; Eriksson & Lindholm, 2007; Schmader, 2002), and non-Asian ethnic minorities (Armenta, 2010; Gonzales, Blanton, & Williams, 2002; Schmader & Johns, 2003; Steele & Aronson, 1995) on tests of verbal and quantitative ability. Most studies investigating the impact of ST on females' mathematics performance have shown that women perform less well in contexts where attention is drawn to negative gender stereotypes related to mathematical ability.

In recent years, ST has received a lot of attention both in research and in the general public as a significant factor explaining the performance of females in STEM subjects. Since Steele and Aronson's (1995) seminal work on ST, there has been a deluge of research and media coverage on the subject. Primary level studies investigating ST in females have found support for the moderating effects of psychological factors (e.g., gender identification, Schmader, 2002; stigma consciousness, Brown & Pinel, 2003) and contextual factors (e.g., sex composition, Inzlicht & Ben-Zeev, 2003; role models, Marx & Roman, 2002). However, the strength of these moderator variables in attenuating ST effects on the quantitative performance of females is not yet quite clear. Specifically, even though ST theory posits the phenomenon as being highly situational (Steele, 1997), the moderating role of context has not been adequately addressed, in both primary level studies and meta-analyses. The present meta-analysis sought to shed clarity upon these areas.

So far, four meta-analyses related to ST have been published (Nguyen & Ryan, 2008; Stoet & Geary, 2012; Walton & Cohen, 2003; Walton & Spencer, 2009), three of which exclusively examined between group differences on performance under ST conditions. Walton and Cohen (2003) found that members of non-stereotyped groups (i.e., males and Asians) performed better than their control counterparts when they were made aware of negative stereotypes surrounding out-group members (females, or non-Asian ethnic minorities). Walton and Cohen (2009) found a considerably moderate gender performance gap in favor of males ($k = 22$, Cohen's $d = 0.48$), while Stoet and Geary (2012) found ST to have

significant negative effects only in primary level studies that had used analysis of covariance (ANCOVA) to analyze data ($d = –0.61$, $p < .001$) and not non-ANCOVA studies ($d = –0.17$, $p = .09$). Because between-groups differences are not the focus of this article, interested readers are referred to Stoet & Geary, (2012), Walton and Cohen, (2003), and Walton & Spencer, (2009) for a more detailed review of the aforementioned.

Nguyen and Ryan (2008) examined the moderating effects of type of priming, test difficulty, type of stereotype threat removal strategy, and domain identification, on females and racial minorities. The authors' within-groups meta-analysis found a small combined ST effect (Cohen's $d = 0.26$), and a much smaller mean effect on women's performance ($d = 0.21$). Like Nguyen and Ryan's study, the current within-groups meta-analysis examined the moderating effect of priming on the performance of females under ST. However, in the present study, new studies were added to the analysis of this moderator (the overlap between both meta-analyses on this moderator was only 15 studies). The major difference between the two meta-analyses lies with how the effect sizes (ES) were computed. In computing ES, the researchers treated ST removed (STR) conditions (conditions where ST was nullified prior to test taking) sometimes as controls and other times as experimental groups. The lack of consistency in how STR conditions were treated (i.e., alternating STR conditions as either control or experimental conditions depending on the type of study) could have confounded both the magnitude and interpretation of the effect sizes. The present study remedied this problem by consistently treating the STR conditions as control groups.

Additionally, the current study also extended previous ST meta-analyses by examining contextual factors not previously investigated and focusing solely on comparing the performance of females in treatment and control groups in quantitative or visuo-spatial tasks where the literature indicates sex bias in favor of males.

ST Moderators Investigated in the Present Study

The goal of the present study was to investigate the moderating effects of contextual factors on females under ST. More specifically, for females exposed to math-related ST, we sought to address the following questions: (a) Is the impact of ST greater in testing contexts where gender stereotypes are overt or covert?; (b) Is ST moderated by sex composition?; and finally (c) for U.S. samples, do ST effects vary by region?

Priming as a moderator of ST. In ST experiments, the nature of the testing environment has often been manipulated by activating ST through priming. Priming, which is defined as the activation of knowledge structures (e.g., stereotypes) in a situational context (Bargh, 1994), can be done explicitly by drawing one's

attention directly to the stereotype, or implicitly by activating stereotypes in more subtle forms. The broader literature on automaticity of behavior reports that priming exerts a passive influence on behavior (Bargh, 1994; Higgins, 1989). Bargh and colleagues (1996) implicitly activated stereotypes of the elderly by asking participants to solve a scrambled-sentence task containing words relevant to the elderly stereotype. They then measured how long it took participants to walk down a hallway after completing the task and found that those who had been implicitly primed with the elderly stereotype walked much more slowly compared to their control counterparts.

However, studies also show that priming outcomes might vary depending on whether it is explicit or implicit. According to reactance theory (Brehm, 1966), individuals assert their freedom more forcefully when they perceive a threat to it. Empirical studies have found that blatantly drawing one's attention to a negative stereotype pertinent to one's group tends to produce a contrast (reactance) effect where individuals engage in behavior that contradicts the stereotype (Moskowitz & Skurnik, 1999). To the contrary, subtly directing one's attention to a negative stereotype through implicit priming has been found to produce the opposite effect (assimilation). Here, stereotyped individuals engage in behaviors that are consistent with the stereotype (Moskowitz & Skurnik, 1999). Other studies have also demonstrated that when faced with negative gender stereotypes regarding their performance in stereotypically masculine domains, women acted in ways that disconfirmed the stereotype when it was blatantly (as opposed to implicitly) expressed (Kray, Thompson, & Galinsky, 2001).

In ST experiments, explicit priming has often involved mentioning that the test was diagnostic of ability (Gonzales, Blanton, & Williams, 2002; Good, Aronson, & Harder, 2008; Schmader, 2002), that there were gender differences on the quantitative test being administered (Armenta, 2010; Cadinu, Maas, Rosabianca, Figerio, & Latinotti, 2003; Cadinu, Maas, Rosabianca, & Kiesner, 2006; Elizaga & Markman, 2008; Picho & Stephens, 2012), or that the test was being administered in order to understand gender differences in quantitative performance/ability (Schmader, Johns, & Barquissau, 2004). Implicit priming for gender ST, on the other hand, has often involved the use of subtle cues to activate gender stereotypes, like asking young girls to color pictures of a doll or an Asian holding chopsticks (Huguet & Regner, 2007), or requesting young women to indicate their gender prior to taking a math test (Steele & Aronson, 1995). Jamieson and Harkins (2012) implicitly primed for gender by asking female participants to write about a day in the life of a female Northeastern University student named "Ashley" for 5 minutes before taking a standardized mathematics test. They found that gender prime participants answered fewer comparison problems correctly than their control-group counterparts.

Based on studies that have demonstrated reactance and assimilation effects as a function of type of priming in task performance among members of stigmatized groups, we hypothesized that samples where implicit priming had been

used to activate ST would generate larger negative effects than samples where ST had been primed for explicitly. Specifically, because of the tendency of implicit priming to produce assimilation effects among members of stereotyped groups, we hypothesized that females subjected to implicit priming would act in ways that confirmed the negative gender stereotype, and hence exhibit larger negative ST effects on math performance, compared to samples where explicit priming had been used.

Sex composition. Studies have shown that some environments are more capable of enhancing ST susceptibility among females than others. Research has found exaggerated ST effects in testing contexts where females were the minority (Inzlicht & Ben-Zeev, 2000; Murphy, Steele, & Gross, 2007). Research by Murphy and colleagues (2007) revealed that factors like a setting's features and numerical representation made females more vulnerable to ST. Similarly, when Inzlicht and Ben-Zeev (2000) varied the sex composition of the experimental groups prior to eliciting ST, the average performance of females, not males, varied significantly as a function of group composition. Specifically, females performed best in homogeneous versus heterogeneous settings; they also performed better when they were the majority in heterogeneous contexts and worse when they were the minority in heterogeneous contexts. Picho and Stephens (2012) also reported differential ST effects among high school female students in coed and single sex schools in Uganda, with non- significant ST effects for females in the single sex school, but large effects for females in the co-ed school ($d = 0.11$ *vs.* 0.74). Huguet and Regner (2009) demonstrated these same gender effects in middle-school-aged children; negative ST effects were only present in mixed gender testing conditions. Gneezy, Niederle, and Rustichini (2003) also found that women performed worse in competitive environments only when they believed they were competing against men. They also found that women performed equally to men when they were in non-competitive environments or competitive environments where they believed that they were only competing with other women. These results demonstrate the active, though possibly implicit "probability weighing" that occurs in ST conditions; that is, women determine whether or not they might be negatively stereotyped and these determinations might mediate ST effects. Based on these and other similar studies (Sekaquaptewa & Thompson, 2003), we hypothesized more negative effects for females in testing conditions where they formed the minority and smaller ST effects in homogeneous samples, or samples where females formed the majority.

Region as a potential moderator of ST. ST is situational, occurring in contexts where negative stereotypes about a particular stigmatized group are made salient (Steele, 1997). Although gender stereotypes in ST experiments are activated through priming, these stereotypes most likely reflect stereotypes already present

within the broader socio-cultural societal context. In this broader context, stereotypes appear to arise, in part, from societal expectations regarding gender roles within a given society. According to social role theory (Eagly, 1987), gender roles are based on shared expectations about what members of a group actually do (descriptive norms) and/or should do (injunctive norms). It follows, then, that gender stereotypes would be stronger in cultures where gender roles were more distinct.

