Journal of Applied Psychology
2008, Vol. 93, No. 6, 1314–1334

0021-9010/08/$12.00 DOI: 10.1037/a0012702

Does Stereotype Threat Affect Test Performance of Minorities and Women? A Meta-Analysis of Experimental Evidence

Hannah-Hanh D. Nguyen
California State University, Long Beach

Ann Marie Ryan
Michigan State University

A meta-analysis of stereotype threat effects was conducted and an overall mean effect size of |.26| was found, but true moderator effects existed. A series of hierarchical moderator analyses evidenced differential effects of race- versus gender-based stereotypes. Women experienced smaller performance decrements than did minorities when tests were difficult: mean *d*s = |.36| and |.43|, respectively. For women, subtle threat-activating cues produced the largest effect, followed by blatant and moderately explicit cues: *d*s = |.24|, |.18|, and |.17|, respectively; explicit threat-removal strategies were more effective in reducing stereotype threat effects than subtle ones: *d*s = |.14| and |.33|, respectively. For minorities, moderately explicit stereotype threat-activating cues produced the largest effect, followed by blatant and subtle cues: *d*s = |.64|, |.41|, and |.22|, respectively; explicit removal strategies enhanced stereotype threat effects compared with subtle strategies: *d*s = |.80| and |.34|, respectively. In addition, stereotype threat affected moderately math-identified women more severely than highly math-identified women: *d*s = |.52| and |.29|, respectively; low math-identified women suffered the least from stereotype threat: *d*= |.11|. Theoretical and practical implications of these findings are discussed.

Keywords: stereotype threat effects, meta-analysis, cognitive ability test performance, gender gap in math scores, racial gap in test scores

Since Steele and Aronson's (1995) seminal experiments, the research literature on stereotype threat effects on test performance has steadily grown. According to the theory, *stereotype threat* refers to the "predicament" in which members of a social group (e.g., African Americans, women) "must deal with the possibility of being judged or treated stereotypically, or of doing something that would confirm the stereotype" (p. 401; Steele & Aronson, 1998). For instance, when stereotyped group members take standardized ability tests, such as in educational admission or employment selection contexts, their performance may be partially undermined when they encounter cues of a salient negative stereotype in the testing environment (e.g., women are not good at math, or ethnic minorities are inferior in intellectual abilities; Steele, Spencer, & Aronson, 2002).

Hannah-Hanh D. Nguyen, Department of Psychology, California State University, Long Beach; Ann Marie Ryan, Department of Psychology, Michigan State University.

This article was partly based on Hannah-Hanh D. Nguyen's dissertation research under Ann Marie Ryan's supervision. The meta-analysis was supported with Hannah-Hanh D. Nguyen's National Science Foundation Graduate Research Fellowship and Michigan State University Competitive Doctoral Enrichment Fellowship.

We thank the researchers who generously shared their work and additional statistical data, especially Gregory Walton. We also thank Irene Sze and Emily Harris for their coding assistance, Huy A. Le for his meta-analytic advice, and Linda Jackson, Neal Schmitt, and Ilies Remus for their committee guidance.

Correspondence concerning this article should be addressed to Hannah-Hanh D. Nguyen, Department of Psychology Room 319, California State University, Long Beach, 1250 Bellflower Boulevard, Long Beach, CA 90840. E-mail: hnguyen@csulb.edu

The social message that the theory of stereotype threat conveys is powerful: Members of stigmatized social groups may be constantly at risk of underperformance in testing contexts, and the risks may be partially caused by situational factors (i.e., beyond established factors such as poverty, parental style, socialization, etc.; Steele, 1997). For a *stigmatized social group*, we adopt the widely used definition of Crocker and Major (1989): belonging to a social category about which others hold negative attitudes, stereotypes, and beliefs. According to Devos and Banaji (2003), the contribution of the stereotype threat theory and literature is that it predicts (and empirically tests) the relationship between negative in-group stereotypes and group members' behavioral changes (e.g., diminished task or test performance in a stereotyped evaluative domain), not only attitudinal or affective changes. The present meta-analytic study aims at investigating the extent to which the activation of stereotype threat is detrimental to stereotyped test takers' performance on cognitive ability tests by aggregating the findings in stereotype threat experiments. In our presentation, we discuss how the laboratory conditions may or may not generalize to employment testing contexts, and hence whether or not any stereotype threat effects produced in these contexts might occur in field settings of interest to organizational psychologists.

Stereotype Threat: Research Paradigm and Empirical Evidence

The primary hypothesis of stereotype threat theory is *performance interference*, or the prediction that stereotyped individuals perform worse on an evaluative task (e.g., African Americans taking a verbal ability test or women taking a mathematics test) in a stereotype-threatening context than they would in a nonthreat-

ening condition (see Steele, 1997; Steele et al., 2002). The basic experimental paradigm involves randomly assigning members of a stereotyped group to a control or threat condition and comparing mean performance of the conditions. Researchers sometimes also include a comparison group to whom the induced negative stereotype is not relevant (e.g., Whites, men). In a seminal experiment, Steele and Aronson (1995; Experiment 1) assigned African American students to one of three conditions of stereotype threat where they were administered a difficult verbal ability test (i.e., only 30% of a pretest sample correctly solved the test). In the stereotype threat condition, participants were told that the test was diagnostic of their intellectual capability; in the other conditions, the test was either described as a problem-solving task or no particular directions were given. Participants in the stereotype threat condition correctly solved fewer test problems compared with those in other conditions, supporting the performance interference hypothesis.

A majority of subsequent researchers replicated and extended the stereotype threat effect on cognitive ability tests for African American or Hispanic test takers (e.g., Cadinu, Maass, Frigerio, Impagliazzo, & Latinotti, 2003; Dodge, Williams, & Blanton, 2001) and on difficult math ability tests for female test takers (e.g., Davies, Spencer, Quinn, & Gerhardstein, 2002; Schmader & Johns, 2003; S. J. Spencer, Steele, & Quinn, 1999). However, some researchers did not find support for the performance interference hypothesis (e.g., McFarland, Kemp, Viera, & Odin, 2003; Oswald & Harvey, 2000–2001; Schneeberger & Williams, 2003; Stricker & Ward, 2004). The mixed findings suggest moderating effects for stereotype threat. Stereotype threat theory does propose three moderators—stereotype relevance, domain identification, and test difficulty—which are described in the next section and investigated in this study.

Conceptual Moderators

Relevance of Stereotypes

Steele (1997) posited that the degree of stereotype threat effects may vary depending on the relevance of a stereotype in the test setting. For example, Shih, Pittinsky, and Ambady (1999) found that, when encountering a stereotype threat-loaded situation (i.e., taking a math test), Asian American women tended to solve more correct problems when the situational cues about their "Asian" identity were accessible than when their "woman" identity was made salient (mean differences were not statistically significant, however). Researchers can activate stereotype threat by manipulating the degree to which a stereotype is salient to individuals in an evaluative testing context by either cuing test takers to the link between a stereotype and a particular evaluative test (i.e., implicit stereotype threat activation) or by declaring that members of a social group tend to perform worse on the test than a comparison group (i.e., explicit stereotype threat activation; see definitions and examples in Table 1).

Stereotype theories in general predict that more implicit threat cues would have a stronger negative effect on task performance than explicit ones (see Bargh, 1997). Explicit stereotype threat cues have been found to cause target test takers to overperform on a test or task (e.g., Kray, Thompson, & Galinsky, 2001; McFarland, Kemp, et al., 2003), a phenomenon called a *stereotype reactance effect.* According to Kray et al. (2001), when a negative stereotype is blatantly and explicitly activated, it might be perceived by test takers as a limit to their freedom and ability to perform, thereby ironically invoking behaviors that are inconsistent with the stereotype.

The implication is that a nonlinear relationship possibly exists between stereotype threat-activating cues and performance. Studies using subtle stereotype threat-activating cues (i.e., via manipulating the testing environment) might produce a larger effect size than those using blatant cues (i.e., via spelling out a stereotype to target test takers) because the stereotype might work on a subconscious level and directly affect targets' test performance (see Levy, 1996). Also, threatened individuals might consciously react against a blatant stereotype. Studies using moderately explicit cues might actually yield the greatest stereotype threat effects: When a general message of subgroup differences in intellectual abilities is explicitly conveyed to target test takers but the direction of these differences is not specified and is instead left open for test takers' interpretation, the stereotype might be direct enough to draw targets' attention, ambiguous enough to cause targets to engage in detrimental off-task thinking (e.g., trying to figure out how the message should be interpreted), but not too blatant to make some targets become motivated to "prove it wrong." This distinction is particularly important to consider if one wishes to determine if laboratory stereotype threat effects generalize to employment testing settings, where blatant or even moderately explicit cues are unlikely to be present. If effects are not found with more subtle cues, then one might question the applicability of this line of research to employment testing contexts.

The relevance between a negative stereotype and a test can also be refuted or removed to reduce observed stereotype threat effects, either implicitly (e.g., by framing a test as a nondiagnostic task) or explicitly (e.g., by disputing said group differences in test performance; see examples in Table 2). Explicit stereotype threat-removal strategies may serve as a catalyst to motivate individuals to avoid being stereotyped; this motivation can in turn inhibit negative stereotypes by shaping activated thoughts into actions toward their goals (see S. J. Spencer, Fein, Strahan, & Zanna, 2005). In other words, explicitly making a stereotype less relevant to a test context might alleviate stereotype threat effects more effectively than implicit threat-removal strategies. This distinction is also an important one to examine in determining the viability of generalizing stereotype threat lab research to employment testing contexts: Hiring organizations are unlikely to enact the more explicit threat-removal strategies used in this line of research.

In this meta-analysis, we examine the type of stereotype activation cue—subtle, moderately explicit, or blatant—and the type of threat-removal strategy (implicit vs. explicit) as potential moderators.

Domain Identification

Stereotype threat theory proposes that only those who strongly identify themselves with a domain with which there is a negative group stereotype are susceptible to the threat of confirming the group-based stigma because the strength of stereotype threat effects depends on "the degree to which one's self-regard, or some component of it, depends on the outcomes one experiences in the domain" (p. 390; Steele et al., 2002). For example, only women who identify with math would experience stereotype threat while

Table 1
Stereotype Threat-Activating Cues

Cue classification	Operational definition	Stereotype threat-activating cue	Example study
Blatant	The message involving a stereotype about a subgroup's inferiority in cognitive ability and/or ability performance is explicitly conveyed to test takers prior to their taking a cognitive ability test. The group-based negative stereotype becomes salient to test takers via a conscious mechanism.	Emphasizing the target subgroup's inferiority on tests (or the comparison subgroup's superiority). For example, stating that Whites tend to perform better than Blacks/Hispanics or that men tend to score higher than women.	Aronson et al. (1999); Cadinu et al. (2003); Schneeberger & Williams (2003)
		Priming targets' group-based inferiority. For example, administering a stereotype threat questionnaire before tests or giving information favoring males before tests.	Bailey (2004); Seagal (2001)
Moderately explicit	The message of subgroup differences in cognitive ability and/or ability performance is conveyed directly to test takers in test directions or via the test-taking context, but the direction of these group differences is left open for test takers' interpretation. The group-based negative stereotype may become salient to test takers via a conscious mechanism.	Race/gender performance differences in general ability tests. For example, stating that generally men and women perform differently on standardized math tests.	R. P. Brown & Pinel (2003); Edwards (2004); H. E. S. Rosenthal & Crisp (2006)
		Race/gender performance differences on the specific test. For example, stating that taking a specific math test produces gender differences, testing minorities' math ability on a White-normed or biased test, stating that certain groups of people perform better than others on math exams.	Keller & Dauenheimer (2003); Pellegrini (2005); Tagler (2003)
Indirect and subtle	The message of subgroup differences in cognitive ability is not directly conveyed; instead, the context of tests, test takers' subgroup membership, or test taking experience is manipulated. The group-based negative stereotype may become salient to test takers via an automatic and/or subconscious mechanism.	Race/gender priming. For example, making a race/gender inquiry prior to tests or race/gender priming by other means (e.g., a pretest questionnaire, a pretest task, a testing environment cue).	Anderson (2001); Dinella (2004); Oswald & Harvey (2000–2001); Schmader & Johns (2003); Spicer (1999); Steele & Aronson (1995)
		Emphasizing test diagnosticity purpose. For example, labeling the test as a diagnostic test or stressing the evaluative nature of the test.	Martin (2004); Marx & Stapel (2006); Ployhart et al. (2003); Prather (2005)

Note. In Walton and Cohen's (2003) meta-analysis, only two levels of classification were employed.

taking a math test (Cadinu et al., 2003, Study 1). That is, negative stereotypes will not be threatening to individuals who do not care about performing well in an area, as success in that domain plays little role in their identity. Surprisingly, only a few studies directly tested domain identification as a moderator of stereotype threat effects, and the results were mixed (Aronson et al., 1999; Cadinu et al., 2003; Leyens, Desert, Croizet, & Darcis, 2000; McFarland, Lev-Arey, & Ziegert, 2003). In this meta-analysis, we examined whether levels of individuals' domain identification might influence the magnitude of stereotype threat effects.