Sociologists in the United States (U.S.) have conducted a plethora of research investigating differences between northern and southern cultures, particularly as it relates to gender roles. Results from these studies have found evidence to suggest that: (a) compared to other regions in the U.S., Southerners are more traditional in their attitudes toward gender roles (Rice & Coates, 1995; Twenge, 1997), especially regarding women in stereotypically masculine vocations like politics (Rice & Coates, 1995); and (b) behavioral expectations (injunctive norms) regarding gender are more clearly defined and culturally prescribed for southern women than they are for men (Suitor & Carter, 1999). Some of these studies have also found differences between and within regions as a function of gender and race. For example, Levant, Majors, and Kelley (1998) found that regardless of region, African American women were significantly more traditional than their European American counterparts. However, African American men in the Northeastern-Mid-Atlantic region had less strong attitudes toward gender roles than their Southern counterparts. Further, within the South itself, African American men were more traditional than European American men. Differences in gender role attitudes have also been found to vary by sub regions, with individuals residing in rural areas maintaining more traditional gender-role attitudes than those in urban areas (Twenge, 1997).

To the best of our knowledge, there have been no primary level studies investigating differential ST effects by region in the U.S. As such, we draw from sociological research to base our hypotheses. Research in this area has found more liberal attitudes towards gender roles in the north compared to the south therefore we hypothesized smaller ST effects in the northeast versus the southern U.S. region. Second, although the majority of this research hardly focused on any other regional differences in gender roles besides north-south regions, we postulated that ST effects would be generally smaller in regions considered to be more liberal regarding to attitudes towards gender roles.

Method

Literature Search

A literature search for ST articles was conducted using major databases including Psycinfo, Proquest, ERIC, web of science, and Ebscohost. Keywords used in conjunction with the words *stereotype threat* were: *women*, *race*,

minorities, and *mathematics*. Unpublished studies were obtained by sending email requests to several listservs, like SEMNET, and the European Association of Social Psychology (EASP). We also attempted to obtain unpublished studies by cold-emailing key researchers in the field for unpublished manuscripts and checking online repositories for conference papers from major Psychology and Education organizations.

Inclusion-exclusion criteria. The initial search yielded 450 articles from 1995–2011. From this pool, studies were selected if: (a) the research design was experimental or quasi-experimental, with results reported for control and experimental groups; (b) women were included in the sample; and (c) the dependent variable reported performance on a quantitative test. Based on these criteria, 170 studies (published, unpublished, and dissertations) were retained. As shown in Figure 1, the retained studies were excluded from the coding process if: (a) control groups did not meet the criteria for this study (i.e., studies where ST was activated in both groups, or studies where two control groups (pure and STR) were compared; (b) no clear ST- priming or activation in experimental groups was reported; (c) scores on the dependent variable were standardized, and raw scores could not be obtained from authors; and (d) means, standard deviations, *n*'s and other metrics on the dependent variable needed to compute effect sizes were not provided and could not be obtained from the authors. Given the focus of the meta-analysis on examining ST effects, ST intervention studies were also excluded. Finally, to avoid duplication of data, journal articles published from dissertations that had also been retained for the meta-analysis were excluded. Here, articles (not dissertations) were eliminated because the latter provided more detailed information on sample characteristics and the nature of the study. This yielded a final sample *k* of 103 independent studies nested in 44 articles and dissertations.

Coding Study Moderators

An elaborate coding form developed by the first author was used to assess sample characteristics, ST moderators that were previously outlined, methodological features of ST studies (e.g., random assignment, domain identification as eligibility criteria for participation in ST studies), and features of the experimental design (e.g., experimenter characteristics, test length, psychometric properties of these instruments measuring ST moderators). A scale was also developed to rate studies on methodological quality. The authors coded all studies independently and met weekly to check reliability for all coded studies. Any areas of disagreement in the coding process were resolved by discussion. Cohen's Kappa's corrected formula was used to calculate reliability; the average reliability estimate for both coders was .87.

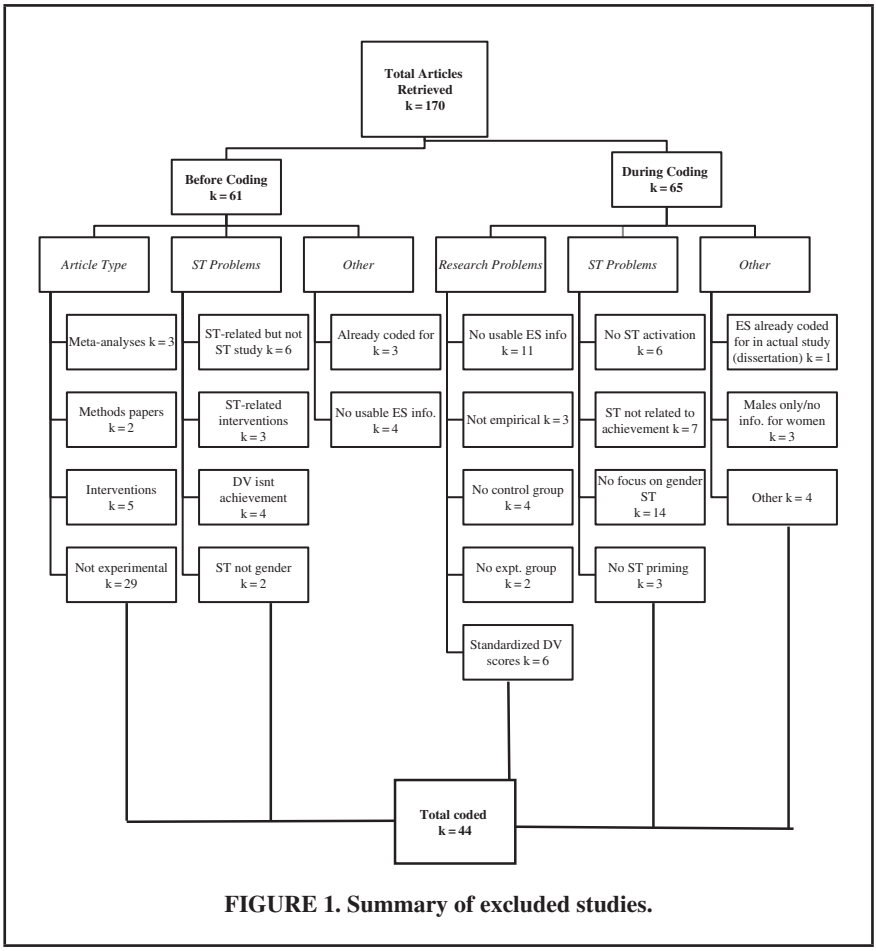

FIGURE 1. Summary of excluded studies.

Coding sex composition. For homogeneous (female-only) samples, effect sizes were computed using statistics presented for control and experimental groups. For heterogeneous (mixed sex) samples, only the means and standard deviations for females in control and experimental conditions were used to compute effect sizes. In the majority of studies using heterogeneous samples, the number of females in each condition was provided. However, in the absence of this information, if a study reported (a) the total number of participants who completed the study by gender and (b) also reported using random assignment of participants into experimental groups, then equal numbers of females were assumed for both groups. When the total number of females completing the study was an odd number, the number of participants in the control group sample was greater than

the experimental group by one. We also coded for the proportion of women in mixed sex samples and used this as a continuous variable in subsequent analyses presented in the results section.

Coding testing environment. Priming was coded as implicit if ST was subtly activated by indirect statements about females' ability in math—e.g., having participants indicate their gender on tests that were reportedly diagnostic of ability (Schmader, 2002; Steele & Aronson, 1995). Two categories for explicit priming were coded: (a) blatant group differences: Studies fell under this category if direct statements about gender differences in math abilities or performance (e.g., Keller & Molix, 2008; Lesko & Corpus, 2006; Schmader & Johns, 2003) had been made; and (b) blatant diagnostic of ability: for studies where participants had been told that the test was diagnostic of personal ability in the domain being assessed (Steele & Aronson, 1995; Wout, Danso, & Jackson, 2009; Scholten, et al., 2005).

Coding U.S. regions. We used 2010 census classification for geographical regions (Northeast, Midwest, South and West) to categorize the states where studies had been conducted. This information was located in the "participants" section of primary level studies where either the college of participants or the geographical region was identified by the researchers. For instance, studies that reported using participants from a "large Midwestern university" were categorized under 2010 census region "Midwest" and so on.

Coding independent data points. The majority of articles reported results of more than one ST experimental study. In all cases, the samples were independent, so effect sizes were calculated separately for each study. In instances where a particular study had one control but two or more experimental groups, effect sizes were computed separately for each experimental group against the control group.

Coding dependent data points. For repeated measures study designs like Dinella (2004), effect sizes and corresponding variances were computed using Wilson's effect size calculator.

Coding ST moderator studies with factorial designs. Studies with two or three factorial designs were split into two or three independent groups respectively, depending on the moderators being measured (as defined by the original researchers themselves). For example, Ambady, Paik, Steele, Owen-Smith, and Mitchell's (2004) 2 (condition- control vs. experimental) $\times$ 2 (construal- individuation vs. non individuation) design was coded for two effect sizes—control

and experimental group for the individuated condition, and the same for the non-individuated condition.

Coding studies where gender was nested in race. The study by Gonzales et al. (2002) reported ST results by condition for each gender (male/female) and race (Whites, Blacks and Hispanics). The study also provided *n*'s for each group, therefore effect sizes were coded separately for samples of White, Black and Hispanic females. The same technique was applied in other studies where gender was nested in race.

Meta-Analysis Procedure

The corrected standardized mean difference, Cohen's *d*, its variance (vard), and weighted variance were computed using an effect size calculator (Huedo-Medina & Johnson, 2011). Cohen's *d* effect size (ES) was calculated by dividing the mean difference by the pooled standard deviation. The mean difference was obtained by subtracting mean scores of the control group from that of the experimental group such that a negative *d* implied better performance by the control group and vice versa for positive effect sizes.

Fixed vs. random effects approach to meta-analysis. Fixed and random effects computational models for meta-analyses are driven by different assumptions regarding the meta-analytic data. The fixed effects approach assumes that all the primary studies included in the meta-analysis come from a single population and therefore are functionally identical, sharing a common ES (Borenstein, Hedges, Higgins, & Rothstein, 2009). Consequently, any differences between studies in ES are attributed to sampling error. On the other hand, the random effects approach assumes that the primary studies included in the meta-analyses do not represent the entire population of studies, but rather are samples from the population of studies (Kalaian & Kasim, 2008). Hence there are differences between studies in ES beyond those due to sampling error (Kalaian & Kasim, 2008).

We used stereotype threat theory (STT) to guide our assumptions about ST effect sizes and subsequently our selection of a computational meta-analytic model. STT conceptualizes ST as a highly situational phenomenon (Steele, 1997), and primary level studies have also demonstrated differential ST effects for samples from different populations e.g., STEM vs. non-STEM students (Crisp, Bache, & Maitner, 2009; Werhun, 2007). As such, the notion that the studies included in the current meta-analysis represented an entire population of ST studies with a shared common effect size seemed unlikely. We therefore conducted moderator analysis in the random effects framework in the statistical analysis software Stata 10.0.

Results

Results were based on a total sample of 5,588 females, 2,820 of who were in control groups. We applied Cohen's (1992) useful guidelines on interpreting standardized ES, where d's = 0.2, 0.5 and 0.8 are small, medium, and large, respectively. Study effect sizes and moderators investigated in the present study are shown in Table 1. The overall mean ST effect (mean ES) for all studies was small, $d = -0.24$ (CI_{95}: -0.35, -0.14) with a between study variance of 0.21 ($p < .001$). The Q-statistic revealed between study heterogeneity (χ^2 (*102*) = 378.6, $p < .001$), and the I^2 statistic (73.3%) indicated far more variation in the distribution of effect sizes than sampling error alone would predict.

Publication Bias

Publication bias refers to concerns over the validity of systematic reviews that have included few or no unpublished studies. The premise of this phenomenon is that meta-analytic findings that hardly contain unpublished studies might be exaggerated (with a bias towards larger or positive treatment effects) because studies reporting statistically significant results are more likely to be published than those that are not (Borenstein, Hedges, Higgins, & Rothstein, 2009). We conducted Egger's regression test that revealed small study bias of which publication bias might be a potential cause ($p = .02$). In spite of having a significantly larger number of published than unpublished studies in this review, we remain cautious about interpreting the results of Egger's test as conclusive evidence of publication bias for two reasons: first, asymmetry tests like the aforementioned tend, and have been shown, to generate false positives for publication bias in the presence of large, statistically significant between- studies heterogeneity (Ioannidis & Trikolos, 2007), which was the case in this study. Second, the distribution of effect sizes in the funnel plot shown in Figure 2 revealed just as many studies with non-significant results reported in the published literature, as there were those with significant results. This suggested that preferential publication of studies based on statistical significance might be unlikely.

Nevertheless, to judge the robustness of results against publication bias, fail-safe numbers, which indicate the number of unpublished studies required to reduce the ES to non-significance (Rosenberg, 2005), were calculated. It should be noted that fail-safe N's do not determine whether meta-analytic results are correct, but rather gauge whether publication bias may be safely ignored (Rosenberg, 2005). Here, large numbers imply more robust results, with the reverse being true for low numbers.

Table 2 shows descriptive statistics obtained using Stata 10.0. As seen in the table, the majority of ST studies were conducted in the United States ($k = 63$), and most studies used samples from the college population ($k = 88$), particularly with students in non-STEM related domains ($k = 98$). Also, the majority of study samples were homogeneous or females-only samples ($k = 70$).

TABLE 1. Studies Analyzed in the Meta-Analysis ($k = 103$)

Study no.	Author (s)	Sample	Sample size	d	Sex comp.	Prop. women	Priming	Country	U.S. region
1	Ambady, Paik, Steele, et al. (2004), Study 1a	Caucasian undergrads. One grad student	20	0.43	Single sex	1	Implicit	USA	Northeast
2	Ambady, Paik, Steele, et al. (2004), Study 1a	Caucasian undergrads. One grad student	20	−0.86	Single sex	1	Implicit	USA	Northeast
3	Ambady, Paik, Steele, et al. (2004), Study 1b	Caucasian undergrads	20	−0.83	Single sex	1	Implicit	USA	Northeast
4	Ambady, Paik, Steele, et al. (2004), Study 1b	Caucasian undergrads	19	0.53	Single sex	1	Implicit	USA	Northeast
5	Armenta (2010)	Asian American undergrads	26	1.88	Mixed sex	.57	Explicit	USA	West
6	Armenta (2010)	Hispanic undergrads	26	−0.82	Mixed sex	.57	Explicit	USA	West
7	Beaton, Tougas, Rinfret, et al. (2007), Study 1	Math identified French Canadian undergrads	44	0.17	Single sex	1	Implicit	Canada	n/a
8	Beaton, Tougas, Rinfret, et al. (2007), Study 2	Math identified French Canadian undergrads	45	−0.11	Single sex	1	Implicit	Canada	n/a
9	Bell, Spencer, Iserman, & Logel (2003)	Engineering undergrads	18	−0.59	Mixed sex	.60	Explicit	USA	South
10	Brodish (2007), Study 1	High gender identified undergrads with strong math skills	88	−0.40	Single sex	1	Explicit	USA	Midwest
11	Brodish (2007), Study 2	High gender identified undergrads with strong math skills	58	−0.33	Single sex	1	Explicit	USA	Midwest
12	Brodish & Devine (2009), Study 1	High math and gender identified undergrads	47	−0.35	Single sex	1	Implicit	N/R	Midwest

13	Brodish & Devine (2009), Study 2	High math and gender identified undergrads	47	−0.47	Single sex	1	Implicit	N/R	Midwest
14	Cadinu, Maass, Frigerio, et al. (2003), Study 1	Psychology students	26	−1.08	Single sex	1	Explicit	Italy	n/a
15	Cadinu, Maass, Frigerio, et al. (2003), Study 2	Psychology students	38	0.17	Single sex	1	Explicit	Italy	n/a
16	Cadinu, Maass, Lombardo, et al. (2006), Study 1	External locus of control high school students	36	−0.36	Mixed sex	.52	Explicit	Italy	n/a
17	Cadinu, Maass, Lombardo, et al. (2006), Study 2	Internal locus of control high school students	42	−1.33	Mixed sex	.52	Explicit	Italy	n/a
18	Cadinua, Maass, Rosabianca, et al. (2005)	Psychology students	60	−0.54	Single sex	1	Explicit	Italy	n/a
19	Campbell & Collaer (2009), Study 1	Undergrad students	70	0.04	Mixed sex	.55	Explicit	USA	Northeast
20	Campbell & Collaer (2009), Study 2	Undergrad students	70	−0.01	Mixed sex	.55	Explicit	USA	Northeast
21	Clark, Eno & Guadagno (2011)	U.S. Southern-born undergrads	41	−0.65	Mixed sex	.89	Explicit	USA	South
22	Cotting (2003), Study 1	Undergrads at an all women college/ university	51	−0.13	Single sex	1	Implicit	USA	Northeast
23	Cotting (2003), Study 2	Undergrads at a historically Black college/university	55	−0.35	Mixed sex	.78	Implicit	USA	Northeast
24	Crisp, Bache & Maitner (2009), Study 1	Engineering undergrads	38	0.71	Single sex	1	Implicit	UK	n/a
25	Crisp, Bache & Maitner (2009), Study 2	Psychology undergrads	40	−0.54	Single sex	1	Implicit	UK	n/a
26	Crisp, Bache & Maitner (2009), Study 3	Engineering undergrads	40	0.44	Single sex	1	Implicit	UK	n/a

(Continued)

TABLE 1. (Continued)

Study no.	Author (s)	Sample	Sample size	*d*	Sex comp.	Prop. women	Priming	Country	U.S. region
27	Crisp, Bache & Maitner (2009), Study 4	Psychology undergrads	40	−0.47	Single sex	1	Implicit	UK	n/a
28	Croizet, Despres, Gauzins, et al. (2004), Study 1	Psychology students	69	−0.72	Mixed sex	.90	Explicit	France	n/a
29	Croizet, Despres, Gauzins, et al. (2004), Study 2	Science students	69	0.66	Mixed sex	.90	Explicit	France	n/a
30	Davies, Spencer, Quinn, et al. (2002)	High math identified undergrads	25	−0.84	Single sex	1	Implicit	Canada	n/a
31	Dinella (2004)	High math and gender identified high school students	266	0.25	Mixed sex	.53	Explicit	USA	West
32	Eriksson & Lindholm (2007)	Swedish, high gender identified, math and engineering undergrads	112	0.33	Single sex	1	Explicit	Sweden	n/a
33	Ford, Ferguson, Brooks, et al. (2004)	Undergrads in Sociology courses	31	0.71	Single sex	1	Explicit	USA	Midwest
34	Gonzales, Blanton & Williams (2002), Study 1	High ethnically identified Hispanics	30	−1.31	Mixed sex	.50	Explicit	USA	Northeast
35	Gonzales, Blanton & Williams (2002), Study 2	High ethnically identified Caucasians	30	1.71	Mixed sex	.50	Explicit	USA	Northeast
36	Good, Aronson & Harder (2007)	STEM undergrads	57	−0.24	Mixed sex	.36	Explicit	USA	South
37	Huguet & Regner (2007), Study 1	High math identified middle school students	20	−0.68	Mixed sex	.50	Implicit	France	n/a