Table 2
Stereotype Threat-Removal Strategies

Strategy	Example study
Explicit	
Give a handout with information favoring women	Bailey (2004)
State that a math test is free of gender bias (men = women)	R. P. Brown & Pinel (2003)
State that Blacks perform better than Whites	Cadinu et al. (2003)
Educate subjects about the stereotype threat phenomenon	Guajardo (2005)
Subtle	
Describe a test as a problem-solving task (no race inquiry before task)	Steele & Aronson (1995)
State that test performance will not be assessed	Wout et al. (n.d.)
Show television commercials with women in astereotypical roles (e.g., engineers)	Davies et al. (2002)

Test Difficulty

Stereotype threat theory suggests that members of a stigmatized social group are most likely to be threatened by a situational stereotype threat cue when a test is challenging. Because the cognitive demands of a difficult test will increase individuals' mental workload, interference from a stereotype will be cognitively more problematic when a test is challenging than when a test does not require as much from the test takers' resources (Steele & Aronson, 1995; Steele et al., 2002). Some researchers selected highly difficult intellectual ability tests to investigate stereotype threat effects (e.g., Croizet et al., 2004; Gonzales, Blanton, & Williams, 2002; Inzlicht & Ben-Zeev, 2003; McIntyre et al., 2003; Schmader, 2002; Steele & Aronson, 1995), whereas other researchers used moderately difficult tests (e.g., Dodge et al., 2001; McKay, Doverspike, Bowen-Hilton, & Martin, 2002; J. L. Smith & White, 2002; Stricker & Ward, 2004). The empirical evidence for the moderating effects of test difficulty is mixed (O'Brien & Crandall, 2003; S. J. Spencer et al., 1999; Stricker & Bejar, 2004). Examining this moderator is also of importance to understanding

the potential generalizability of stereotype threat effects from laboratory to employment testing contexts, as employers often seek to use at least moderately difficult tests to increase selectivity.

Type of Stereotype

The theory of stereotype threat alludes to a universal reaction to stereotype threat by members of any stigmatized social group, implying that findings are generalizable from one stigmatized social group to another (see Steele et al., 2002). In this meta-analysis, we tested the viability of this assumption of universal effects by examining whether the activation of stereotype threat might differentially affect women and ethnic minority test takers in testing contexts. Stereotype relevance might differ because in the United States, where group differences are publicly acknowledged and discussed, advances in employment and higher education for minorities are affected by high-stakes testing (Sackett, Schmitt, Ellingson, & Kabin, 2001). However, the effect of math test performance on women's advancement opportunities is not likely to be as great, given that career opportunities that many women choose are not affected by math scores (see Halpern et al., 2007, for a review). Hence, a race-based stereotype regarding test performance might be more salient to test takers, or might lead to stronger reactions, than a gender-based math stereotype.

Prior Meta-Analysis

Although not their primary focus, Walton and Cohen (2003) conducted a meta-analysis on stereotype threat effects experienced by members of several stigmatized groups (e.g., ethnic minorities, women, older adults, individuals of lower socioeconomic status). They found a small overall effect size (mean $d = |.29|$, $k = 43$), which was moderated by stereotype relevance and domain identification. Walton and Cohen did not examine test difficulty as a moderator but opted instead to examine only studies that used a difficult test or performance situation. Although Walton and Cohen's results on stereotype threat effects are informative, there is room for improvement on their approach. We replicated and extended their work in five ways: (a) examining test difficulty as a moderator, (b) examining differential effects of different group-based stereotypes, (c) using a more fine-grained classification in examining stereotype-activating cues as a moderator, (d) considering nonindependent data points in a more appropriate manner, and (e) including substantially more studies.

First, Walton and Cohen (2003) reported significant heterogeneity tests of the observed effect sizes for all meta-analytic findings, meaning that there were other uninvestigated moderators that may further explain the variance in their findings. Therefore, we extended Walton and Cohen's work to include test difficulty and type of stereotype as potential moderators. Second, Walton and Cohen meta-analytically cumulated effect sizes from studies with various stereotype threats based on race, gender, age, or socioeconomic status and with various outcome measures. In this meta-analysis, we tested the viability of a universal stereotype threat reaction, considering potential differential effects of two types of stereotypes (race and gender) related to test performance.

Third, although Walton and Cohen (2003) did examine whether subtle versus explicit ways of activating and/or attempting to remove stereotype threat manipulations produced differing results, we extended their work by using a more concretely defined and detailed categorization for both threat activation and threat removal to better address generalizability issues to employment contexts. Fourth, Walton and Cohen's treatment of nonindependent data points was nonstandard: Studies in the data set that yielded hundreds of nonindependent data points each (i.e., identical or overlapping samples on multiple dependent measures; e.g., Stricker, 1998; Stricker & Ward, 1998) were given a weight of 0.5 in the effect size computation, but the reasons and/or implications of such a treatment in regard to the variance estimation of effect sizes (see Hunter & Schmidt, 1990) were neither explained nor discussed. Finally, Walton and Cohen used a small data set of experiments ($k = 43$) and the literature has grown substantially since their study. Less than one half of the studies the researchers meta-analyzed are related to the two group-based stereotypes of interest in employment settings, and many additional studies were included in the present meta-analysis.

In summary, to address these five extensions to Walton and Cohen's (2003) work, we conducted a hierarchical moderator meta-analysis, with each of the stated conceptual moderators—test difficulty, domain identification, and stereotype threat relevance (i.e., activation cues and removal strategies)—meta-analyzed across group-based stereotypes (i.e., race-based vs. gender-based). Furthermore, the primary focus of Walton and Cohen's meta-analysis was that the variance in observed between-subgroups mean test score differences (e.g., men vs. women, Black/Hispanic vs. White) might be partially accounted for by the debilitated performance of the target group members and partially by the comparison group's performance boost (stereotype lift). In this study, we directly examined the estimates of such between-group differences in test performance, considering potential differential effects for different stereotypes.

Method

Literature Search

We conducted a bibliographic search of electronic databases such as PsycINFO and PROQUEST using the combined keywords of *stereotype* and *threat* as search parameters for journal articles and dissertation abstracts dated between 1995 (the publication year of the seminal article by Steele and Aronson) and April 2006. A manual search was conducted by reviewing the reference lists of key articles to find additional citations of unpublished articles. The internet search engines of Google and Google Scholar were used to search for unpublished empirical articles of interest and/or for self-identified stereotype threat researchers. We sent the identified researchers with available e-mail addresses a "cold" e-mail, requesting manuscripts and/or working papers. We also posted the same request on various psychology list-servs. Furthermore, several prominent researchers in the stereotype threat area were contacted for unpublished manuscripts, in-press papers, as well as for other additional sources of research data on stereotype threat effects on cognitive ability test performance.

Inclusion Criteria

To be included in the data set, a research report first had to be an experiment designed to test Steele and Aronson's (1995)

within-subgroup performance interference hypothesis regarding stereotyped minorities' or women's cognitive ability test performance (quantitative, verbal, analytic, and/or nonverbal intelligence). Empirical studies that drew inferences from the theory of stereotype threat but were correlational or based on a different research framework were excluded (e.g., Ben-Zeev, Fein, & Inzlicht, 2005; Chung-Herrera, Ehrhart, Ehrhart, Hattrup, & Solamon, 2005; Cullen, Hardison, & Sackett, 2004; Good, Aronson, & Inzlicht, 2003; Inzlicht & Ben-Zeev, 2000, 2003; Osborne, 2001a, 2001b; Roberson, Deitch, Brief, & Block, 2003).

Second, a report had to operationalize test performance as the number of correct answers. (For studies that used a different index of performance, such as a ratio of correct answers to attempted problems, we converted these indexes from available reported information or contacted study authors for the information.) Third, an article had to yield precise statistics that were convertible to a weighted effect size *d* (e.g., mean test performance differences between women in a stereotype threat condition and those in a control condition). Finally, a report had to be written in English or could be translated into English.

Summary of the Meta-Analytic Data Set

The literature search initially identified a total of 151 published and unpublished empirical reports on stereotype threat effects. Of these reports, 75 were excluded because they did not meet one or more inclusion criteria.[1]

The remaining 76 reports contained 116 primary studies; 67 of which were from published peer-reviewed articles, and 65 of which included a comparison sample (e.g., Whites or men). The study database yielded a total of 8,277 data points from stereotyped groups and a total of 6,789 data points from comparison groups. Table 3 presents an overview of the characteristics of studies included in the full data set. Note that of the 43 primary studies in Walton and Cohen's (2003) meta-analysis, there were only 24 studies overlapping with our data set. In other words, about 79% of our data set (or $k = 92$) was nonoverlapping with that in Walton and Cohen's meta-analysis.

Treatment of Independent Data Points

When an article consisted of multiple single studies, we treated each study as an independent source of effect size estimates. When a single study included a fully replicated design across demographic subgroups (i.e., a conceptually equivalent but statistically independent design), we treated the data as if they were values from different studies. For example, means of cognitive ability test scores from all ethnic subgroups (e.g., Hispanic Americans and African Americans; Sawyer & Hollis-Sawyer, 2005) were statistically independent.

Treatment of Nonindependent Data Points

To be sensitive to potential problems caused by nonindependent data, for the eight studies with multiple measures of cognitive ability in our data set, we used only one independent estimate of effect size per study (i.e., an average effect size across cognitive ability tests for all subsamples per study). Nonindependent data points also occurred when the design of an experiment allowed for multiple effect size estimates to be computed across study conditions. For example, the research design of 14 studies in our data set consisted of one stereotype threat-activated condition and at least two or more stereotype threat-removed conditions and vice versa, resulting in multiple mean effect estimates per study. Following Webb and Sheeran's (2006) practice, we used the largest mean effect size estimate.

Treatment of Studies With a Control Condition

A nonexperimental (control) condition was defined as when a cognitive ability test was administered to test takers without any special directions. When a study design consisted of two conditions of stereotype threat manipulation (i.e., stereotype threat-activation, or STA, vs. stereotype threat-removal, or STR), the study contributed one effect size, $d_{\text{STA-STR}}$, to the data set. When a study design consisted of STA and control conditions, the study contributed one effect size, $d_{\text{STA-Control}}$. When all three conditions (STA, STR, control) were present in a study, the effect size $d_{\text{STA-STR}}$ was chosen to be cumulated. Although this approach might result in an upward bias in interpreting the magnitude of stereotype threat effects across studies (i.e., an estimate of $d_{\text{STA-STR}}$ might be larger than that of $d_{\text{STA-Control}}$), we erred on optimizing the probability of detecting stereotype threat effects and supporting the theory tenets, given the important social implications of stereotype threat.

Treatment of Studies With Stereotype Threat × Moderator Designs

For primary studies employing either the design of Stereotype Threat × Domain Identification or Stereotype Threat × Test Difficulty (e.g., Anderson, 2001), we split these studies into two or three independent subsamples according to the levels of domain identification or test difficulty as defined by the researchers themselves. Each subsample contributed an independent estimate of effect size to the database. For studies with a Stereotype Threat × Nontarget Moderator design (i.e., a moderating factor not investigated in the present meta-analysis), we gathered relevant statistical information across the stereotype threat conditions only.[2]

Treatment of Studies Where Gender Was Nested in Race

Schmader and Johns (2003, Study 2) and Stricker and Ward (2004, Study 2) conducted studies where test takers' gender was nested within race/ethnicity subgroups (i.e., White men/women vs. Latinos/Latinas vs. African American men/women). Because these

[1] The list of all excluded studies and reasons for exclusion is available from Hannah-Hanh D. Nguyen upon request.

[2] An exception was Keller's (2007) experiment, involving both domain identification and test difficulty. This study was coded as five separate substudies: two studies across levels of domain identification and three studies across levels of test difficulty. However, to avoid a violation of the independent error variance assumption, we cumulated only the estimates from Stereotype Threat × Test Difficulty studies for the overall mean effect size (because the domain identification subsets were nested within the subsets of test difficulty studies; Hunter & Schmidt, 1990). Further, each set of substudies across moderator levels contributed estimates to respective moderator analyses of domain identification or of test difficulty.

Table 3
Overview of the Meta-Analysis Database: Characteristics of Included Studies (K = 116)