38	Huguet & Regner (2007), Study 1	High math identified middle school students.	178	−0.50	Mixed sex	.49	Implicit	France	n/a
39	Huguet & Regner (2007), Study 2	High math identified middle school students.	227	0.03	Mixed sex	.49	Implicit	France	n/a
40	Jamieson (2009), Study 1	Undergrads	32	−0.79	Mixed sex	.50	Explicit	USA	Northeast
41	Jamieson (2009), Study 4	Undergrads	36	−0.63	Mixed sex	.50	Implicit	USA	Northeast
42	Jamieson (2009), Study 2	Undergrads	120	−0.79	Single sex	1	Explicit	USA	Northeast
43	Jamieson (2009), Study 3	Undergrads	31	−0.19	Single sex	1	Explicit	USA	Northeast
44	Jamieson (2009)	Undergrads	31	−2.09	Single sex	1	Explicit	USA	Northeast
45	Jamieson (2009)	Undergrads	41	−0.69	Mixed sex	.68	Implicit	USA	Northeast
46	Johns, Schmader, & Martens (2005)	Introductory statistics students	50	−1.02	Mixed sex	.64	Explicit	USA	West
47	Keller (2002)	High school students	37	−0.40	Mixed sex	.49	Explicit	Germany	n/a
48	Keller & Dauenheimer (2003)	High school students	35	−0.49	Mixed sex	.51	Explicit	Germany	n/a
49	Lesko & Corpus (2006)	Undergrads at a selective liberal arts college	68	−0.61	Mixed sex	.56	Explicit	USA	West
50	McIntyre, Paulson, Lord, et al. (2010), Study 1	Undergrads	60	−0.01	Single sex	1	Explicit	USA	South
51	McIntyre, Paulson, Lord, et al. (2010), Study 2	Undergrads	60	−0.59	Single sex	1	Explicit	USA	South
52	McIntyre, Paulson, Lord, et al. (2010), Study 3	Undergrads	60	−0.68	Single sex	1	Explicit	USA	South
53	Oswald & Harvey (2001), Study 1	Undergrads in hostile testing environment	34	−0.70	Single sex	1	Implicit	USA	Midwest
54	Oswald & Harvey (2001), Study 2	Undergrads in non-hostile environment	38	−0.53	Single sex	1	Implicit	USA	Midwest

(Continued)

TABLE 1. (Continued)

Study no.	Author (s)	Sample	Sample size	*d*	Sex comp.	Prop. women	Priming	Country	U.S. region
55	Picho & Stephens (2012), Study 1	African high school students at a coed school	38	−0.74	Mixed sex	.40	Explicit	Uganda	n/a
56	Picho & Stephens (2012), Study 2	African high school students at a single sex-school	51	−0.14	Single sex	1	Explicit	Uganda	n/a
57	Quinn & Spencer (2001)	Undergrad students	18	−0.67	Mixed sex	NR	Explicit	USA	Midwest
58	Rivardo, Rhodes, Camaione, et al. (2011)	Undergrads who passed advanced placement in Calculus 1	39	0.88	Mixed sex	.47	Explicit	USA	Northeast
59	Rosenthal & Crisp (2006), Study 1	Psychology undergrads	32	1.14	Single sex	1	Explicit	UK	n/a
60	Rosenthal & Crisp (2006), Study 2	Psychology undergrads	24	0.35	Single sex	1	Implicit	UK	n/a
61	Rosenthal & Crisp (2006), Study 3	Psychology undergrads	24	1.17	Single sex	1	Implicit	UK	n/a
62	Rosenthal & Crisp (2006), Study 4	Psychology undergrads	32	0.36	Single sex	1	Explicit	UK	n/a
63	Rosenthal & Crisp (2006), Study 5	Psychology undergrads	32	0.47	Single sex	1	Explicit	UK	n/a
64	Rucks (2008), Study 1	Math identified undergrads	35	0.61	Single sex	1	Explicit	USA	Midwest
65	Rucks (2008), Study 2	Math identified undergrads	35	−0.03	Single sex	1	Explicit	USA	Midwest
66	Rucks (2008), Study 3	Math identified undergrads	35	−0.31	Single sex	1	Explicit	USA	Midwest

67	Rydell, Beilock & McConnell (2009), Study 1a	Undergrads	56	−1.36	Single sex	1	Explicit	USA	Midwest
68	Rydell, Beilock & McConnell (2009), Study 1b	Undergrads	56	0.01	Single sex	1	Explicit	USA	Midwest
69	Rydell, Beilock & McConnell (2009), Study 1c	Undergrads	56	−0.21	Single sex	1	Explicit	USA	Midwest
70	Rydell, Beilock & McConnell (2009), Study 2a	Undergrads	50	−1.28	Single sex	1	Implicit	USA	Midwest
71	Rydell, Beilock & McConnell (2009), Study 2b	Undergrads	50	0.05	Single sex	1	Explicit	USA	Midwest
72	Rydell, Beilock & McConnell (2009), Study 2c	Undergrads	50	0.05	Single sex	1	Explicit	USA	Midwest
73	Rydell, Beilock & McConnell (2009), Study 3a	Undergrads	29	−0.37	Single sex	1	Explicit	USA	Midwest
74	Rydell, Beilock & McConnell (2009), Study 3b	Undergrads	29	−1.02	Single sex	1	Explicit	USA	Midwest
75	Rydell, Beilock & McConnell (2009) Study 3c	Undergrads	29	−0.21	Single sex	1	Explicit	USA	Midwest
76	Rydell, Beilock & McConnell (2009), Study 4a	Undergrads	40	−0.03	Single sex	1	Implicit	USA	Midwest

(Continued)

TABLE 1. (Continued)

Study no.	Author (s)	Sample	Sample size	*d*	Sex comp.	Prop. women	Priming	Country	U.S. region
77	Rydell, Beilock & McConnell (2009), Study 4b	Undergrads	40	−1.39	Single sex	1	Implicit	USA	Midwest
78	Rydell, Beilock & McConnell (2009), Study 4c	Undergrads	40	−0.17	Single sex	1	Implicit	USA	Midwest
79	Schmader & Johns (2003), Study 1	Psychology undergrads	28	−0.05	Mixed sex	.47	Explicit	USA	West
80	Schmader & Johns (2003), Study 3	Psychology undergrads	28	−0.88	Single sex	1	Implicit	USA	West
81	Schmader, Johns & Barquissau (2004), Study 1	STEM undergrads	240	−0.09	Mixed sex	.55	Explicit	Netherlands	n/a
82	Schmader, Johns & Barquissau (2004), Study 2	STEM undergrads	229	−0.1	Mixed sex	.55	Explicit	Netherlands	n/a
83	Steele & Aronson (1995), Study 2	White undergrads	20	0.45	Single sex	1	Explicit	USA	West
84	Steele & Aronson (1995), Study 2	Black undergrads	20	−0.79	Single sex	1	Explicit	USA	West
85	Steele & Aronson (1995), Study 4	Black undergrads	22	−0.76	Mixed sex	.64	Implicit	USA	West
86	Taylor, Lord, McIntyre, & Paulson (2011)	Undergrads	76	0.18	Single sex	1	Explicit	USA	South
87	Taylor, Lord, McIntyre, & Paulson (2011)	Undergrads	76	−0.58	Single sex	1	Explicit	USA	South

88	Taylor, Lord, McIntyre & Paulson (2011)	Undergrads	76	−0.60	Single sex	1	Explicit	USA	South
89	Thoman, White, Yamawaki, et al. (2008), Study 1	Math identified undergrads	47	0.50	Single sex	1	Explicit	USA	West
90	Thoman, White, Yamawaki, et al. (2008), Study 2	Math identified undergrads	44	0.16	Single sex	1	Explicit	USA	West
91	Weger, Hooper, Meir & Hopthrow (2012)	Psychology students	36	−0.06	Single sex	1	Explicit	UK	n/a
92	Werhun (2007), Study 2	Undergrads in moderate math intensive majors	20	−1.06	Single sex	1	Explicit	Canada	n/a
93	Werhun (2007), Study 2	Math identified STEM students	24	0.43	Single sex	1	Explicit	Canada	n/a
94	Werhun (2007), Study 2	Non-STEM undergrads	28	0.22	Single sex	1	Explicit	Canada	n/a
95	Werhun (2007), Study 3	Undergrads in moderate math intensive majors	67	−0.45	Single sex	1	Explicit	Canada	n/a
96	Werhun (2007), Study 3 STEM (HCS)	STEM undergrads	65	0.45	Single sex	1	Explicit	Canada	n/a
97	Werhun (2007), Study 3	Non-STEM undergrads	71	0.43	Single sex	1	Explicit	Canada	n/a
98	Wout, Danso, Jackson, et al. (2008), Study 1a	Undergrads	30	−1.06	Single sex	1	Explicit	USA	Midwest
99	Wout, Danso, Jackson, et al. (2008), Study 1b	Undergrads	29	−0.57	Single sex	1	Explicit	USA	Midwest
100	Wout, Danso, Jackson, et al. (2008) Study 1c	Undergrads	28	−0.73	Single sex	1	Explicit	USA	Midwest

(Continued)

TABLE 1. (Continued)

Study no.	Author (s)	Sample	Sample size	*d*	Sex comp.	Prop. women	Priming	Country	U.S. region
101	Wout, Danso, Jackson, et al. (2008), Study 2a	Undergrads	37	−0.75	Single sex	1	Explicit	USA	Midwest
102	Wout, Danso, Jackson, et al. (2008), Study 2b	Undergrads	37	−0.80	Single sex	1	Explicit	USA	Midwest
103	Zand Sholten, Wicherts, Elsenburg et al. (2005)	Psychology undergrads	254	0.08	Mixed sex	.68	Explicit	Netherlands	n/a

Note: N/A = Not applicable, N/R = Not reported.