Study no.	Study: Author	Study: No.	Status [a]	Stereotyped group	Sample size	Effect size	Comparison group included?	DI preselected?
1	Ambady et al. (2004)	1 of 2	Published	Female undergrads	20	−0.57	No	No
2	Ambady et al. (2004)	2 of 2	Published	Female undergrads	20	−0.67	No	No
3	**Anderson (2001)**	1 of 1	Unpublished	Female undergrads	604	−0.96	Yes	No
4	Aronson et al. (1999)	1 of 2	Published	White undergrads	23	−1.46	No	Yes
5	Aronson et al. (1999)	2 of 2	Published	White undergrads	26	−0.99	No	Yes
6	Aronson et al. (1999)	2B of 2	Published	White undergrads	23	−2.74	No	Yes
7	Bailey (2004)	1 of 1	Unpublished	Female undergrads	44	−0.09	Yes	No
8	**J. L. Brown et al. (n.d.)**	2 of 2	Unpublished	African American undergrads	28	−0.62	Yes	No
9	R. P. Brown & Day (2006)	1 of 1	Published	African American undergrads	34	0.38	Yes	No
10	R. P. Brown & Josephs (1999)	1 of 3	Published	Female undergrads	65	−0.09	Yes	No
11	R. P. Brown & Josephs (1999)	2 of 3	Published	Female undergrads	35	−1.17	Yes	No
12	R. P. Brown & Pinel (2003)	1 of 1	Published	Female undergrads	46	−0.53	No	Yes
13	Cadinu et al. (2003)	1 of 2	Published	Female undergrads (Italian)	25	−0.10	No	Yes
14	Cadinu et al. (2003)	1B of 2	Published	Female undergrads (Italian)	38	0.023	No	Yes
15	Cadinu et al. (2003)	2 of 2	Published	African American soldiers	50	−0.19	No	No
16	Cadinu et al. (2005)	1 of 1	Published	Female undergrads (Italian)	60	0.015	No	No
17	G. L. Cohen & Garcia (2005)	2 of 3	Published	African American undergrads	41	0.74	No	No
18	Cotting (2003)	1 of 1	Unpublished	Female undergrads	51	−0.53	No	No
19	Cotting (2003)	1B of 1	Unpublished	African American undergrads	55	0.44	No	No
20	**Davies et al. (2002)**	1 of 2	Published	Female undergrads	25	−0.71	Yes	Yes
21	Davies et al. (2002)	2 of 2	Published	Female undergrads	34	0.27	Yes	Yes
22	Dinella (2004)	1 of 1	Unpublished	Female high school students	232	0.11	Yes	No
23	Dodge et al. (2001)	1 of 1	Unpublished	African American undergrads	93	0.045	Yes	No
24	Edwards (2004)	1 of 1	Unpublished	Female undergrads and graduates	79	−0.78	No	No
25	Elizaga & Markman (n.d.)	1 of 1	Unpublished	Female undergrads	145	−0.38	No	No
26	Foels (1998)	1 of 1	Unpublished	Female undergrads	33	−0.78	No	No
27	Foels (1998)	1B of 1	Unpublished	Female undergrads	32	−0.3	No	No
28	**Foels (2000)**	1 of 1	Unpublished	Female undergrads	71	−0.7	Yes	No
29	Ford et al. (2004)	2 of 2	Published	Female undergrads	31	−1.7	No	No
30	Gamet (2004)	1 of 1	Unpublished	Female undergrads	51	−1.51	No	No
31	Gresky et al. (n.d.)	1 of 1	Unpublished	Female undergrads	23	−0.32	Yes	Yes
32	Gresky et al. (n.d.)	1B of 1	Unpublished	Female undergrads	37	−0.45	Yes	Yes
33	Guajardo (2005)	1 of 2	Unpublished	Female undergrads	56	0.03	Yes	No
34	Guajardo (2005)	2 of 2	Unpublished	Female undergrads	30	−0.52	Yes	No
35	Harder (1999)	1 of 2 (pilot)	Unpublished	Female undergrads	36	−0.66	Yes	No
36	Harder (1999)	2 of 2	Unpublished	Female undergrads	19	−0.04	No	Yes
37	Johns et al. (2005)	1 of 1	Published	Female undergrads	46	0.27	Yes	No
38	**Josephs et al. (2003)**	1 of 1	Published	Female undergrads	39	−0.79	Yes	Yes
39	**Keller (2002)**	1 of 1	Published	Female secondary school students (German)	37	−0.62	Yes	No
40	Keller (2007)	1 of 1	Published	Female secondary school students (German)	19	−0.10	Yes	No
41	Keller (2007)	1B of 1	Published	Female secondary school students (German)	18	−0.95	Yes	No
42	Keller (2007)	1C of 1	Published	Female secondary school students (German)	18	−1.11	Yes	No
43	Keller & Bless (n.d.)	2 of 3	Unpublished	Female undergrads (German)	66	−0.57	No	No
44	**Keller & Dauenheimer (2003)**	1 of 1	Published	Female secondary school students (German)	33	−0.38	Yes	No

(*table continues*)

Table 3 *(Continued)*

Study no.	Study: Author	Study: No.	Status [a]	Stereotyped group	Sample size	Effect size	Comparison group included?	DI preselected?
45	Lewis (1998)	1 of 1	Unpublished	African American undergrads	71	−0.12	Yes	No
46	Martens et al. (2006)	1 of 2	Published	Female undergrads	22	−0.67	No	Yes
47	**Martens et al. (2006)**	2 of 2	Published	Female undergrads	38	−0.11	Yes	No
48	Martin (2004)	2 of 2	Unpublished	African American undergrads	100	0.54	No	No
49	Martin (2004)	2B of 2	Unpublished	African American undergrads	102	−0.93	No	No
50	Marx & Stapel (2005)	1 of 1	Published	Female undergrads (Dutch)	48	−1.07	Yes	No
51	Marx & Stapel (2006)	1 of 3	Published	Female undergrads (Dutch)	29	−1.22	Yes	No
52	Marx & Stapel (2006)	3 of 3	Published	Female undergrads (Dutch)	28	−1.24	Yes	No
53	Marx et al. (2005)	3 of 4	Published	Female undergrads (Dutch)	27	0.56	No	No
54	Marx et al. (2005)	3B of 4	Published	Female undergrads (Dutch)	25	−0.16	No	No
55	Marx et al. (2005)	4 of 4	Published	Female undergrads (Dutch)	25	−0.14	No	No
56	McFarland, Kemp, et al. (2003)	1 of 1	Unpublished	Female undergrads	126	−0.035	No	No
57	McFarland, Lev-Arey, & Ziegert (2003)	1 of 1	Published	African American undergrads	50	−0.22	Yes	No
58	McIntyre et al. (2003)	1 of 2	Published	Female undergrads	116	−0.52	Yes	No
59	McIntyre et al. (2003)	2 of 2	Published	Female undergrads	74	−0.49	Yes	No
60	McIntyre et al. (2005)	1 of 1	Published	Female undergrads	81	−0.98	Yes	No
61	**McKay (1999)**	1 of 1	Unpublished	African American undergrads	103	0.91	Yes	No
62	Nguyen et al. (2003)	1 of 1	Published	African American undergrads	80	0.05	Yes	No
63	Nguyen et al. (2004)	1 of 1	Unpublished	Female undergrads	114	0.057	Yes	No
64	O'Brien & Crandall (2003)	1 of 1	Published	Female undergrads	58	−0.305	Yes	No
65	Oswald & Harvey (2000–2001)	1 of 1	Published	Female undergrads	34	−0.06	No	No
66	Pellegrini (2005)	1 of 1	Unpublished	Hispanic undergrads (female)	60	−1.03	No	No
67	Philipp & Harton (2004)	1 of 1	Unpublished	Female undergrads	38	−1.21	Yes	No
68	Ployhart et al. (2003)	1 of 1	Published	African American undergrads	48	−0.59	Yes	No
69	Ployhart et al. (2003)	1B of 1	Published	African American undergrads	48	−0.57	Yes	No
70	Prather (2005)	1 of 1	Unpublished	Female undergrads	114	−0.67	No	No
71	Rivadeneyra (2001)	1 of 1	Unpublished	Latino high school students	116	−0.96	No	No
72	H. E. S. Rosenthal & Crisp (2006)	2 of 3	Published	Female undergrads (British)	24	−1.46	No	No
73	H. E. S. Rosenthal & Crisp (2006)	3 of 3	Published	Female undergrads (British)	29	−0.99	No	No
74	H. E. S. Rosenthal & Crisp (2006)	3B of 3	Published	Female undergrads (British)	27	−2.74	No	No
75	**Salinas (1998)**	1 of 2	Unpublished	Mexican American undergrads	27	−0.09	Yes	No
76	**Salinas (1998)**	2 of 2	Unpublished	Mexican American undergrads	56	0.38	Yes	No
77	Sawyer & Hollis-Sawyer (2005)	1 of 1	Published	African American undergrads	66	−0.09	Yes	No
78	Sawyer & Hollis-Sawyer (2005)	1B of 1	Published	Hispanic undergrads	47	−1.17	Yes	No
79	Schimel et al. (2004)	2 of 3	Published	Female undergrads	46	−0.53	No	Yes
80	**Schmader (2002)**	1 of 1	Published	Female undergrads	32	−0.19	Yes	No
81	Schmader & Johns (2003)	1 of 3	Published	Female undergrads	28	−0.62	Yes	Yes
82	Schmader & Johns (2003)	2 of 3	Published	Latino American undergrads	33	−0.10	Yes	No

Table 3 *(Continued)*

Study no.	Study: Author	Study: No.	Status [a]	Stereotyped group	Sample size	Effect size	Comparison group included?	DI preselected?
83	Schmader & Johns (2003)	3 of 3	Published	Female undergrads	28	0.023	No	Yes
84	Schmader et al. (2004)	2 of 2	Published	Female undergrads	68	0.015	No	No
85	Schneeberger & Williams (2003)	1 of 1	Unpublished	Female undergrads	61	0.74	Yes	No
86	**Schultz et al. (n.d.)**	1 of 2	Unpublished	Hispanic American undergrads	44	−0.533	Yes	No
87	**Schultz et al. (n.d.)**	2 of 2	Unpublished	Hispanic American undergrads	40	0.44	Yes	No
88	Seagal (2001)	6 of 6	Unpublished	African American and Latino undergrads	101	−0.71	Yes	No
89	**Sekaquaptewa & Thompson (2002)**	1 of 1	Published	Female undergrads	80	0.27	Yes	No
90	C. E. Smith & Hopkins (2004)	1 of 1	Published	African American undergrads	160	0.11	No	No
91	J. L. Smith & White (2002)	1 of 2	Published	White undergrads (male)	47	0.045	No	No
92	J. L. Smith & White (2002)	2 of 2	Published	Female undergrads	23	−0.78	No	No
93	**S. J. Spencer et al. (1999)**	2 of 3	Published	Female undergrads	30	−0.78	Yes	Yes
94	S. L. Spencer (2005)	1 of 1	Unpublished	Female undergrads	40	−0.38	No	No
95	Spicer (1999)	2 of 2	Unpublished	African American undergrads	39	−0.3	No	Yes
96	Spicer (1999)	2B of 2	Unpublished	African American undergrads	39	−0.7	No	Yes
97	**Steele & Aronson (1995)**	1 of 4	Published	African American undergrads	38	−1.7	Yes	No
98	**Steele & Aronson (1995)**	2 of 4	Published	African American undergrads	20	−1.51	Yes	No
99	**Steele & Aronson (1995)**	4 of 4	Published	African American undergrads	22	−0.32	Yes	No
100	**Sternberg et al. (n.d.)**	1 of 2	Unpublished	Female high school students	27	−0.45	Yes	No
101	**Sternberg et al. (n.d.)**	2 of 2	Unpublished	Female high school students	96	0.03	Yes	No
102	**Stricker & Ward (2004)**	1 of 2	Published	African American high school students	122	−0.52	Yes	No
103	**Stricker & Ward (2004)**	1B of 2	Published	Female high school students	730	−0.66	Yes	No
104	**Stricker & Ward (2004)**	2 of 2	Published	African American undergrads	468	−0.04	Yes	No
105	Tagler (2003)	1 of 1	Unpublished	Female undergrads	136	0.27	Yes	No
106	van Dijk et al. (n.d.)	1 of 1	Unpublished	Female undergrads (Dutch)	38	−0.79	Yes	No
107	van Dijk et al. (n.d.)	1B of 1	Unpublished	Female undergrads (Dutch)	38	−0.57	Yes	No
108	von Hippel et al. (2005)	4 of 4	Published	White undergrads (Australian)	56	−0.38	No	No
109	Walsh et al. (1999)	2 of 2	Published	Female undergrads (Canadian)	96	−0.62	Yes	No
110	Walters (2000)	1 of 2	Unpublished	African American undergrads	49	−0.10	No	Yes
111	Wicherts et al. (2005)	1 of 3	Published	Minority high school students (Dutch)	138	−0.95	Yes	No
112	Wicherts et al. (2005)	3 of 3	Published	Female undergrads (Dutch)	95	−1.11	Yes	No
113	Wout et al. (n.d.)	1 of 4	Unpublished	African American undergrads	57	−0.12	No	No
114	Wout et al. (n.d.)	2 of 4	Unpublished	African American undergrads	29	−0.67	No	No
115	Wout et al. (n.d.)	3 of 4	Unpublished	African American undergrads	24	−0.11	No	No
116	Wout et al. (n.d.)	4 of 4	Unpublished	African American undergrads	26	0.54	No	No

Note. Studies presented in bold font ($k = 24$) are those that overlap with Walton and Cohen's (2003) data set. DI = domain identification.
[a] Published articles are those that appeared in peer-reviewed journal articles, including those in press; unpublished articles refer to dissertations, theses, conference papers, and working manuscripts.

studies mainly aimed at examining race-based stereotype threat effects, only the effect sizes as a function of race/ethnicity and stereotype threat activation contributed data points to the overall meta-analytic data set.[3]

Treatment of Studies With Large Sample Sizes

There were a few studies with substantially larger sample sizes than those in the majority of other studies in the meta-analysis (e.g., Anderson, 2001; Stricker & Ward, 2004, Studies 1 and 2). Meta-analytic results with and without the estimates of effect sizes from these studies were similar so we report findings including all studies.[4]

Coding of Studies

We coded levels of test difficulty (e.g., difficult, easy) and domain identification based on investigators' description and/or evidence in source reports. When test takers' domain identification was on a continuous scale in some reports (e.g., Bailey, 2004; Edwards, 2004; Ployhart, Ziegert, & McFarland, 2003), we did not find sufficient statistical information to convert the data into categorical subgroup data.

For the type of stereotype threat-activation cues, we coded the data on three levels: blatant, moderately explicit, and subtle (see Table 1). A similar coding practice was used for the moderator of stereotype threat-removal strategies (see Table 2). We coded the condition of a cognitive ability test without any special directions as a control condition, following Fisher's (1925) definitions of research groups.

Studies were also coded for demographic characteristics of samples, such as whether the stereotype activated was based on race or gender. Because stereotype threat manipulation and test takers' race/ethnicity or gender were correlated in many studies, a series of hierarchical moderator analyses was needed to assess the potential impact of confounding on the moderator analyses. To accomplish this, we first broke down the stereotype threat effect estimates for manipulation conditions by test takers' race/ethnicity or by gender, and then we undertook a moderator analysis by the race/ethnicity of test takers (minorities vs. Whites) or by gender (women vs. men) within the stereotype threat manipulation conditions.