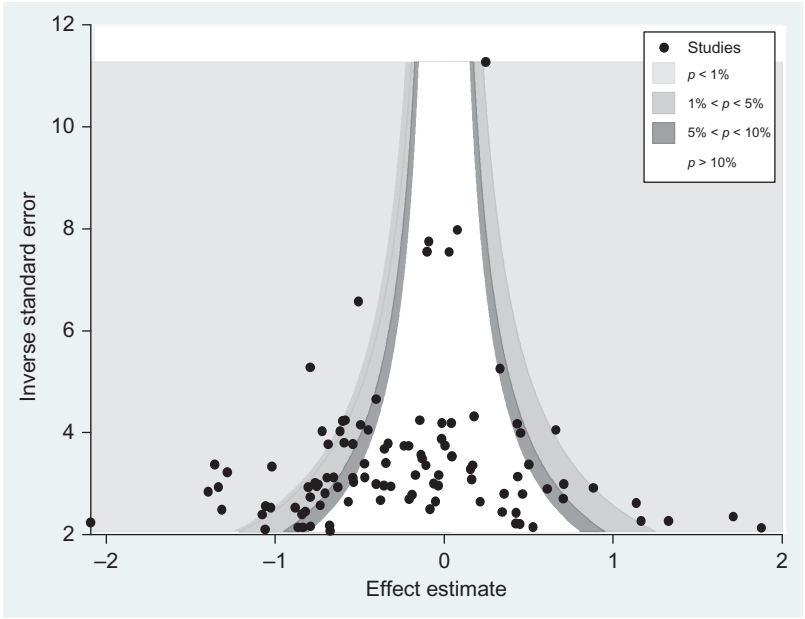

FIGURE 2. Funnel plot of ST studies (color figure available online).

ST Moderator Analyses

Moderator meta-analyses were conducted separately for sex composition, U.S. region, and testing environment (see table 3). 95% confidence intervals about study ES were also computed. For any given category confidence intervals that did not include zero indicated 95% confidence that the mean ST effect (denoted by ES) was not zero. Confidence intervals including zero indicated undependable ST effects.

Is the impact of ST greater in testing contexts where gender stereotypes are overt or covert? As shown in Table 3, mean ST effects were greater for samples tested in contexts where negative stereotypes were covert ($d = –0.28$) versus overt ($d = -0.23$). The variance about the ES in overt stereotype environments (denoted by explicit priming) was much larger than that of samples tested in contexts with covert stereotypes. This indicated greater variability in ES for studies in the former category. The non-overlap of confidence intervals with zero for both categories suggested sturdy ST effects but the difference in ES between these two conditions was not statistically significant ($p > .05$).

TABLE 2. Study Characteristics

Sample characteristic	*k*	*d*	Var_d	95 %	CI_{95}
Overall mean	103	−0.24	0.21	−0.35	−0.14
Education level					
College	88	−0.24	0.26	−0.37	−0.11
High school	9	−0.30	0.11	−0.57	−0.02
Middle school	3	−0.30	0.11	−0.76	0.15
Elementary school	1	−0.14	–	−0.61	0.32
Academic domain					
STEM	6	−0.06	0.36	−0.61	0.47
Non-STEM	95	−0.25	0.20	−0.36	−0.14
Math identification used as selection criteria					
No	89	−0.25	0.25	−0.38	−0.13
Yes	14	−0.16	0.04	−0.33	0.01
Region					
North America	73	−0.29	0.24	−0.42	−0.16
East Africa	2	−0.41	0.09	−1.01	0.18
Western Europe	16	0.01	0.11	−0.20	0.23
Southern Europe	5	−0.61	0.22	−1.11	−0.10
Northern Europe	1	0.33	–	−0.04	0.71

Note. Some studies did not report required information, resulting in missing values (and discrepancy in *k*'s) for select categories. Var_d = variance of mean effect size (ES).

TABLE 3. Moderator Analyses Results

Variable	*k*	*d*	Var_d	95	CI	Fail safe *N*
Overall	103	−0.23	0.21	−0.35	−0.14	157
Sex composition						
Single sex	70	−0.22	0.22	−0.37	−0.10	48
Mixed sex	33	−0.26	0.19	−0.43	−0.09	8
Priming						
Explicit	74	−0.23	0.22	−0.36	−0.10	42
Implicit	29	−0.28	0.14	−0.46	−0.10	16
U.S. region						
Northeast	17	−0.29	0.42	−0.65	0.06	0
Midwest	28	−0.42	0.14	−0.61	−0.24	41
South	8	−0.38	0.05	−0.62	−0.15	3
West	12	−0.16	0.36	−0.55	0.24	0

Note. Var_d = variance of mean effect size (ES).

Sex composition as a moderator of ST. Mean ST effects were larger for mixed sex samples ($d = –0.26$) than they were for single sex samples ($d = –0.22$). Both types of samples also had confidence intervals that did not contain zero, which suggested dependable ST effects. However, differences between these groups were not statistically significant ($p > .05$). We also used the proportion of women variable to further examine the effect of heterogeneity on the performance of females exposed to ST in mixed sex samples. Results showed that for every unit increase in the proportion of women in the study samples, the mean ES increased (females exposed to ST did better) by 0.02, but this increase was not statistically significant ($p > .05$). The variance in studies conducted in contexts where gender stereotypes were overt had been quite large; we examined studies in this category and found that ST effects were greater for studies presenting the test as diagnostic of personal ability ($d = –0.35$, CI $_{95} = –0.62$ –0.09), versus explicitly stating that there are gender differences on the test ($d = –0.26$, CI $_{95} = –0.43$ –0.10). The differences between these sub-groups also did not reach statistical significance.

Geographical region as a moderator of ST in the U.S. Mean ES for different regions in the United States showed the largest effects in the Midwest ($d = -0.42$), followed by the south ($d = -0.38$) and smallest in Western U.S. ($d = -0.16$). However, only the results for the Midwest and South were reliable, as their confidence intervals did not include zero. Also, fail-safe numbers were also low, indicating that publication bias could not be safely ignored in the interpretation of ES for all other regions but the Midwest. Moderator analyses revealed no significant regional differences in ST effects ($p > .05$). The large variance in ES for studies conducted in the Northeast and Western United States indicated a lot more variability in ST effects in these regions. To investigate this heterogeneity further, we analyzed data by sub divisions for each of the main US regions using the 2010 census categories. However, very few studies had reported information that would enable classification into given sub-divisions for any given region (see Table 4). This made it impossible to conduct fine-grained analyses of ST effects by sub-division for main regions in the United States. All the same, the trends in ES revealed sub-divisional variations in ST effects (see Table 4), for instance in the West, mean ST effects ES were more negative for studies conducted in the mountain region compared to the Pacific.

Subgroup Analyses

Results from this meta-analysis revealed a significant amount of heterogeneity in ST effects among studies that could not be explained by the moderators investigated. Subgroup analyses were conducted to explore potential sources of heterogeneity further.

We started by investigating whether heterogeneity could be explained by the methodological quality of studies included in the meta-analysis. The

TABLE 4. ST Effects by Sub-Division in the U.S.

Region	*k*	*d*	CI_{95}
Northeast	17	−0.29	−0.65 0.06
New England	10	−0.61	−0.99–0.22
Middle Atlantic	4	−0.04	−1.05 0.97
Midwest	28	−0.42	−0.61–0.24
East North Central	6	−0.19	−0.51 0.12
West North Central	4	−0.48	−0.81–0.15
South	8	−0.38	−0.62–0.15
East South Central	1	−0.61	–
West South Central	7	−0.36	−0.61–0.10
West	10	0.02	−0.46 0.51
Mountain	2	−0.46	−1.27 0.35
Pacific	8	0.09	−0.69 0.87

Note. Discrepancies with some *k*'s for the Northeast and Midwest are due to unavailable information regarding the state where the study was conducted. (Some studies only reported geographical region for samples more broadly e.g. students from a large Midwestern/Northeastern university).

methodological quality form scores range from zero to a maximum of 22 points. Scores of studies included in the analyses ranged from 5–19. We created a dichotomous variable using 12 (arbitrarily chosen) as a cut off score for good quality studies. Most studies ($k = 89$) had quality scores above 12. We found that the mean ES for better quality studies was half that of studies with average methodological rigor (i.e., $k = 89$, $d = –0.20$, CI $_{95} = –0.32$ $–0.09$ cf. $k = 14$, $d = –0.46$, CI $_{95} = –0.73$ $–0.18$). However, the differences in mean ST effects were not statistically significant ($p > .05$).

The funnel plot examined earlier also indicated heterogeneity in effect sizes (ES), with several studies showing a reversal in the expected ST trend (i.e., positive ES denoting better performance of females exposed to ST). This trend, showing that females in ST experimental conditions ST outperformed their control counterparts, was evident in 32 studies. A critical qualitative review of these studies indicated that this trend could have been attributed to emerging patterns in the interaction of alternate positive social identities and contextual factors present in the testing environment.