Coders and Agreement

Three coders (Hannah-Hanh D. Nguyen and two trained assistants) coded target variables; each study was coded by at least two coders and periodically cross-checked. Objective statistics and continuous variables coded include sample size, variable cell means and standard deviations, *t*-test values, and/or *F*-test values. When there was insufficient statistical information to compute the estimate of effect size for a study, coders tried to contact source authors for additional information before marking the data as missing. The interrater agreement rates for continuous and objective variables were between 91% and 100%; disagreements were discussed and resolved. For categorical variables, we computed a series of interrater agreement index kappas following Landis and Koch's (1977) rules. Kappa values ranged from 0.49 to 0.95, indicating moderately good to very satisfactory interrater agreement levels. The lower kappa values were associated mainly with the classification of stereotype threat conditions given how researchers might differ in labeling these conditions in the primary studies (e.g., an STR condition might be referred to as a control group). Disagreements were discussed and resolved.

Meta-Analytic Procedure

We employed the meta-analysis procedure of Hunter and Schmidt (1990, 2004) and conducted an overall meta-analysis to cumulate findings across studies, as well as conducting separate meta-analyses with subsets of studies to examine moderator effects. Specifically, we cumulated the average population effect size δ (corrected for measurement error) and computed variance var(δ) across studies, weighted by sample size, using the Meta-Analysis of *d*-Values Using Artifact Distributions software program (Schmidt & Le, 2005).

We converted descriptive statistics, *t*-test estimates, or *F*-test estimates into the effect size Cohen's *d* (i.e., mean difference between cell means in standard score form) using Thalheimer and Cook's (2002) software program, which was based on Rosnow and Rosenthal's (1996) and Rosnow, Rosenthal, and Rubin's (2000) formulas. Reliability information on cognitive ability tests was sporadically reported in the source reports; therefore, study effect sizes could not be corrected individually for measurement error; we used artifact distributions instead.

Following Hunter and Schmidt's (1990, 2004) recommendation, to judge whether substantial variation due to moderators exists, we used the standard deviation, SD_δ, estimated from var(δ) to construct the 90% credibility intervals (CrI) around δ as an index of true variance due to moderators (Whitener, 1990). When the credibility intervals were large (e.g., greater than 0.11; Koslowsky & Sagie, 1993) and overlapped zero, we interpreted them as indicating the presence of true moderators and inconclusive meta-analytic findings. V%, or the ratio of sampling error variance to the observed variance in the corrected effect size, was also calculated. When most of the observed variance is due to sampling error (i.e., V% $>$75%), it is less likely that a true moderator exists and explains the observed variance in effect sizes. Following recent meta-analytic practices (e.g., Roth, Bobko, & McFarland, 2005; Zhao & Seibert, 2006), we also reported 95% confidence intervals (CI; the likely amount of error in an estimate of a single value of mean effect size due to sampling error) in our meta-analyses. The interpretation of 95% CI excluding zero means that we can be 95% confident that mean δ is not zero.

---

[3] Stricker and Ward (2004, Study 1) was an exception to this rule, as the stereotype threat cues were both race-based and gender-based (i.e., race and gender inquiries prior to tests). Therefore, it was conceptually sound to code the outcomes of this study separately as a function of race or gender; that means the study contributed some nonindependent estimates of effect size to the data set. However, the proportion of these nonindependent data points was not large in the data set (i.e., 842 data points altogether, or 10.7%).

[4] The results from meta-analytic data sets without large sample-size studies are available from Hannah-Hanh D. Nguyen upon request.

Testing for Publication Bias

Our meta-analysis database consisted of a relatively balanced number of published and unpublished reports (54.8% and 45.2%, respectively). Nevertheless, fail-safe *N* analyses were conducted to test a potential file-drawer bias in each meta-analysis. Hunter and Schmidt (1990) provided a formula to calculate fail-safe *N*, which indicates the number of missing studies with zero-effect size that would have to exist to bring the mean effect size down to a specific level. In the present review, mean d_{critical} was arbitrarily set to 0.10, which constitutes a negligible effect size (see J. Cohen, 1988). In the interest of space, we discussed file-drawer analyses only where potential problems were indicated.

Additionally, we used Light and Pillemer's (1984) "funnel graph" technique of plotting sample sizes versus effect sizes. In the absence of bias, the plot should resemble a symmetrical inverted funnel. There may be a problem of publication bias when there is a cutoff of small effects for studies with a small sample size. In other words, because only large effects reach statistical significance in small samples, a publication bias or other types of location biases are present when only large effects are reported by studies with a small sample size (i.e., an asymmetrical and skewed shape). On the contrary, there are no biases if an exclusion of null results is not visible on the funnel graph.

Results

Within-Group Stereotype Threat Effects

Overall effect. Tables 4 and 5 present the results of our hierarchical moderating meta-analyses. The overall effect size was mean $d = |.26|$ ($K = 116$, $N = 7{,}964$; see Table 4), which was comparable to the finding of mean $d = |.29|$ in Walton and Cohen's (2003) study. However, the variance of effect sizes was nonzero (V% was about 26%) and the CrI shows that there was a 90% probability that the true effect size was between −0.85 and 0.29—a range of *d* values overlapping zero. These values indicate that true moderators existed.

Group-based stereotypes. We separately analyzed the mean effect sizes for studies with an ethnic/racial group-based stereotype of intellectual inferiority and studies with a gender-based stereotype of mathematical ability inferiority, something not considered in Walton and Cohen (2003). Lines 3 and 4 in Table 4 show differential stereotype threat effects in that the mean effect size was greater in the ethnicity/race-based stereotype subset than in the gender-based stereotype subset (mean *d*s = |.32| and |.21|, respectively). The nonoverlapping 95% CIs indicate reliable effects. Although the subset variance values decreased compared with the variance of the entire set of *d* values, they were still

Table 4
Hierarchical Moderator Analyses of Domain Identification and Test Difficulty

Variable	*K*	*N*	Mean *d*	Var *d*	Var *e*	Mean δ	Var δ	% var *SE*	V%	90% CrI	95% CI	Fail-safe *N*
Overall effect size	116	7,964	−.258	.227	.060	−.281	.198	26.26	26.33	−0.85, 0.29	−0.38,−0.16	415
Group-based stereotype												
Race/ethnicity	44	2,988	−.324	.186	.060	−.353	.149	32.50	32.63	−0.85, 0.14	−0.49,−0.19	187
Female	72	4,935	−.208	.241	.059	−.227	.216	24.64	24.68	−0.82, 0.37	−0.35,−0.08	222
Test difficulty by group-based stereotype												
Overall												
Difficult	48	2,161	−.394	.396	.092	−.429	.361	23.20	23.29	−1.20, 0.34	−0.62,−0.02	237
Moderately difficult	24	1,560	−.190	.153	.063	−.208	.107	40.86	40.92	−0.63, 0.21	−0.38,−0.02	70
Easy	9	308	.083	.199	.119	.091	.095	59.74	59.74	−0.30, 0.49	−0.23, 0.40	2
Minority test takers												
Difficult	12	549	−.425	.157	.091	−.464	.078	57.84	58.11	−0.82,−0.11	−0.71,−0.18	63
Moderately difficult	10	647	−.181	.073	.063	−.198	.012	86.27	86.37	−0.34,−0.06	−0.38, 0.00	28
Female test takers												
Difficult	33	1,508	−.363	.500	.090	−.395	.487	18.04	18.10	−1.29, 0.50	−0.66,−0.10	153
Moderately difficult	13	890	−.175	.195	.059	−.191	.162	30.34	30.38	−0.71, 0.32	−0.45, 0.09	36
Easy	9	308	.083	.199	.119	.091	.095	59.74	59.74	−0.30, 0.49	−0.23, 0.40	2
Domain identification by group-based stereotype[a]												
Overall												
High	12	478	−.316	.210	.103	−.344	.127	49.21	49.32	−0.80, 0.11	−0.63,−0.03	50
Medium	9	313	−.371	.290	.120	−.404	.203	41.00	41.10	−0.98, 0.17	−0.79, 0.01	42
Female test takers[b]												
High	9	380	−.287	.201	.097	−.313	.123	48.44	48.54	−0.76, 0.14	−0.63, 0.03	35
Medium	6	212	−.518	.204	.119	−.565	.100	58.29	58.59	−0.97,−0.16	−0.96,−0.12	37

Note. *K* = Number of effect sizes (*d* values); *N* = total sample size; mean *d* = sample size weighted mean effect size; var *d* = sample size weighted observed variance of *d* values; var *e* = variance attributed to sampling error variance; mean δ = mean true effect size; var δ = true variance of effect sizes; % var *SE* = percent variance in observed *d* values due to sampling error variance; V% = percent variance accounted for in observed *d* values due to all corrected artifacts; 90% CrI = 90% of mean δ (credibility interval); 95% CI = 95% of mean δ (confidence interval); fail-safe *N* = number of missing studies averaging null findings that would be needed to bring mean *d* down to .10, from Hunter and Schmidt's (1990) effect size file-drawer analysis.
[a] Domain identification levels: *High* = strongly identified with academic or cognitive ability domains; *Medium* = moderately identified. [b] Only the subsets of female test takers were meta-analyzed here because there were insufficient race-based studies contributing effect size estimates (see Arthur et al., 2003).

Table 5
Hierarchical Moderator Analyses of Stereotype Threat Relevance (Activation and Removal)

Variable	*K*	*N*	Mean *d*	Var *d*	Var *e*	Mean δ	Var δ	% var *SE*	V%	90% CrI	95% CI	Fail-safe *N*
Stereotype threat-activating (STA) cues by group-based stereotype												
Minority test takers												
Overall	38	2,724	−.295	.185	.057	−.322	.151	30.89	31.00	−0.82, 0.18	−0.47, −0.15	150
Blatant	6	436	−.405	.077	.057	−.441	.024	73.37	73.86	−0.64, −0.24	−0.68, −0.16	30
Moderately explicit	7	277	−.639	.058	.108	−.696	0	100	100		−0.89, −0.45	52
Subtle	25	2,011	−.224	.201	.051	−.244	.179	25.10	25.16	−0.79, 0.30	−0.44, −0.03	81
Female test takers												
Overall	73	4,947	−.205	.240	.060	−.223	.214	25.06	25.10	−0.82, 0.37	−0.35, −0.08	223
Blatant	22	1,279	−.172	.390	.070	−.188	.381	17.90	17.92	−0.98, 0.60	−0.47, 0.11	60
Moderately explicit	20	1,138	−.184	.181	.072	−.201	.130	39.63	39.67	−0.66, 0.26	−0.41, 0.02	57
Subtle	32	2,564	−.239	.193	.051	−.261	.169	26.42	26.49	−0.8, 0.27	−0.43, −0.07	108
Stereotype threat-removal (STR) strategies by group-based stereotype												
Minority test takers												
Overall	30	1,661	−.415	.245	.075	−.452	.201	30.53	30.69	−1.03, 0.12	−0.65, −0.22	155
Explicit	5	157	−.800	.053	.140	−.870	0	100	100		−1.09, −0.58	45
Subtle	25	1,504	−.375	.248	.068	−.408	.213	27.62	27.75	−1.00, 0.18	−0.62, −0.16	119
Female test takers												
Overall	61	3,310	−.233	.337	.075	−.254	.311	22.34	22.38	−0.97, 0.46	−0.41, −0.07	203
Explicit	31	1,626	−.135	.285	.078	−.147	.245	27.17	27.19	−0.78, 0.49	−0.35, 0.07	73
Subtle	30	1,684	−.329	.368	.731	−.358	.350	19.88	19.95	−1.12, 0.40	−0.59, 0.36	129

Note. *K* = Number of effect sizes (*d* values); *N* = total sample size; mean *d* = sample size weighted mean effect size; var *d* = sample size weighted observed variance of *d* values; var *e* = variance attributed to sampling error variance; mean δ = mean true effect size; var δ = true variance of effect sizes; % var *SE* = percent variance in observed *d* values due to sampling error variance; V% = percent variance accounted for in observed *d* values due to all corrected artifacts; 90% CrI = 90% of mean δ (credibility interval); 95% CI = 95% of mean δ (confidence interval); fail-safe *N* = number of missing studies averaging null findings that would be needed to bring mean *d* down to .10, from Hunter and Schmidt's (1990) effect size file-drawer analysis.

nonzero. There was a 90% probability that the true mean effect size of the race-based subset was between a zero-included range of −0.85 and 0.14, whereas there was a 90% probability that the true mean effect size of the gender-based subset was between a zero-included range of −0.82 and 0.37. Subset V% values were 33% and 25% for minorities and women subsets, respectively. Taken together, these values suggested further moderator meta-analyses.

Test difficulty. We next meta-analyzed test difficulty as a moderator across stereotypes. Table 4 shows that stereotype-threatened minorities performed more poorly than did nonthreatened minorities when cognitive ability tests were highly difficult (mean $d = |.43|$, a reliable effect with a nonoverlapping 95% CI) than when tests were moderately difficult (mean $d = |.18|$, zero-included 95% CI). (There were no studies using easy tests to investigate stereotype threat effects among minority test takers.) The credibility intervals did not overlap zero, meaning that these findings were conclusive.

Similar to ethnic minority test takers, women underperformed when a math test was highly difficult (mean $d = |.36|$), more so than when a math test was moderately difficult (mean $d = |.18|$). However, when the test was easy, women tended to improve their test performance slightly (mean $d = |.08|$). The nonoverlapping 95% CIs indicate that only the high difficulty finding was reliable. The V% and zero-included 90% CrIs indicate that true moderators still existed. The smaller file-drawer *N* values indicate that the findings on medium and low difficulty levels were not conclusive.