Alternate positive social identities. Most positive effect sizes came from studies where the alternate, positive identities of the participants could have been either inadvertently introduced into the experimental design or deliberately activated as part of the experimental manipulation ($k = 6$). An example of the latter was the study of by Rydell, McConnell and Beilock (2009) where college students in

experimental conditions where multiple identities (i.e. a positive stereotype that college students were smarter than non-college students, and a negative stereotype that women were bad at math) were activated simultaneously were compared to a no identity activated (control) condition before taking a math test. Studies where positive identities might have been indirectly introduced into the experimental design involved (a) participant-specific characteristics e.g., samples of Asian females in ST experiments ($k = 1$) where the positive ethnic stereotype could have counteracted the negative female stereotype regarding quantitative ability, and (b) contextual features of the testing context itself ($k = 2$). With respect to the latter, single-sex samples were often ethnically diverse, which might have boosted the performance of the females with the alternate positive (in this case, ethnic) identity (e.g., Caucasian or Asian females) taking the test under the same ST conditions as other minority females ($k = 5$). Indeed, when we decomposed effect sizes for same sex samples by ethnicity in studies that facilitated this analysis e.g., Armenta (2010), we found reversed ST effects (positive effect sizes) for White women and negative ST effects for Hispanics. This suggested higher performance for Caucasian females compared to their Hispanic counterparts under the same conditions where ST was activated. Finally, studies that blurred inter-group boundaries by having females focus on gender similarities and not differences in other contexts ($k = 5$) also yielded positive ES. Therefore, it seemed that the majority of ST reversal effects were associated with the introduction of alternate positive social identities, which could have counteracted ST effects.

Contextual factors. Other results showing reverse ST effects were based on studies with student samples with low math identification ($k = 1$), samples where members of stigmatized groups did not have solo status ($k = 1$), with samples from STEM-related domains ($k = 3$—half of the STEM studies included in the meta analysis) and countries with very low scores on Hofstede's masculinity index ($k = 5$). Even for studies reporting ST effects in the expected direction (per STT), further analyses pointed to contextual factors like geography and education level as potential contributors to differential ST effects. For instance, as shown in Table 2, ST had larger mean negative effects for samples from Southern Europe and East Africa compared to North American samples. The mean ES for studies conducted in Sweden and the Netherlands was almost zero, with females in experimental groups performing better than their control counterparts. Also, ST effects appeared to be more negative for samples in the middle school ($d = –0.30$) and the high school ($d = –0.30$), compared to the college samples ($d = –0.24$). A minimum of 12 studies per sub-group is required to achieve a power of 0.9 to conduct moderator analyses (Borenstein et al, 2009). Unfortunately, extremely low sample *k*'s for most sub-categories of education level and geographical region translated to low power which precluded moderator analyses exploring these differences further.

Discussion

Findings from this meta-analysis revealed that on average, females under ST performed nearly a quarter of a standard deviation below their non-ST counterparts. Interpreted in the context of standardized math exams like the GRE-Quantitative test (GRE-Q), which most ST studies have used in ST research, the ramifications become much clearer. ST aside, there has been a gender gap favoring males on standardized mathematics/quantitative tests for decades. For SAT-Math scores, this gap has remained constant for over 35 years (Halpern et al., 2007). In 2007 for example, males scored an average of 34 points higher (*M (SD)* = 533 (114)) than females (*M (SD)* = 499 (111)) on the SAT-M (Hyde et. al., 2008). The performance gap is reportedly bigger for the GRE-Q, where males score an average of 70 points higher than females (Coley, 2001). For the 2006-2007 general US student test-taking population, males (*M(SD)* = 599 (141) outscored females (*M(SD)* = 521 (138)) on the GRE-Q by 78 points on average (Educational Testing Service, 2008).

Most ST studies use items from the GRE-Q, and in these experiments, the mean performance of females in the control groups often serves as a baseline, reflecting the expected mean performance of females under ordinary testing situations on such tests. Alternatively, control scores could also be viewed as a proxy for mean performance on the GRE when females are not subjected to ST. In the current study, females in ST experimental groups performed nearly a quarter of a standard deviation below their control counterparts. Thus, the mean ST effect found in the current study ($d = -0.24$) might be akin to stereotype threatened females performing a quarter of a standard deviation (or approximately 33 points) below the expected average performance on the GRE-Q, which is lower than the mean male GRE-Q score to begin with. That the mean scores of ST affected females could potentially fall below the average score of female GRE test takers in the general population suggests that the ST could exacerbate what is already a disadvantageous situation. Viewed in this context, the magnitude of ST on math performance could have a deleterious effect upon admittance into selective graduate programs, especially in STEM disciplines, given the emphasis placed on quantitative ability as a factor of paramount importance.

Results also showed that ST was not moderated by the nature of testing environment or sex composition of the participants, and that females did not benefit more from test setting situations that were homogeneous, or testing contexts where they formed the majority. Although regional differences in ST effects did not reach statistical significance for U.S. samples in this study, we believe that this might have been due to potential sub-cultural variations within each given region. However, efforts to explore these differences at a more fine-grained level were futile as the number of studies reporting more specific geographic information were so few that further analysis was impossible. Nonetheless, we still believe that this area still holds promise as a potential moderator of ST and merits further investigation at the primary level.

That none of the features of the experimental design in the elaborate coding form could explain the between-study variance in ST effects suggests at least in part that the mediating mechanisms of ST are not yet fully known, and that the answer could lie with non-experimental contextual factors or other situational elements not assessed in primary level studies. Specifically, the interaction of broader contextual factors with psychological factors possessed by participants in ST studies could have explained some of the variance in ST effects. Be that as it may, both quantitative and qualitative trends in the data showed potential interactions between multiple participant identities activated during the ST experiment and differential ST effects by geographical region and level of education. Earlier on, we surmised that contextual factors related to geography might have been useful in explaining some of the variance in ST studies. Although the discussion focused on US geographical regions, we would like to add that this could also be the case for studies conducted in other countries culturally different from the United States. Indeed, this study showed a trend of reverse to near-zero ST effects for primary studies conducted in countries with small gender gaps, like the Netherlands and Sweden for example (e.g., Eriksson & Lindholm, 2007; Scholten et al., 2005). That ST effects were weak and reversed in studies conducted on females in Scandinavian countries and were larger in African samples is plausible given that gender roles are strong and distinct in Africa (Martineau, 1997), and weak to non-existent in Scandinavian countries, which have the smallest gender gap in the world (Hausmann, Tyson, Bekhouche, & Zahidi, 2011). We surmise that in countries where culturally, gender roles are weak, ST would hardly be expected to impact the performance of females because the occurrence of ST requires the individual to be aware of negative stereotypes (i.e., high stigma consciousness) about one's group in a given domain, which might not be the case for individuals from these cultures. This is because at a cultural level (within egalitarian societies), these gender stereotypes might be so weak that they engender low (as opposed to high) levels of stigma consciousness regarding gender stereotypes in women. Consequently, this would make women less likely to pick up environmental cues related to the negative stigma surrounding females' math ability. Further, because these stereotypes are not part of the consciousness of the broader culture, introducing them in experimental testing contexts primed for ST might hardly be discernible or even perceived as self-relevant enough to participants to severely impact performance in this area.

This study was limited in a few ways, which temper the results reported. First, the majority of primary ST studies have been conducted in the United States. This was reflected in our meta-analysis where nearly two thirds of the studies were from the United States. Therefore, caution should be taken in the interpretation of ST effects, as they may not be easily generalizable to countries that are more culturally distinct from the United States.

Second, the analysis of potential contextual moderators (like region) of ST was precluded by insufficient study samples, *k*, (and hence power) required to

investigate those factors further. The lack of power resulting from insufficient *k*'s for certain sub-categories within these factors prohibited more fine-grained analyses. Moreover, the problem of inadequate studies at the meta-analytic level appeared to reflect similar problems inherent in the literature. That is, key ST moderator variables (whether contextual or psychological) have been investigated by only a few primary experimental studies. Consequently, limited investigation of key ST moderators at the primary level bound the validity, generalizability, or even the power to detect their effects at the meta-analytic level as well. This could be easily remedied by replication research in areas where few studies have been conducted. We suspect that the paucity of ST research on context could be partly attributed to the fact that the majority of ST research has been largely experimental, and this mode of inquiry might be somewhat limited in its capacity to provide insight into the contextual mechanisms that might mediate or moderate ST. Thus we reckon that a more effective approach to understanding ST, especially with respect to its impact on individuals in authentic academic environments, might be mixed methods research that incorporates qualitative inquiry into context. The latter would be more useful in exploring questions surrounding the role of context in mediating or moderating ST.

Conclusion

The current meta-analysis has shown that ST does exert a negative effect on the quantitative performance of females, and that this effect could significantly hinder entry into graduate programs in science disciplines. Based on the synthesis of ST research over a 17-year period, we contend that research on ST and the mechanisms that drive it are barely out of its nascent stage. Certainly, over the past decade and a half, a lot of work in this area has been done, but this line of research has expanded more in breadth than in depth. That is, a lot of ground has been covered with respect to uncovering factors that mediate and moderate the phenomenon, but the literature is plagued with insufficient replication studies investigating these factors more fully. To that end, the present study has brought the need for the replication of several under-investigated factors identified as ST moderators at the empirical level to the forefront, and also highlighted the need for more research to explore the interactive nature of multiple identities and contextual factors in moderating ST. Currently, most ST studies repeatedly cite the same one or two studies that have investigated ST moderators, which might inadvertently yield misguided perceptions regarding the strength of any given ST moderators. Cadinu and colleagues' (2006) study for example, is perhaps one of a handful of studies that has investigated and reported locus of control as a moderator of the phenomenon. The risk in citing results from under or unreplicated studies is that the moderator effects might be exaggerated, or that the findings themselves might be attenuated by the randomness introduced in the research

study. A lack of replication would therefore make it difficult to develop a comprehensive theory of ST. This could have serious implications for ST interventions and the researchers pursuing this line of work.

First, effective interventions are grounded in solid theory supported by the best empirical evidence; therefore, structuring interventions based on under replicated ST moderators might lead to ineffective interventions. It appears that at this time, given the current state of empirical research on ST moderators, any interventions to alleviate ST would be premature. It might be worthwhile for researchers to first focus on replicating findings across several key factors empirically identified as ST moderators, and developing research on the contextual component of STT before developing interventions to combat the phenomenon.

Based on the non-significant findings of several ST empirically identified moderators, we contend that the current meta-analysis has successfully demonstrated what is not missing, i.e. what does not seem to moderate females' performance in math under ST conditions. Therefore, it is incumbent upon future researchers to explore the mechanisms that drive ST more fully. A good first step in this direction would be to investigate the interaction between multiple identities, as well as several levels of context both broad and narrow (i.e., national, educational and academic domain), which have thus far been under-investigated, as implicated in the current study. In particular, future research might aim to demonstrate patterns of ST on the basis of geography; we predict that findings might be indicative of stronger stereotype presence supported in what we typically consider to be cultures (both regional and national) that are less egalitarian in their expectations and perceptions of gender roles. Researchers are also encouraged to examine ST within the larger context of factors affecting females' performance in mathematics and treat it as one of many factors that curtail the achievement and advancement of females in these domains. Inter-disciplinary research tying other factors to females' performance in these domains might yield more useful information pertinent to narrowing the gender gap in STEM and boosting the performance of females in these domains.

AUTHOR NOTES

Katherine Picho is affiliated with the Connecticut Institute for Clinical and Translational Science at the University of Connecticut Health Center. **Ariel Rodriguez** is affiliated with the University of Hartford. **Lauren Finnie** is affiliated with the University of Hartford.

REFERENCES

References marked with an asterisk indicate studies included in the meta-analysis.

*Ambady, N., Paik, S. K., Steele, J., Owen-Smith, A., & Mitchell, J. P. (2004) Deflecting negative self-relevant stereotype activation: The effects of individuation. *Journal of Experimental Social Psychology*, *40*. 401–408. doi:10.1016/j.jesp.2003.08.003

Ambady, N., Shih, M., Stephanie, A., & Pittinsky, T.L. (2001). Stereotype susceptibility in children: Effects of identity activation on quantitative performance. *Psychological Science*, *12*, 385–390. doi:10.1111/1467-9280.00371

Ardila, A., Rosselli, M., Matute, E., & Inozemtseva, O. (2011). Gender differences in cognitive development. *Journal of Developmental Psychology*, *47*, 984–990. doi:10.1037/a0023819

Armenta, B. (2010). Stereotype boost and stereotype threat effects: The moderating role of ethnic identification. *Cultural Diversity and Ethnic Minority Psychology*, *16*, 94–98. doi:10.1037/a0017564

*Aronson, J., & Steele, C. M. (2005). Stereotypes and the fragility of human competence, motivation, and self-concept. In C. Dweck & E. Elliot (Eds.), *Handbook of competence and motivation* (pp. 436–456). New York, NY: Guilford Publications.

Baenninger, M., & Newcombe, N. (1995). Environmental input to the development of sex-related differences in spatial and mathematical ability. *Learning and Individual Differences*, *7*, 363–379. doi:10.1016/1041-6080(95)90007-1

Bargh, J. A. (1994). The four horsemen of automaticity. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (pp. 1–40). Hillsdale, NJ: Erlbaum.

Bargh, J. A., Chen, M., & Barrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*, 230–244. doi:10.1037/0022-3514.71.2.230

*Beaton, A., Tougas, F., Rinfret, N., & Huard, N. (2007). Strength in numbers? Women and mathematics. *European Journal of Psychology of Education*, *22*, 291–306. doi:10.1007/BF03173427

Beede, D., Julian, T., Langdon, D., McKittrick, G., Khan, B., & Doms, M. (2011). *Women in stem: A gender gap to innovation* (ESA #04-11). U.S. Department of Commerce, Economics, and Statistics Administration. Retrieved from http://www.esa.doc.gov/sites/default/files/reports/documents/womeninstemagaptoinnovation8311.pdf

*Bell, A. E., Spencer, S. J., Iserman, E., & Logel, C. E. R. (2003). Stereotype threat and women's performance in Engineering. *Journal of Engineering Education, 92,* 307–312. Retrieved from http://www.jee.org/2003/october/782.pdf

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley & Sons LTD

Brehm, J. W. (1966). *A theory of psychological reactance*. New York, NY: Academic Press.

*Brodish, A. B. (2007). *Stereotype threat and achievement goals: An integrative approach*. (Unpublished doctoral dissertation). University of Wisconsin, Madison, WI.

*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

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*, 626–633. doi:10.1016/S0022-1031(03)00039-8

Burelli, J., Arena, C., Shettle, C., & Fort, D. (1996). *Women, minorities, and persons with disabilities in science and engineering: 1996* (NSF 96-311). Arlington, VA: National Science Foundation, Division of Science Resources Studies. Retrieved from http://www.nsf.gov/statistics/nsf96311/intro.pdf

*Cadinu, M., Maas, A., Rosabianca, A., Figerio, S., & Latinotti, S. (2003). Stereotype Threat: The effect of expectancy performance. *European Journal of Social Psychology*, *33*, 267–285. doi: 10.1002/ejsp.145

*Cadinu, M., Maas, A., Rosabianca, A., & Kiesner, J. (2005). Why do women underperform under stereotype threat? Evidence for the role of negative thinking. *Psychological Science*, *16*, 572–578. doi:10.1111/j.0956-7976.2005.01577.x

*Cadinu, M., Maas, A., Rosabianca, A., Lombardo, M., & Figerio, S. (2006). Stereotype threat: The moderating role of locus of control beliefs. *European Journal of Social Psychology*, *36*, 183–197. doi:10.1002/ejsp.303

*Campbell, S. M., & Collaer, L. M. (2009). Stereotype threat and gender differences in performance on a novel visuospatial task. *Psychology of Women Quarterly*, *33*, 437–444. doi:10.1111/j.1471-6402.2009.01521.x

*Clark, J. K., Eno, C. A., & Guadagno, R. E. (2011). Southern discomfort: The effects of stereotype threat on the intellectual performance of southerners. *Self and Identity*, *10*, 248–262. doi:10.1080/15298861003771080

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. doi:10.1037/0033-2909.112.1.155

Coley, R. J. (2001). *Differences in the gender gap: comparisons across racial/ethnic groups in education and work*. Princeton, NJ: Educational Testing Services. Retrieved from http://www.ets.org/Media/Research/pdf/PICGENDER.pdf

*Cotting, D. I. (2003). *Shedding light in the black box of stereotype threat: The role of emotion*. (Unpublished dissertation). City University of New York, New York, NY.

*Crisp, R. J., Bache, L. M., & Maitner, A. T. (2009). Dynamics of social comparison in counter-stereotypic domains: Stereotype boost, not stereotype threat for women engineering majors. *Social Influence*, *4*, 171–184. doi:10.1080/15534510802607953

*Croizet, J. C., Despres, G., Gauzins, M. E., Huguet, P., Leyens, J. P., & Méot, A. (2004). Stereotype threat undermines intellectual performance by triggering a disruptive mental load. *Personality and Social Psychology Bulletin*, *30*, 721–731. doi: 10.1177/0146167204263961

*Davies, P. G., Spencer, S. J., Quinn, D. M., & Gerhardstein, R. (2002). Consuming images: How television commercials that elicit stereotype threat can restrain women academically and professionally. *Personality and Social Psychology Bulletin*, *28*, 1615–1628. doi: 10.1177/014616702237644

Dinella, L. M. (2004). *A development perspective of stereotype threat on high school mathematics*. (Unpublished dissertation). Arizona State University, Phoenix, AZ.

Eagly, A. H. (1987). *Sex differences in social behavior: A social-role interpretation*. Hillsdale, NJ: Erlbaum

Educational Testing Services. (2008). *Graduate Record Examination: Factors that can influence performance on the GRE General Test 2006–2007*. Retrieved from http://www.ets.org/Media/Tests/GRE/pdf/gre_0809_factors_2006-07.pdf

*Elizaga, R. A., & Markman, K. D. (2008). Peers and performance: How in-group and outgroup comparisons moderate stereotype threat effects. *Current Psychology*, *27*, 290–300. doi: 10.1007/s12144-008-9041-y

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

*Ford, T. E., Ferguson, M. A., Brooks, J. L., & Hagadone, K. M. (2004). Coping sense of humor reduces effects of stereotype threat on women's math performance. *Personality and Social Psychology Bulletin*, *30*, 643–653. doi: 10.1177/0146167203262851

Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics*, *118*, 1049–1074. doi: 10.1162/00335530360698496

*Good, C., Aronson, J., & Harder, J.A. (2008). Problems in the pipeline: Stereotype threat and women's achievement in high-level math courses. *Journal of Applied Developmental Psychology 29*, 17–28. doi: 10.1016/j.appdev.2007.10.004

*Gonzales, P. M., Blanton, H., & Williams, K. J. (2002). The effects of stereotype threat and double-minority status on the test performance of Latino women. *Personality and Social Psychology Bulletin*, *28*, 659–670. doi: 10.1177/0146167202288010

Halpern, D. F. (1992). *Sex differences in cognitive abilities* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum, Associates, Inc.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest*, *8*, 1–51. doi: 10.1111/j.1529-1006.2007.00032.x

Hausman, R., Tyson, L. D., & Zahidi, S. (2011). *The global gender-gap report 2011*. Geneva, Switzerland: World Economic Forum. Retrieved from http://www3.weforum.org/docs/WEF_GenderGap_Report_2011.pdf

Higgins, E. T. (1989). Knowledge accessibility and activation: Subjectivity and suffering from unconscious sources. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 75–123). New York, NY: Guilford Press.

Huedo-Medina, T. B., & Johnson, B. T. (2011). *Estimating the standardized mean difference effect size and its variance from different data sources: A spreadsheet*. Storrs, CT: Authors.