Domain identification. As shown in Table 4, there were no discernible differences in stereotype threat effects between highly and moderately domain-identified samples (mean ds = |.32| and |.37|, ks = 12 and 9, respectively). The results were inconsistent with those in Walton and Cohen's (2003) meta-analysis (i.e., mean $d_{\text{domain-identified}} = |.68|$ vs. mean $d_{\text{not-identified}} = |.29|$). However, our hierarchical meta-analyses with studies on gender-based stereotypes showed a different pattern of findings: Highly math-identified women experienced smaller stereotype threat effects (mean $d = |.29|$, $k = 9$) than did moderately low math-identified women (mean $d = |.52|$, $k = 6$). The nonoverlapping 95% CIs show that only the result for the moderate domain identification subset was reliable. Smaller V% values and zero-included CrIs indicate that these findings were inconclusive because of other moderators that may explain the variance in the data over and above study artifacts. Furthermore, lower fail-safe *N* values show that these meta-analytic findings might not be conclusive. There were only three studies of ethnic minorities in each level, so the analyses for minority subsets were not conducted (see Arthur, Bennett, Edens, & Bell, 2003).

Stereotype threat relevance: Threat-activating cues. Table 5 shows that stereotype threat-activating cues affected minority test takers' test performance (mean $d = |.30|$) more than that of women (mean $d = |.21|$). However, Table 5 shows that when the negative stereotype was based on race, the largest mean effect size was produced for moderately explicit threat-activating cues (mean $d = |.64|$) compared with other types of threat-activating cues (blatant cues: mean $d = |.41|$; subtle cues: mean $d = |.22|$). As the V% and CI values in Table 5 indicate, the findings for the race-based subset for blatant cue conditions were conclusive, but for the subtle cues subset further moderator analyses were still needed.

As shown in Table 5, the negative stereotype concerning women's mathematical ability yielded a different pattern of findings from the race-based stereotype. Studies using moderately explicit cues yielded a comparable mean effect size (mean $d = |.18|$) to that in studies using blatant cues (mean $d = |.17|$). The zero-included 95% CIs indicate nonreliable effects. Studies employing subtle stereotype threat cues yielded the largest mean effect size (mean $d = |.24|$), and the nonoverlapping 95% CI indicates a reliable effect. (The effect size differences among these subsets were trivial though.) V% values and the 90% CrIs suggest that other moderators would further explain the variance in these *d* values to reach conclusive findings. Our findings show a more complex pattern than the results in Walton and Cohen's (2003) meta-analysis: Walton and Cohen had found that overall, explicit stereotype threat activation produced a greater effect size (mean $d = |.57|$) than when the activation was not explicit (mean $d = |.27|$). Their findings were consistent with those found for minorities in the present study but inconsistent with those found for female test takers.

Stereotype threat relevance: Threat-removal strategies. Table 5 shows that stereotype threat-removal strategies differentially affected minority test takers (mean $d = |.42|$, nonoverlapping 95% CI) and female test takers' performance (mean $d = |.23|$, nonoverlapping 95% CI). Stereotype threat-removal strategies seemed to work better on women's math test performance than on minorities' test performance, at least at the mean level.

Hierarchical meta-analyses showed that minority test takers seemed to benefit more from subtle or indirect threat-removal strategies than from direct, explicit ones (i.e., smaller stereotype threat effects: mean $d = |.38|$ and mean $d = |.80|$, respectively). Study artifacts explained all variance in the explicit-removal strategy subset of *d* values, indicating that this finding of interest was conclusive, although one should be cautious about generalizing this finding because there was a smaller fail-safe *N* of 45. However, study artifacts explained only 28% of the variance in the subtle removal-strategy subset of *d* values, and the 90% CrI overlapped zero, indicating true moderators and inconclusive findings.

Table 5 also reveals that female test takers benefited more from explicit stereotype threat-removal strategies (mean $d = |.14|$, zero-included 95% CI) than from subtle strategies (mean $d = |.33|$, nonoverlapping 95% CI). Low V% values and zero-overlapping 90% CrIs indicate the effects of other true moderators. Again, the pattern of Walton and Cohen's (2003) meta-analytic findings was more consistent with our result pattern for minorities than with that for women (i.e., Walton and Cohen found that studies explicitly removing stereotype threat produced greater stereotype threat, mean $d = .45$, than studies that did not, mean $d = .20$).

Supplemental Bias Analysis

As shown in Figure 1, the funnel plot for the full meta-analytic data set resembles a relatively symmetrical inverted funnel, indicating the absence of publication bias in the data set. Further, the relationship between effect size estimates and study sample sizes was positive and statistically significant ($r = .23$, $p < .05$). When four primary studies, each with a sample size larger than 200, were excluded from the data set (Anderson, 2001; Dinella, 2004; Stricker & Ward, 2004, Study 1B & Study 2), a similar pattern of findings was also found.

One additional question is whether studies that yielded either positive effect size estimates or estimates clustering around the zero point ($k = 29$, 25% of the data set) have differential characteristics from studies where the *d* values supported the hypothesis of performance interference (i.e., a negative effect size). Examining the general characteristics of samples in subsets of studies at different levels of effect size estimates, we found no clearly defining characteristics that might distinguish studies that found no stereotype threat effects or positive effects from studies that found the effects.[5]

Between-Group Stereotype Threat Effects

As shown in Table 6, the overall between-group effect values increased from a mean effect size $d = |.44|$ in test-only, control conditions to a mean effect size $d = |.53|$ in stereotype threat-activated conditions. When interventions or threat-removal strategies were implemented, stereotyped test takers underperformed on cognitive ability tests compared with reference test takers (mean $d = |.28|$). The nonoverlapping 95% CIs indicate reliable effects. The 90% CrI values and V% estimates indicate true moderator effects. The zero-included CrIs for mean *d*s in stereotype threat-activated conditions and stereotype threat-removed conditions mean that these findings were not conclusive. The credibility interval for mean *d* in control conditions did not overlap zero, however.

Subsequent hierarchical meta-analyses across group-based stereotypes were conducted. As shown in Table 6, in control conditions, ethnic minority test takers underperformed compared with majority test takers: The between group mean *d* is |.56|. The nonoverlapping 95% CI indicates a reliable effect. On the average, ethnic minority test takers' test scores were approximately at the 30th percentile of majority groups' mean test scores, which is relatively consistent with the literature (e.g., the overall mean standardized differences for *g* are 1.10 for the Black–White comparison and 0.72 for the Latino–White comparison; Roth, Bevier, Bobko, Switzer, & Tyler, 2001). Note that study artifacts explained all observed variance in the *d* values in this subset, suggesting that no further moderator analyses should be conducted for this subset.

Table 6 also reveals that, in control conditions, women underperformed compared with men on mathematical ability tests; the between-group mean effect size was mean $d = |.26|$, which is consistent with the literature (see a review by Halpern et al., 2007). The nonoverlapping 95% CI indicates a reliable effect. On average, women's mean math test scores were approximately at the 40th percentile of men's mean math test scores. Study artifacts explained all of the variance in *d* values, suggesting no other moderators for this subset. Although the fail-safe *N* value was not very large (47), similar overall gender differences in math test

[5] Whereas there was only one non-American sample in the "non-effect" group of studies (3.5%), there were 23 non-American samples (26.5%) in the "stereotype threat effect" group, suggesting that non-American authors (or American authors who used non-American samples) might be more likely to publish significant findings that were consistent with the hypothesis of performance interference in American journals than non-significant findings or findings that were contradictory to the hypothesis.

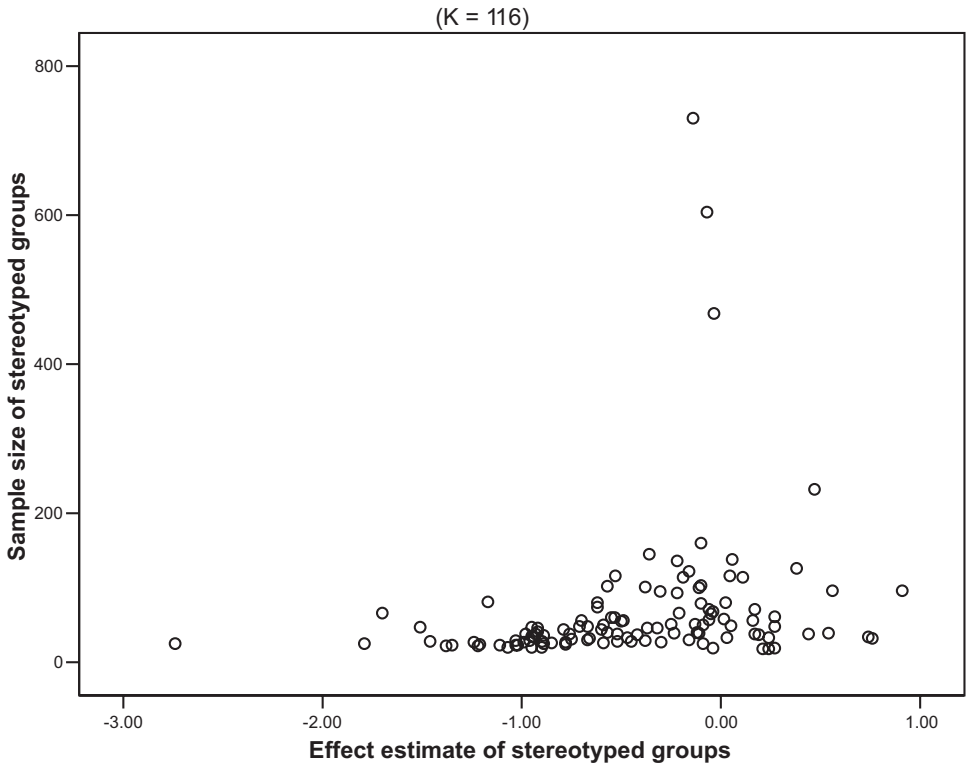

Figure 1. The funnel graph of stereotype threat effects on target test takers' cognitive ability test performance. The effect size estimates are plotted against study sample sizes ($r = .23$, $p < .05$).

performance were observed in the testing literature; therefore, a file-drawer problem was possible but not plausible.

In stereotype threat-activated conditions, ethnic minority test takers underperformed compared with majority test takers, and the between-group mean effect size was mean $d = |.67|$. The nonoverlapping 95% CI indicates a reliable effect. When stereotype threat was activated, on average, ethnic minority test takers' mean scores were approximately at the 25th percentile of majority groups' mean test scores, which was worse than the results in control conditions. Study artifacts explained about 62% of the observed variance in *d* values, suggesting that further moderator effects should be investigated for this subset. The 90% CrI did not overlap zero, indicating a conclusive result.

Also in stereotype threat-activated conditions, women underperformed compared with men on mathematical ability tests, and the between-group mean effect was mean $d = |.39|$. The nonoverlapping 95% CI indicates a reliable effect. In other words, when stereotype threat was activated, women's mean math test scores were approximately at the 34th percentile of men's mean scores. Study artifacts explained only 26% of the variance in *d*, suggesting other true moderator effects. The zero-included credibility interval indicates an inconclusive finding.

In stereotype threat-removed conditions, ethnic minority test takers underperformed compared with majority test takers; the between-group mean effect size was mean $d = |.38|$, a reliable effect based on the nonoverlapping 95% CI value. On average, when stereotype threat effects were removed, ethnic minority test takers' mean test scores were approximately at the 34th percentile of majority groups' mean test scores. However, study artifacts explained about 38% of the observed variance in the *d* values in the subset findings, suggesting moderator effects. The zero-included 90% CrI indicates that this finding was not conclusive.

Furthermore, in stereotype threat-removal conditions, women underperformed compared with men on mathematical ability tests, and the between-group mean effect size was mean $d = |.23|$, a reliable effect based on the nonoverlapping 95% CI value. On the average, women's math test scores were approximately at the 41st percentile of men's mean math scores when stereotype threat-removal strategies were implemented. Study artifacts explained 73% of the variance in *d* values, suggesting no other moderators. The 90% CrI did not include zero, indicating a meaningful effect.

Discussion

In integrating more than 10 years of experimental research on stereotype threat effects on stereotyped test takers' cognitive ability test performance, we found that the overall performance of stereotyped test takers might suffer from a situational stereotype threat: Our overall effect size of |.26| was consistent with the finding in Walton and Cohen (2003). However, there was considerable variability in the effect sizes (i.e., one fourth of studies in our data set showed zero or positive effects); therefore we examined several conceptual moderators (test difficulty, domain identification, and the activation or removal of stereotype relevance), each of which was analyzed separately for race-based and gender-

Table 6
Hierarchical Meta-Analytic Findings of Between-Group Mean Test Performance Across Stereotype Threat Levels

Condition	*K*	*N*	Mean *d*	Var *d*	Var *e*	Mean δ	Var δ	% var *SE*	V%	90% CrI	95% CI	Fail-safe *N*
Control												
Overall	23	3,620	−.440	.080	.030	−.480	.060	34.88	35.48	−0.79, −0.17	−0.61, −0.31	124
Minority vs. majority[a]	10	1,695	−.564	.0195	.025	−.615	0	100	100		−0.71, −0.47	66
Women vs. men	13	1,803	−.264	.025	.029	−.288	0	100	100		−0.38, −0.17	47
Stereotype threat-activating												
Overall	62	5,937	−.530	.160	.040	−.580	.130	27.80	28.22	−1.05, −0.11	−0.69, −0.42	391
Minority vs. majority[a]	23	2,498	−.686	.065	.039	−.747	.065	60.29	61.95	−0.97, .53	−0.86, −0.57	181
Women vs. men	39	3,330	−.392	.186	.048	−.428	.163	26.08	26.27	−0.94, .09	−0.58, −0.24	192
Stereotype threat-removing												
Overall	46	2,603	−.280	.130	.070	−.300	.070	54.76	54.90	−0.64, .04	−0.41, −0.17	175
Minority vs. majority[a]	14	848	−.377	.182	.068	−.410	.135	37.42	37.60	−0.88, .06	−0.65, −0.13	67
Women vs. men	32	1,765	−.232	.101	.074	−.252	.032	73.02	73.15	−0.48, −0.02	−0.37, −0.11	106

Note. *K* = Number of effect sizes (*d* values); *N* = total sample size; mean *d* = sample size weighted mean effect size; var *d* = sample size weighted observed variance of *d* values; var *e* = variance attributed to sampling error variance; mean δ = mean true effect size; var δ = true variance of effect sizes; % var *SE* = percent variance in observed *d* values due to sampling error variance; V% = percent variance accounted for in observed *d* values due to all corrected artifacts; 90% CrI = 90% of mean δ (credibility interval); 95% CI = 95% of mean δ (confidence interval); fail-safe *N* = number of missing studies averaging null findings that would be needed to bring mean *d* down to .10, from Hunter and Schmidt's (1990) effect size file-drawer analysis.
[a] Five primary studies from the entire data set that used White test takers as the stereotyped group were excluded in these analyses (Aronson et al., 1999; J. L. Smith & White, 2002; von Hippel et al., 2005) so that only minority subgroups' *d* values were meta-analyzed.

based stereotypes. We focus our discussion on the implications of these key moderating relationships.

Moderators

The theory of stereotype threat assumes a uniform pattern of target reactions to the activation (or removal) of a salient negative stereotype when an evaluative ability test is administered. Although our meta-analytic findings suggest that under a situational stereotype threat, both minority and women test takers tended to perform poorly on different types of cognitive ability tests compared with others under no threat, we also found the observed effects appeared to be greater among studies using a race/ethnicity-based stereotype than among studies using a gender-based stereotype. Although these findings should not be directly interpreted (because the zero-included credibility intervals required further moderating tests), they lend initial credence to our proposition that the type of group-based stereotype is an important moderator of stereotype threat effects. In fact, most of our subsequent lower order moderator analytic results further supported this proposition.

In terms of test difficulty, although both racial/ethnic-based and gender-based stereotypes seemed to interact with this moderator in a similar fashion (i.e., more difficult tests produced larger effect sizes), stereotype threat effects were more severe for ethnic minorities than for female test takers when a test was highly difficult. One possible explanation is the methodological inconsistency in how test difficulty is operationalized in the literature. For "very difficult" math tests in gender-based studies, many researchers selected a specific advanced type of standardized quantitative ability test (e.g., GRE Calculus only; Aronson et al., 1999), whereas other content domains were considered as constituting "moderately difficult" or "easy" tests (e.g., algebra, trigonometry, and geometry; O'Brien & Crandall, 2003; S. J. Spencer et al., 1999). The construct of test difficulty might be confounded with math subdomains. It remains unclear in these studies whether stereotype threat effects were manifested at a high level of difficulty or whether they were observed with certain types of math ability problems (e.g., advanced calculus) but not with other types.

In some studies of race-based stereotypes, researchers reviewed test score distributions or pilot-tested test items to see whether or not a test was difficult enough for their sample (e.g., Steele & Aronson, 1995; Stricker & Bejar, 2004; Wicherts, Dolan, & Hessen, 2005). These practices also might induce variance beyond the construct of difficulty. Also, the theory of stereotype threat suggests stereotype threat effects will occur when the test is difficult because of the cognitive demands of taking the test. However, one might posit that difficult items can be seen by a test taker as showing that the stereotype is true and affecting subsequent motivation. Future studies need to adopt a more consistent and appropriate method to operationalize test difficulty and measure potential mediating mechanisms to clarify these differential findings for race and gender stereotypes—a good example is Stricker and Bejar's (2004) approach (reducing the difficulty of the same types of test items). Furthermore, of interest would be studies that examine samples that range in ability (and hence the test is more difficult for some test takers than others) as well as studies that contrast effects on easy versus hard items in the same test.

In terms of test takers' domain identification, the lack of substantial direct investigation of ethnic minorities' domain identification rendered meta-analyses impossible. Similar to the overall pattern of findings in Walton and Cohen (2003), we found that less math-identified women did not suffer much from stereotype threat in terms of math test performance compared with more strongly math-identified women. However, in a departure from Walton and Cohen, we found that moderately math-identified women were surprisingly affected more severely by stereotype threat than highly math-identified ones, which is inconsistent with the theory tenet but may be suggestive of stereotype reactance among highly identified individuals. Investigators might inadvertently lose informative data when implementing the strongest experimental design by screening in only strongly identified individuals (e.g., math

majors). Not only does future research need to expand to include more studies on domain identification effects with ethnic minorities, but researchers also need to include the full spectrum of identification to more accurately determine effects.

Similar to test difficulty, the variance in domain identification effects might be partially explained by the inconsistent operationalization of the construct in the literature. Test takers' domain identification was either directly assessed using self-report measures (e.g., R. P. Brown & Pinel, 2003; Spicer, 1999), indirectly inferred from objective measures such as high standardized cognitive ability test scores (e.g., Anderson, 2001; Quinn & Spencer, 2001; Schmader & Johns, 2003), or assessed with both approaches (e.g., Davies et al., 2002; Harder, 1999). Defining individuals' domain identification indirectly from their existing standardized cognitive ability test scores might be problematic. The performance interference hypothesis of stereotype threat would predict that a negative stereotype might negatively affect target stereotyped individuals in a highly diagnostic testing situation. Prescreening participants based on their strong performance on prior cognitive ability tests in the hope that these high performers would subsequently underperform on another cognitive ability test might result not only in a restriction of range but also in a circular conceptualization of the construct. Future research needs to reach a consensus on the operational definition of domain identification.

The pattern of differential evidence of stereotype threat effects was most apparent when we considered the moderating effect of stereotype threat relevance. As mentioned, the relevance of a negative stereotype might be activated with threat cues or removed with various strategies. We extended Walton and Cohen's (2003) work by categorizing stereotype threat-activating cues as blatant, moderately explicit, or subtle. Stereotype threat theory implies that the more explicit threat cues would produce a stronger stereotype threat effect, and Walton and Cohen's findings supported this prediction. Consistent with this, for minority test takers, we found that subtle stereotype threat cues produced smaller stereotype threat effects compared with other conditions. However, we also found that moderately explicit threat-activating cues produced a greater mean effect size than blatant cues for minority test takers. These interesting findings lend partial credence to the theory of stereotype reactance, which posits that stereotyped individuals might perceive a blatant negative stereotype as a limit to their freedom and ability to perform, thereby ironically invoking behaviors that are inconsistent with the stereotype (see Kray et al., 2001).

In contrast, for female test takers, explicit threat-activation cues (both blatant and moderate) generally produced smaller mean effect sizes than subtle cues. The findings seemed to support Levy's (1996) position that explicit priming of a negative stereotype might produce a weaker effect than subtle priming because the latter might bypass individuals' conscious psychological mechanisms to directly affect task performance.

Before further discussing possible explanations for these differential outcomes, we should note our findings for the other side of the coin: What happens when researchers actively removed the link between a stereotype and an ability test? Explicit stereotype threat-removal strategies were more effective than subtle ones in reducing stereotype threat effects for women, supporting Shih et al.'s (1999) notion that the direct activation of a positive in-group stereotype (e.g., women are better on a specific math test than men) might buffer the effect of the stereotype and even cause a performance boost for some female test takers. However, for ethnic minorities, explicit stereotype threat-removal strategies counterintuitively led to stronger stereotype threat effects compared with subtle strategies, a pattern consistent with the overall moderator effect found by Walton and Cohen (2003). In other words, actively removing stereotype threat seemed to be not as effective for minority test takers.

Because studies with minorities and women rely on different stereotypes and different dependent variables, it may seem unsurprising that different effects were found, even though prior research has treated these as conceptually interchangeable manifestations of stereotype threat. Further, the moderator of stereotype relevance might affect test performance of women and minorities via different mechanisms. For example, telling test takers that ethnic minorities in general perform better than Whites on a certain cognitive ability test (an explicit threat-removal strategy) might actually introduce performance interference to the testing context. In other words, direct and explicit statements might create a performance pressure for these test takers: Should they do poorly, they would not be able to confirm the positive in-group image. Therefore, ethnic minorities' test performance might suffer because of the same psychological mechanisms as experienced by individuals of a "model minority" status (see Cheryan & Bodenhausen, 2000). Explicit interventions aimed at refuting a negative stereotype about minorities' intellectual inferiority might backfire, inadvertently worsening stereotype threat effects instead of alleviating them.

On the other hand, it is possible that female test takers reacted favorably to explicit threat-removals because they might not experience as much performance pressure. As a sociocultural factor, an intellectual-based negative stereotype may carry more distress for ethnic minority targets than a math-based stereotype for women; hence, minorities might paradoxically underperform even when encountering a refuting message, whereas female test takers might take the threat removal message at face value. Although most empirical efforts to understanding why stereotype threat effects generally take place are not successful (see a review by J. L. Smith, 2004), future research needs to focus on investigating possible differential psychological mechanisms underlying group-based targets' reactions to stereotype threat, both activated and removed.

Another possibility is that how stereotype relevance is activated or removed would somehow tap into unique characteristics associated with ethnicity and gender, invoking differential behavioral reactions. *Rejection sensitivity* is defined "as a cognitive–affective processing dynamic . . . whereby people anxiously expect, readily perceive, and intensely react to rejection in situations in which rejection is possible" (Mendoza-Denton, Downey, Purdie, Davis, & Pietrzak, 2002, p. 897). Ethnic minorities have a lifetime history of being subjected to group-based discrimination, mistreatment, prejudice, and exclusion from salient domains (e.g., higher academic education, employment, and certification), either directly or vicariously (see Essed, 1991). When the outcome is important and where one would possibly experience rejection based on one's group membership (Higgins, 1996), minority test takers might more readily recognize and/or interpret situational threat cues as rejection cues than female test takers who might not have the same life experiences (i.e., women successfully pursuing careers not related to math; cf. Halpern et al., 2007). Future research may want

to investigate the hypothesis that ethnic minorities may be more vulnerable to stereotype threat cues than women or other target social groups in cognitive ability testing situations because of their higher level of rejection expectations, as suggested by Mendoza-Denton, Page-Gould, and Pietrzak (2005).

Another related question is whether experimenter effects (R. Rosenthal, 1964) would additionally influence the relationship between stereotype threat relevance and targets' test performance (i.e., contributing to mild, subtle situational cues of stereotype threat). The presence of an out-group experimenter (e.g., men or Whites) might enhance the relevance of stereotype threat cues given to women or ethnic minorities in testing situations, whereas the presence of an in-group experimenter might create more credibility to a message of stereotype threat-removal. Although the predictive effect of experimenter expectancy bias might be small (e.g., less than 3% of the variance in standardized test results; Janssen, 1973), it is possible that African Americans or women might subconsciously experience discomfort with out-group experimenters according to research on in-group/out-group preference (e.g., Pedersen, Walker, & Glass, 1999). So far little evidence has been gathered to address these specific questions within a stereotype threat paradigm, although a few investigators safeguarded against experimenter effects by using experimenter(s) of the same race/ethnicity or gender as that of stereotype threat targets (e.g., Blascovich, Spencer, Quinn, & Steele, 2001, Study 1; Cadinu et al., 2003; Croizet & Claire, 1998; Nguyen, O'Neal, & Ryan, 2003), using double-blind procedures (e.g., Aronson et al., 1999; Quinn & Spencer, 2001; Shih et al., 1999), or minimizing experimenter presence in test administration (e.g., Inzlicht & Ben-Zeev, 2000, 2003). Future research should investigate experimenter effects and/or standardize research procedures.

There are other possible moderators that may further shed light on the effects of stereotype threat effects that we could not meta-analyze in the scope of this study. For example, individuals' defense mechanisms, such as intellectual disidentification, discounting, and disengagement, might mitigate the effects of stereotype threat on cognitive ability test performance (e.g., Major, Feinstein, & Crocker, 1994; Nguyen, Shivpuri, Ryan, & Langset, 2004; Schmader, Major, & Granzow, 2001). Social identity such as racial/ethnic identity (see Sellers, Rowley, Chavous, Shelton, & Smith, 1997) or gender identity (e.g., Schmader, 2002) might make certain target test takers more vulnerable to stereotype threat.

Between-Group Meta-Analytic Findings

Stereotype threat theory is commonly believed to provide a partial explanation of the observed between-group gaps in cognitive ability test performance (e.g., minorities vs. majority; women vs. men; e.g., see Halpern et al., 2007). Assumptions are that (a) the subgroup performance gaps in standardized cognitive ability testing are partially due to stereotype threat effects and (b) removing or refuting stereotype threat reduces or even eliminates these performance gaps (i.e., women performing equally well as men on difficult math tests, minority test takers performing equally well as Whites on various cognitive ability tests). However, Sackett and colleagues (Sackett, 2003; Sackett, Hardison, & Cullen, 2005; Sackett et al., 2001) proposed that any observed reduction in mean test score differences between minority and majority test takers under stereotype threat may be a product of artificial experimental treatments of stereotype threat and/or of how cognitive ability test performance is analyzed (i.e., mean scores controlled for prior cognitive ability).

In our meta-analysis, the minority–majority mean effect sizes were greater in stereotype threat-activated conditions than in control conditions, but the effect size of interest was indeed reduced in the threat-removal condition. However, the pattern of findings for male–female math score differences seemed to support Sackett and colleagues' (Sackett, 2003; Sackett et al., 2001, 2005) position: The performance gaps between men and women in the control and stereotype-removed conditions appeared similar, whereas the mean effect size of interest was higher in the threat-activating condition. With the variability in the effect sizes accounted for by true moderators, we refrain from reaching conclusions. Nevertheless, these meta-analytic results call for future research to better understand the variance in the between-groups mean effect sizes as a function of stereotype type.

In sum, stereotype threat effects are not a ubiquitous phenomenon, because moderators (both those theoretically proposed and unexamined factors) clearly play a role. Despite the inconclusiveness regarding true effect sizes for several analyses, we suggest treating the phenomenon as if it were a real occurrence (i.e., a nonzero effect size), as most mean effects were suggestive of the existence of stereotype threat effects, and a substantial proportion of the data tends to align with the theory.