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

Huguet, P., & Regner, I. (2009). Counter stereotypic beliefs in math do not protect school girls from stereotype threat. *Journal of Experimental Social Psychology*, *45*, 1024–1027. doi: 10.1016/j.jesp.2009.04.029

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

Inzlicht M., & Ben-Zeev, T. (2000). A threatening intellectual environment: Why females are susceptible to experiencing problem-solving deficits in the presence of males. *Psychological Science*, *11*, 365–371. doi: 10.1111/1467-9280.00272

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*, 796–805. doi: 10.1037/0022-0663.95.4.796

Ioannidis, J. P. A., & Trikalinos, T. A. (2007). The appropriateness of asymmetry tests for publication bias in meta-analyses: A large survey. *Canadian Medical Association Journal*, *176*, 1091–1096. doi: 10.1503/cmaj.060410

*Jamieson, J. P. (2009). *The role of motivation in blatant stereotype threat, subtle stereotype threat, and stereotype priming*. Psychology Dissertations. Paper 10. Retrieved from http://hdl.handle.net/2047/d20000007

Jamieson, J. P., & Harkins, S. G. (2012). Distinguishing between the effects of stereotype priming and stereotype threat on math performance. *Group Processes and Intergroup Relations*, *15*, 291–304. doi: 10.1177/1368430211417833

*Johns, M., Schmader, T., & Martens, A. (2005). Knowing is half the battle: Teaching stereotype threat as a means of improving women's math performance. *Psychological Science 16*, 175–179. doi: 10.1111/j.0956-7976.2005.00799.x

Kalaian, S. A., & Kassim, R. M. (2008). Multilevel methods for Meta-Analysis. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 315–343). Charlotte, NC: Information Age Publishing.

*Keller, J. (2002). Blatant stereotype threat and women's math performance: Self-handicapping as a strategic means to cope with obtrusive negative performance expectations. *Sex Roles*, *47*, 193–198. doi: 10.1023/A:1021003307511

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

Keller, J., & Molix, L. (2008). When women can't do math: The interplay of self-construal, gender identification, and stereotypic performance standards. *Journal of Experimental Social Psychology*, *44*, 437–444. doi: 10.1016/j.jesp.2007.01.007

Kray, L. J, Galinsky, A., & Thompson, L. (2001). Battle of the sexes: Gender stereotype confirmation and reactance in negotiations. *Journal of Personality and Social Psychology*, *80*, 942–958. doi: 10.1037/0022-3514.80.6.942

*Lesko, A. C., & Corpus, J. H. (2006). Discounting the difficult: How high math-identified women respond to stereotype threat. *Sex Roles*, *54*, 113–125. doi: 10.1007/s11199-005-8873-2

Levant, R. H., Majors, R. G., Kelley, M. L. (1998). Masculinity ideology among young African American and European American women and men in different regions of the United States. *Cultural Diversity and Mental Health*, *4*, 227–236. doi: 10.1037/1099-9809.4.3.227

Levine, S. C., Huttenlocher, J., Taylor, A., & Langrock, A. (1999). Early sex differences in spatial skill. *Developmental Psychology*, *35*, 940–949. doi: 10.1037/0012-1649.35.4.940

*Martens, A., Johns, M., Greenburg, M., & Schimel, J. (2006). Combating stereotype threat: The effect of self-affirmation on women's intellectual performance. *Journal of Experimental Social Psychology*, *42*, 236–243. doi: 10.1016/j.jesp.2005.04.010

Martineau, R. (1997). Women and education in South Africa: Factors influencing women's educational progress and their entry into traditionally male-dominated fields. *The Journal of Negro Education, 66,* 383–395. Retrieved from http://www.jstor.org/stable/2668166

Marx, D. M., & Roman, J.S. (2002). Female role models: Protecting women's math test performance. *Personality and Social Psychology Bulletin*, *28*, 1183–1193. doi: 10.1177/01461672022812004

*McIntyre, R. B., Paulson, R. M., Taylor, C.A., Morin, A. L., & Lord, C. G. (2010). Effects of role model deservingness on overcoming performance deficits induced by stereotype threat. *European Journal of Social Psychology*, *41*, 301–311. doi: 10.1002/ejsp.774

Moskowitz, G. B., & Skurnik, I. (1999). Contrast effects as determined by the type of prime: Trait versus exemplar primes initiate processing strategies that differ in how accessible constructs are used. *Journal of Personality and Social Psychology*, *76*, 911–927. doi: 10.1037/0022-3514.76.6.911

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*, 1314–1334. doi: 10.1037/a0012702

*Oswald, D. L., & Harvey, R. D. (2001). Hostile environments, stereotype threat, and math performance among undergraduate women. *Current Psychology*, *19*, 338–356. doi: 10.1007/s12144-000-1025-5

*Picho, K., & Stephens, J. M. (2012). Culture, context and stereotype threat: A comparative analysis of single-sex and coed schools in Uganda. *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. doi: 10.1111/0022-4537.00201

Rice, T. W., & Coates, D. L. (1995). Gender role attitudes in the Southern United States. *Gender and Society*, *9*, 744–756. doi: 10.1177/089124395009006007

*Rivardo, M. G., Rhodes, M. E., Camaione, T. C., & Legg, J. M. (2011). Stereotype threat leads to reduction in number of math problems women attempt. *North American Journal of Psychology, 13,* 5–16.

Rosenberg, M. S. (2005). The file drawer problem revisited: A general weighted method for calculating fail-safe numbers in meta-analysis. *Evolution*, *59*, 464–468. doi: 10.1554/04-602

*Rosenthal, H. E. S., & Crisp, R. J. (2006). Reducing stereotype threat by blurring intergroup boundaries. *Personality and Social Psychology Bulletin*, *32*, 501–511. doi: 10.1177/0146167205281009

*Rucks, L. J. (2008). *Me, women, and math: The role of personal and collective threats in the experience of stereotype threat*. (Doctoral dissertation). Ohio State University, Columbus, OH. Retrieved from http://etd.ohiolink.edu/send-pdf.cgi/Rucks%20Lana%20J.pdf?osu1204661976

*Rydell, R. J., McConnell, A. J., Beilock, S. L. (2009). Multiple social identities and stereotype threat: Imbalance, accessibility and working memory. *Journal of Personality and Social Psychology*, *96*, 949–966. doi: 10.1037/a0014846

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

Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology*, *88*, 934–947. doi: 10.1037/0022-3514.85.3.440

Schmader, T., Johns, M. & Barquissau, M. (2004). The cost of accepting gender differences: The role of stereotype endorsement in women's experience in math domains. *Sex Roles*, *50*, 835–850. doi: 10.1023/B:SERS.0000029101.74557.a0

*Scholten, A. Z., Wicherts, J. M., Elsenburg, F., Stoffels, M., Delsing, G., Dotsch, T., & Mulders, M. (2005). *Het effect van stereotype dreiging aangaande de scores van meisjes op een wiskundetest* [Stereotype threat effects on the scores of girls on a math test] (Internal Report OP4810). Amsterdam, Netherlands: University of Amsterdam.

Sekaquaptewa, D., & Thompson, M. (2003). Solo status, stereotype threat and performance expectancies: Their effects on women's performance. *Journal of Experimental Social Psychology*, *39*, 68–74. doi: 10.1016/S0022-1031(02)00508-5

*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*, *68*, 797–811. doi: 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*, 93–102. doi: 10.1037/a0026617

Suitor, J. J. & Carter, R. S. (1999). Jocks, nerds, babes, and thugs: A research note on regional differences in adolescent gender norms. *Gender Issues*, *3*, 87–101. doi: 10.1007/s12147-999-0005-9

*Taylor, C. A., Lord, C. G., McIntyre, R. B., & Paulson, R. M. (2011). The Hillary Clinton effect: When the same role model inspires or fails to inspire improved performance under stereotype threat. *Group Processes Intergroup Relations*, *14*, 447–459. doi: 10.1177/136843021038268

Terlecki, M. S., Newcombe, N. S., & Little, M. (2007). Durable and generalized effects of spatial experience on mental rotation: Gender differences in growth patterns. *Applied Cognitive Psychology*, *22*, 996–1013. doi: 10.1002/acp.1420

*Thoman, D. B., White, P. H., Yamawaki, N., & Koishi, H. (2008). Variations of gender-math stereotype content affect women's vulnerability to stereotype threat. *Sex Roles*, *58*, 702–712. doi: 10.1007/s11199-008-9390-x

Twenge, J. M. (1997). Attitudes toward women, 1970–1995. *Psychology of Women Quarterly*, *21*, 35–51. doi: 10.1111/j.1471-6402.1997.tb00099.x

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

Walton, G. M., & Spencer, S. J. (2009). Latent ability: Grades and test scores systemically underestimate the intellectual ability of negatively stereotyped students. *Psychological Science, 20*, 1132–1139. doi: 10.1111/j.1467-9280.2009.02417.x

*Weger, U. W., Hooper, N., Meir, B. P., & Hopthrow, T. (2012). Mindful maths: Reducing the impact of stereotype threat through a mindfulness exercise. *Consciousness and Cognition*, *21*, 471–475. doi: 10.1016/j.concog.2011.10.011

*Werhun, C. D. (2007). *The limitations of stereotype threat: Not all math and science women are threatened by stereotypes*. (Doctoral dissertation). University of Toronto, Toronto, Canada. Retrieved from psycinfo. (Order No. AAINR28095)

*Wicherts, J. M. (2007). *Another study of the effects of stereotype threat on sex differences in mathematics test performance*. Internal report. Amsterdam, Netherlands: University of Amsterdam.

Wilder, G., & Powell, K. (1989). *Sex differences in test performance: A survey of the literature*. New York, NY: College Board Publications.

*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*, 792–799. doi: 10.1016/j.jesp.2007.07.005

Received May 27, 2012
Accepted October 2, 2012