Practical Implications

Normative practices dictate that a cognitive ability testing setting should be devoid of any overt effort to divert test takers' attention away from the test. Blatant and moderately explicit stereotype-activating cues are not likely to be present in typical employment testing contexts. Subtle cues such as asking individuals their race/gender before taking a test are a relatively common practice in employment settings. Although our findings were equivocal, subtle cues activating stereotypes can and should be avoided in high-stakes testing situations.

One fascinating finding is that explicit stereotype threat-removal strategies might not have intended effects for minority test takers, suggesting that well-intentioned intervention programs designed to proactively deal with stereotype threat effects should carefully monitor the degree to which their "positive" message has effects. Further, we are hard pressed to imagine an employer implementing the types of removal strategies that have been studied (see Table 2) in an actual hiring context. Even the more subtle strategy of describing a test as nondiagnostic could not be implemented, as ethically employers need to convey to test takers what is being tested, why it is being tested, and how test results will be used (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Further research on stereotype-removal strategies that can actually be employed in personnel selection contexts is warranted.

Findings regarding test difficulty indicate that taking highly challenging tests is the most likely predictor of stereotype threat effects for ethnic minority test takers (and for women to a less certain extent). Given the use of difficult screen-in cognitive ability tests to select top candidates (e.g., for prestigious scholarships, for top employment positions), there is a possibility that

ethnic minorities and women might underperform in these settings compared with their true ability if subtle stereotype threat cues exist in the testing environment. Stereotype threat effects are also manifested when cognitive ability tests are only moderately difficult (e.g., consisting of items of mixed difficulty), which suggests applicability to many tests used in educational or employment testing settings.

Because typical employment and academic test-taking contexts do not involve explicit or even moderately explicit cues but likely involve subtle ones (e.g., asking test takers to report demographics prior to taking a test), understanding the true magnitude of the effect of subtle cues is important. Even when a mean stereotype threat effect size d is as small as |.18|, it still indicates that almost a one-fifth standard deviation separates two group means (e.g., stereotyped-activated group means vs. stereotype-removed group means). If a minority student took the SAT and his or her true cognitive ability were at the national mean level, he or she might underperform by about 50 points due to subtle stereotype threat cues. Danaher and Crandall (2008) reanalyzed Stricker and Ward's (2004) data and argued that even with a stereotype threat effect size as small as $d < .20$, in the context of high-stakes educational testing, this effect size might still translate to an additional 5.9% of women achieving a passing score for calculus. Combining this fact with the meta-analytic results concerning situational cues and threat removal strategies, at-risk test takers (stereotyped group members) should be aware of the possibility of underperformance in high-stakes testing situations.

On the basis of the meta-analytic findings in the present review, researchers, practitioners, educators, social policy-makers, and test takers themselves should be cautious in generalizing and applying stereotype threat-based knowledge acquired from one target social group (e.g., women performing on math tests) to generate research ideas or make decisions regarding another target group in a different testing setting (e.g., ethnic minorities taking cognitive ability tests). Also, test types may be confounded with social groups. For example, minorities' test performance was investigated either in a single or multiple ability domains (e.g., Steele & Aronson, 1995, used a GRE verbal test; J. L. Smith & White, 2002, used a GRE math test; other researchers employed a mixed-ability domain test, such as McFarland, Lev-Arey, & Ziegert, 2003, and Stricker & Ward, 2004). Could specific domains of cognitive ability tests moderate stereotype threat effects for minorities? Could different types of math tests yield different stereotype threat effects for women? Unfortunately, the issue of nonindependent data points does not allow these research questions to be examined meta-analytically at the present time.

Limitations

The present meta-analytic review has several limitations. First, the inclusion criteria might be defined and applied too strictly, as only about half of empirical articles in the preliminary database were retained in the final meta-analytic set. However, we felt each of these criteria was critical for the scope of our study. Although a vigilant process of literature search was undertaken, publication biases still existed in a few subsets of the data. These cases were clearly described so that readers can exercise caution in drawing conclusions about the findings.

The nonzero variance in some findings indicates that further moderator analyses might be necessary. Although in this meta-analysis most subsets of effect sizes used to investigate stereotype threat moderators with minority test takers were sufficiently large (i.e., having no fewer than five effect sizes per cell; see Arthur, Bennett, Edens, & Bell, 2003), further dividing the data in some of these subsets to search for additional moderators to fully explain data variance would result in trivial findings (i.e., due to greater sampling error). For instance, domain identification might be conceptually nested in test difficulty, which might be in turn nested in stereotype threat-activating cues. Unfortunately, testing such a fully nested, hierarchical model of moderator analyses is beyond the scope of this data set. Few empirical studies test all of these moderators; therefore, we call for researchers to incorporate all of the boundary moderators in their investigations.

One important limitation of the literature base is the lack of sufficient, successful studies on mediating mechanisms. That is, we do not know exactly what psychological processes stereotype-activating cues trigger (e.g., a drop in efficacy or motivation, an increase in anxiety or off-task thinking), and therefore there are limits to understanding and addressing stereotype threat effects. We feel that the acceptance of stereotype threat effects as more than a laboratory phenomenon will require further specification and investigation of the psychological processes that underlie the effects.

Last, the results from studies employing other group-based stereotypes (e.g., associated with ageism, social class, study majors) and/or measuring other domains of ability (e.g., athletic ability, work-related abilities, working memory capacity) were not cumulated in the present review, as we focused on the two stereotypes of concern to many personnel testing contexts. Given the differential patterns of findings between stereotypes relevant to women and ethnic minorities in the present study, we speculate that members of other stereotyped groups might also react differently to stereotype threat activation. This research question remains to be explored in future studies.

Conclusion

Stereotype threat theory is a high-impact framework, and the literature has received much attention from social scientists and the public in the past decade. We meta-analyzed the reported effects of stereotype threat activation on one specific outcome: cognitive ability test performance or math test performance of ethnic minority test takers or women, taking into account most relevant conceptual and methodological moderators. The meta-analytic findings not only revealed the complexity of the stereotype threat phenomenon (i.e., that it manifests differentially under various conditions) but also suggested a research direction for future studies where all of the important boundary constructs should be incorporated in research designs to more accurately represent the true effects of stereotype threat in employment testing contexts.

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). Detecting negative self-relevant stereotype activation: The ef-

fects of individuation. *Journal of Experimental Social Psychology, 40,* 401–408.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

*Anderson, R. D. (2001). *Stereotype threat: The effects of gender identification on standardized test performance.* Unpublished doctoral dissertation, James Madison University.

*Aronson, J., Lusting, M. J., Good, C., Keough, K., Steele, C. M., & Brown, J. (1999). When White men can't do math: Necessary and sufficient factors in stereotype threat. *Journal of Experimental Social Psychology, 35,* 29–46.

Arthur, W., Bennett, W., Edens, P. S., & Bell, S. T. (2003). The effectiveness of training in organizations: A meta-analysis of design and evaluation features. *Journal of Applied Psychology, 88,* 234–245.

*Bailey, A. A. (2004). *Effects of stereotype threat on females in math and science fields: An investigation of possible mediators and moderators of the threat–performance relationship.* Unpublished doctoral dissertation, Georgia Institute of Technology.

Bargh, J. A. (1997). The automaticity of everyday life. In R. S. Wyer Jr. (Ed.), *Advances in social cognition* (Vol. 10, pp. 1–61). Mahwah, NJ: Erlbaum.

Ben-Zeev, T., Fein, S., & Inzlicht, M. (2005). Arousal and stereotype threat. *Journal of Experimental Social Psychology, 41,* 174–181.

Blascovich, J., Spencer, S. J., Quinn, D., & Steele, C. (2001). African Americans and high blood pressure: The role of stereotype threat. *Psychological Science, 12,* 225–229.

*Brown, J. L., Steele, C. M., & Atkins, D. (n.d.). *Performance expectations are not a necessary mediator of stereotype threat in African American verbal test performance.* Unpublished manuscript.

*Brown, R. P., & Day, E. A. (2006). The difference isn't Black and White: Stereotype threat and the race gap on Raven's Advanced Progressive Matrices. *Journal of Applied Psychology, 91,* 979–985.

*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,* 246–257.

*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.

*Cadinu, M., Maass, A., Frigerio, S., Impagliazzo, L., & Latinotti, S. (2003). Stereotype threat: The effect of expectancy on performance. *European Journal of Social Psychology, 33,* 267–285.

*Cadinu, M., Maass, 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.

Cheryan, S., & Bodenhausen, G. (2000). When positive stereotypes threaten intellectual performance: The psychological hazards of "model minority" status. *Psychological Science, 11,* 399–402.

Chung-Herrera, B. G., Ehrhart, M. B., Ehrhart, K. H., Hattrup, K., & Solamon, J. (2005). A new vision of stereotype threat: Testing its effects in a field setting. In K. M. Weaver (Ed.), *Proceedings of the 65th annual meeting of the Academy of Management*.

*Cohen, G. L., & Garcia, J. (2005). "I am us": Negative stereotypes as collective threats. *Journal of Personality and Social Psychology, 89,* 566–582.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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

Crocker, J., & Major, B. (1989). Social stigma and self-esteem: The self-protective properties of stigma. *Psychological Review, 96,* 608–630.

Croizet, J.-C., & Claire, T. (1998). Extending the concept of stereotype threat to social class: The intellectual underperformance of students from low socioeconomic backgrounds. *Personality and Social Psychology Bulletin, 24,* 588–594.

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

Cullen, M. J., Hardison, C. M., & Sackett, P. R. (2004). Using SAT–grade and ability–job performance relationships to test predictions derived from stereotype threat theory. *Journal of Applied Psychology, 89,* 220–230.

Danaher, K., & Crandall, C. S. (2008). Stereotype threat in applied settings reexamined. *Journal of Applied Social Psychology, 38,* 1639–1655.

*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.

Devos, T., & Banaji, M. R. (2003). Implicit self and identity. In M. R. Leary & J. P. Tangney (Eds.), *Handbook of self and identity* (pp. 153–175). New York: Guilford Press.

*Dinella, L. M. (2004). *A developmental perspective on stereotype threat and high school mathematics.* Unpublished doctoral dissertation, Arizona State University.

*Dodge, T. L., Williams, K. J., & Blanton, H. (2001, April). *Motivational mediators of the stereotype threat effect.* Paper presented at the 16th annual meeting of the Society for Industrial and Organizational Psychology, San Diego, CA.

*Edwards, B. D. (2004). *An examination of factors contributing to a reduction in race-based subgroup differences on a constructed response paper-and-pencil test of achievement.* Unpublished doctoral dissertation, Texas A&M University.

*Elizaga, R. A., & Markman, K. D. (n.d.). *Peers and performance: How in-group and out-group comparisons moderate stereotype threat effects.* Unpublished manuscript.

Essed, P. J. M. (1991). *Understanding everyday racism.* Newbury Park, CA: Sage.

Fisher, R. A. (1925). *Statistical methods for research workers.* Edinburgh, United Kingdom: Oliver & Boyd.

*Foels, R. (1998, June). *Women's math ability: An investigation of stereotype threat.* Poster presented at the Society for the Psychological Study of Social Issues conference, Ann Arbor, MI.

*Foels, R. (2000, February). *Disidentification in the face of stereotype threat.* Paper presented at the Society for Personality and Social Psychology conference, Nashville, TN.

*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.

*Gamet, M. M. (2004). *Stereotype threat and the effects on women in mathematical tasks.* Unpublished manuscript.

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

Good, C., Aronson, J., & Inzlicht, M. (2003). Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Journal of Applied Developmental Psychology, 24,* 645–662.

*Gresky, D. M., Eyck, L. L. T., & Lord, C. G. (n.d.). *Effects of salient multiple identities on women's performance under mathematics stereotype threat.* Unpublished manuscript.

*Guajardo, G. A. (2005). *Modifying stereotype relevance and altering affect attributions to reduce performance suppression on cognitive abil-*

ity selection tests. Unpublished master's thesis, Northern Illinois University.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest, 8,* 1–51.

*Harder, J. A. (1999). *The effect of private versus public evaluation on stereotype threat for women in mathematics.* Unpublished doctoral dissertation, University of Texas at Austin.

Higgins, E. T. (1996). Knowledge activation: Accessibility, applicability and salience. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 136–168). New York: Guilford Press.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Newbury Park, CA: Sage.

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.

Inzlicht, M., & Ben-Zeev, T. (2003). Do high-achieving female students underperform in private? The implication of threatening environments on intellectual processing. *Journal of Educational Psychology, 95,* 796–805.

Janssen, J. P. (1973). The experimenter's expectation effect: An artifact of non-standardized experimental conditions? An investigation concerning Rosenthal's experimenter bias under standardized group experimental conditions. *Psychologische Beitrage, 15,* 230–248.

*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.

*Josephs, R. A., Newman, M. L., Brown, R. P., & Beer, J. M. (2003). Status, testosterone, and human intellectual performance: Stereotype threat as status concern. *Psychological Science, 14,* 158–163.

*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.

*Keller, J. (2007). Stereotype threat in classroom settings: The interactive effect of domain identification, task difficulty, and stereotype threat on female students' math performance. *British Journal of Educational Psychology, 77,* 323–338.

*Keller, J., & Bless, H. (n.d.). *When positive and negative expectancies disrupt performance: Regulatory focus as a catalyst.* Unpublished manuscript.

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

Koslowsky, M., & Sagie, A. (1993). On the efficacy of credibility intervals as indicators of moderator effects in meta-analytic research. *Journal of Organizational Behavior, 14,* 695–699.

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

Landis, J., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

Levy, B. (1996). Improving memory in old age by implicit self-stereotyping. *Journal of Personality and Social Psychology, 71,* 1092–1107.

*Lewis, P. B. (1998). *Stereotype threat, implicit theories of intelligence, and racial differences in standardized test performance.* Unpublished doctoral dissertation, Kent State University.

Leyens, J.-P., Desert, M., Croizet J.-C., & Darcis, C. (2000). Stereotype threat: Are lower status and history of stigmatization preconditions of stereotype threat? *Personality and Social Psychology Bulletin, 26,* 1189–1199.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research.* Cambridge, MA: Harvard University Press.

Major, B., Feinstein, J., & Crocker, J. (1994). Attributional ambiguity of affirmative action. *Basic and Applied Social Psychology, 15,* 113–141.

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

*Martin, D. E. (2004). *Stereotype threat, cognitive aptitude measures, and social identity.* Unpublished doctoral dissertation, Howard University.

*Marx, D. M., & Stapel, D. A. (2005). It's all in the timing: Measuring emotional reactions to stereotype threat before and after taking a test. *European Journal of Social Psychology, 35,* 1–12.

*Marx, D. M., & Stapel, D. A. (2006). Distinguishing stereotype threat from priming effects: On role of the social self and threat-based concerns. *Journal of Personality and Social Psychology, 91,* 243–254.

*Marx, D. M., Stapel, D. A., & Muller, D. (2005). We can do it: The interplay of construal orientation and social comparisons under threat. *Journal of Personality and Social Psychology, 88,* 432–446.

*McFarland, L. A., Kemp, C. F., Viera, L., Jr., & Odin, E. P. (2003, April). *Stereotype threat and male–female differences in test performance.* Paper presented at the 18th annual conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

*McFarland, L. A., Lev-Arey, D. M., & Ziegert, J. C. (2003). An examination of stereotype threat in a motivational context. *Human Performance, 16,* 181–205.

*McIntyre, R. B., Lord, C. G., Gresky, D. M., Eyck, L. L. T., Frye, G. D. J., & Bond, C. F., Jr. (2005). A social impact trend in the effects of role models on alleviating women's mathematics stereotype threat. *Current Research in Psychology, 10,* 116–136.

*McIntyre, R. B., Paulson, R. M., & Lord, C. G. (2003). Alleviating women's mathematics stereotype threat through salience of group achievements. *Journal of Experimental Social Psychology, 39,* 83–90.

*McKay, P. F. (1999). *Stereotype threat and its effect on the cognitive ability test performance of African-Americans: The development of a theoretical model.* Unpublished doctoral dissertation, University of Akron.

McKay, P. F., Doverspike, D., Bowen-Hilton, D., & Martin, Q. D. (2002). Stereotype threat effects on the Raven Advanced Progressive Matrices scores of African Americans. *Journal of Applied Social Psychology, 32,* 767–787.

Mendoza-Denton, R., Downey, G., Purdie, V., Davis, A., & Pietrzak, J. (2002). Sensitivity to status-based rejection: Implications for African American students' college experience. *Journal of Personality and Social Psychology, 83,* 896–918.

Mendoza-Denton, R., Page-Gould, E., & Pietrzak, J. (2005). Mechanisms for coping with race-based rejection expectations. In S. Levin & C. van Laar (Eds.), *Stigma and group inequality: Social psychological approaches.* New York: Erlbaum.

*Nguyen, H.-H. D., Shivpuri, S., Ryan, A. M., & Langset, K. (2004, April). *Relations of stereotype threat effects to assessment domains and self-identity.* Paper presented at the 19th annual meeting of the Society for Industrial and Organizational Psychology, Chicago, IL.

*Nguyen, H.-H. D., O'Neal, A., & Ryan, A. M. (2003). Relating test-taking attitudes and skills and stereotype threat effects to the racial gap in cognitive ability test performance. *Human Performance, 16,* 261–294.

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

Osborne, J. W. (2001a). Testing stereotype threat: Does anxiety explain race and sex differences in achievement? *Contemporary Educational Psychology, 26,* 291–310.

Osborne, J. W. (2001b). Unraveling underachievement among African

American boys from an identification with academic perspective. *The Journal of Negro Education, 68,* 555–565.

*Oswald, D. L., & Harvey, R. D. (2000–2001). Hostile environments, stereotype threat, and math performance among undergraduate women. *Current Psychology: Developmental, Learning, Personality, Social, 19,* 338–356.

Pedersen, A., Walker, I., & Glass, C. (1999). Experimenter effects on the ingroup preference and self-concept of urban Aboriginal children. *Australian Journal of Psychology, 51,* 82–89.

*Pellegrini, A. V. (2005). *The impact of stereotype threat on intelligence testing in Hispanic females.* Unpublished doctoral dissertation, Carlos Albizu University.

*Philipp, M. C., & Harton, H. C. (2004, January). *The role of social dominance in stereotype threat effects.* Paper presented at the annual meeting of the Society for Personality and Social Psychology, Austin, TX.

*Ployhart, R. E., Ziegert, J. C., & McFarland, L. A. (2003). Understanding racial differences on cognitive ability tests in selection contexts: An integration of stereotype threat and applicant reactions research. *Human Performance, 16,* 231–259.

*Prather, H. M. (2005). *Controlling the threat of stereotypes: The effectiveness of mental control strategies in increasing female math ability test performance.* Unpublished doctoral dissertation, George Washington University.

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.

*Rivadeneyra, R. (2002). *The influence of television on stereotype threat among adolescents of Mexican descent.* Unpublished doctoral dissertation, University of Michigan.

Roberson, L., Deitch, E. A., Brief, A. P., & Block, C. J. (2003). Stereotype threat and feedback seeking in the workplace. *Journal of Vocational Behavior, 62,* 176–188.

*Rosenthal, H. E. S., & Crisp, R. J. (2006). Reducing stereotype threat by blurring intergroup boundaries. *Journal of Personality and Social Psychology, 32,* 501–511.

Rosenthal, R. (1964). The effect of the experimenter on the results of psychological research. *Bulletin of the Maritime Psychological Association, 13,* 1–39.

Rosnow, R. L., & Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychological Methods, 1,* 331–340.

Rosnow, R. L., Rosenthal, R., & Rubin, D. B. (2000). Contrasts and correlations in effect-size estimation. *Psychological Science, 11,* 446–453.

Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., III, & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54,* 297–330.

Roth, P. L., Bobko, P., & McFarland, L. A. (2005). A meta-analysis of work sample test validity: Updating and integrating some classic literature. *Personnel Psychology, 58,* 1009–1037.

Sackett, P. R. (2003). Stereotype threat in applied selection settings: A commentary. *Human Performance, 16,* 295–309.

Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2005). On interpreting research on stereotype threat and test performance. *American Psychologist, 60,* 271–272.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist, 56,* 302–318.

*Salinas, M. F. (1998). *Stereotype threat: The role of effort withdrawal and apprehension on the intellectual underperformance of Mexican-Americans.* Unpublished doctoral dissertation, University of Texas at Austin.

*Sawyer, T. P., Jr., & Hollis-Sawyer, L. A. (2005). Predicting stereotype threat, test anxiety, and cognitive ability test performance: An examination of three models. *International Journal of Testing, 5,* 225–246.

*Schimel, J., Arndt, J., Banko, K. M., & Cook, A. (2004). Not all self-affirmations were created equal: The cognitive and social benefits of affirming the intrinsic (vs. extrinsic) self. *Social Cognition, 22,* 75–99.

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

*Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology, 85,* 440–452.

*Schmader, T., Johns, M., & Barquissau, M. (2004). The costs of accepting gender differences: The role of stereotype endorsement in women's experience in the math domain. *Sex Roles, 50,* 835–850.

Schmader, T., Major, B., & Granzow, R. H. (2001). How African-American college students protect their self-esteem. *Journal of Social Issues, 57,* 116–119.

Schmidt, F. L., & Le, H. A. (2005). The Hunter–Schmidt meta-analysis programs package (Version 1.1) [Computer software]. Available from http://www.testpublishers.org/Documents/FrankSchmidtSoftware.pdf

*Schneeberger, N. A., & Williams, K. (2003, April). *Why women "can't" do math: The role of cognitive load in stereotype threat research.* Paper presented at the 18th meeting of the Society for Industrial and Organizational Psychology, Orlando, FL.

*Schultz, P. W., Baker, N., Herrera, E., & Khazian, A. (n.d.). *Stereotype threat among Hispanic-Americans and the moderating role of ethnic identity.* Unpublished manuscript.

*Seagal, J. D. (2001). *Identity among members of stigmatized groups: A double-edged sword.* Unpublished doctoral dissertation, University of Texas at Austin.

*Sekaquaptewa, D., & Thompson, M. (2002). Solo status, stereotype threat, and performance expectancies: Their effects on women's performance. *Journal of Experimental Social Psychology, 39,* 68–74.

Sellers, R. M., Rowley, S. A., Chavous, T. M., Shelton, J. N., & Smith, M. A. (1997). Multidimensional Inventory of Black Identity: A preliminary investigation of reliability and construct validity. *Journal of Personality and Social Psychology, 73,* 805–815.

Shih, M., Pittinsky, T. L., & Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological Science, 10,* 80–83.

*Smith, C. E., & Hopkins, R. (2004). Mitigating the impact of stereotypes on academic performance: The effects of cultural identity and attributions for success among African American college students. *Western Journal of Black Studies, 28,* 312–321.

Smith, J. L. (2004). Understanding the process of stereotype threat: A review of mediational variables and new performance goal directions. *Educational Psychology Review, 16,* 177–206.

*Smith, J. L., & White, P. H. (2002). An examination of implicitly activated, explicitly activated, and nullified stereotypes on mathematical performance: It's not just a woman's issue. *Sex Roles, 47,* 179–191.

Spencer, S. J., Fein, S., Strahan, E. J., & Zanna, M. P. (2005). The role of motivation in the unconscious: How our motives control the activation of our thoughts and shape our actions. In K. D. Williams & J. P. Forgas (Eds.), *Social motivation: Conscious and unconscious processes* (pp. 113–129). New York: Cambridge University Press.

*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.

*Spencer, S. L. (2005). *Stereotype threat and women's math performance: The possible mediating factors of test anxiety, test motivation and self-efficacy.* Unpublished doctoral dissertation, Rutgers, The State University of New Jersey.

*Spicer, C. V. (1999). *Effects of self-stereotyping and stereotype threat on*

intellectual performance. Unpublished doctoral dissertation, University of Kentucky.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52,* 613–629.

*Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69,* 797–811.

Steele, C. M., & Aronson, J. (1998). Stereotype threat and the test performance of academically successful African Americans. In C. Jencks & M. Phillips (Eds.), *The Black–White test score gap* (pp. 401–427). Washington, DC: Brookings.

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (pp. 379–440). San Diego, CA: Academic Press.

*Sternberg, R. J., Jarvin, L., Leighton, J., Newman, T., Moon, T., Callahan, C., & Grigorenko, E. L. (n.d.). *Girls can't do math?: The disidentification effect and gifted high school students' math performance.* Unpublished manuscript.

Stricker, L. J. (1998). *Inquiring about examinee's ethnicity and sex: Effects on AP Calculus AB examination performance.* (College Board Rep. No. 98–1; ETS Research Rep. No. 98–5). New York: College Entrance Examination Board.

Stricker, L. J., & Bejar, I. I. (2004). Test difficulty and stereotype threat on the GRE General test. *Journal of Applied Social Psychology, 34,* 563–597.

Stricker, L. J., & Ward, W. C. (1998). *Inquiring about examinee's ethnicity and sex: Effects on computerized placement tests performance*. (College Board Rep. No. 98–2; ETS Research Rep. No. 98–9). New York: College Entrance Examination Board.

*Stricker, L. J., & Ward, W. C. (2004). Stereotype threat, inquiring about test takers' ethnicity and sex, and standardized test performance. *Journal of Applied Social Psychology, 34,* 665–693.

*Tagler, M. J. (2003). *Stereotype threat: Prevalence and individual differences.* Unpublished doctoral dissertation, Kansas State University.

Thalheimer, W., & Cook, S. (2002, August). *How to calculate effect sizes from published research articles: A simplified methodology.* Retrieved May 20, 2008, from http://work-learning.com/effect_sizes.htm

*van Dijk, A., Koenders, H., Korenhof, I. H., Mulder, H. R., & de Vries, H. (n.d.). *The moderating role of group membership activation on stereotype lift and threat.* Unpublished manuscript.

*von Hippel, W., von Hippel, C., Conway, L., Preacher, K. J., Schooler, J. W., & Radvansky, G. A. (2005). Coping with stereotype threat: Denial as an impression management strategy. *Journal of Personality and Social Psychology, 89,* 22–35.

*Walsh, M., Hickey, C., & Duffy, J. (1999). Influence of item content and stereotype situation on gender differences in mathematical problem solving. *Sex Roles, 41,* 219–240.

*Walters, A. M. (2000). *Stereotype threat: An examination of process.* Unpublished doctoral dissertation, University of Florida.

Walton, G. M., & Cohen, G. L. (2003). Stereotype lift. *Journal of Experimental Social Psychology, 39,* 456–467.

Webb, T. L., & Sheeran, P. (2006). Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence. *Psychological Bulletin, 132,* 249–268.

Whitener, E. M. (1990). Confusion of confidence intervals and credibility intervals in meta-analysis. *Journal of Applied Psychology, 75,* 315–321.

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

*Wout, D. A., Shih, M. J., Jackson, J. S., & Sellers, R. M. (n.d.). *Targets as perceivers: How Blacks determine if they will be stereotyped.* Unpublished manuscript.

Zhao, H., & Seibert, S. E. (2006). The Big Five personality dimensions and entrepreneurial status: A meta-analytical review. *Journal of Applied Psychology, 91,* 259–271.

Received May 23, 2007
Revision received February 12, 2008
Accepted April 28, 2008 ■