American Psychological Association

Journal of Applied Psychology

© 2019 American Psychological Association
0021-9010/19/$12.00

http://dx.doi.org/10.1037/apl0000420

Stereotype Threat Effects in Settings With Features Likely Versus Unlikely in Operational Test Settings: A Meta-Analysis

Oren R. Shewach, Paul R. Sackett, and Sander Quint
University of Minnesota – Twin Cities

The stereotype threat literature primarily comprises lab studies, many of which involve features that would not be present in high-stakes testing settings. We meta-analyze the effect of stereotype threat on cognitive ability tests, focusing on both laboratory and operational studies with features likely to be present in high stakes settings. First, we examine the features of cognitive ability test metric, stereotype threat cue activation strength, and type of nonthreat control group, and conduct a focal analysis removing conditions that would not be present in high stakes settings. We also take into account a previously unrecognized methodological error in how data are analyzed in studies that control for scores on a prior cognitive ability test, which resulted in a biased estimate of stereotype threat. The focal sample, restricting the database to samples utilizing operational testing-relevant conditions, displayed a threat effect of $d = -.14$ ($k = 45$, $N = 3{,}532$, $SD_{\delta} = .31$). Second, we present a comprehensive meta-analysis of stereotype threat. Third, we examine a small subset of studies in operational test settings and studies utilizing motivational incentives, which yielded d-values ranging from .00 to $-.14$. Fourth, the meta-analytic database is subjected to tests of publication bias, finding nontrivial evidence for publication bias. Overall, results indicate that the size of the stereotype threat effect that can be experienced on tests of cognitive ability in operational scenarios such as college admissions tests and employment testing may range from negligible to small.

Keywords: meta-analysis, personnel selection, standardized testing, stereotype threat, subgroup differences

Supplemental materials: http://dx.doi.org/10.1037/apl0000420.supp

Stereotype threat has been extensively examined since Steele and Aronson's (1995) seminal experiments. Stereotype threat is defined as "the situation in which there is a negative stereotype about a person's group, and he or she is concerned about being judged or treated negatively on the basis of this stereotype" (Spencer, Logel, & Davies, 2016, p. 416). There are now hundreds of primary studies measuring the effects of this phenomenon on outcomes such as cognitive ability, working memory, athletic performance, academic performance, negotiation, and decision-making. Stereotype threat is most frequently examined in the context of cognitive ability testing (e.g., quantitative or verbal ability) within female and ethnic minority samples.

A central hypothesis of this literature states that experiencing stereotype threat interferes with test performance, resulting in an underestimation of ability for the stereotyped group (Spencer, Steele, & Quinn, 1999; Steele & Aronson, 1995). This line of evidence was used in an amicus curiae brief in the United States Supreme Court case *Fisher v. University of Texas* in support of the use of race in college admissions. The central argument is: "a substantial body of research by social scientists has revealed that standardized test scores and grades often underestimate the true academic capacity of members of certain minority groups" (Aronson et al., 2015, p. 4).

There have been multiple meta-analyses of the stereotype threat phenomenon on cognitive ability tests. Table 1 documents these. Some are early efforts, with a relatively small number of studies (Walton & Cohen, 2003); others focus on a narrow domain (e.g., women in math; Picho, Rodriguez, & Finnie, 2013). The current study can be seen as a complement to two prior meta-analyses. Whereas Flore and Wicherts (2015) examine threat effects in primary and secondary school, we examine effects in adults. The study can also be seen as a complement to Nguyen and Ryan (2008), which is the largest meta-analysis to date, with 116 samples. They included both high-school and adult samples; again, we focus solely on adult samples. With 212 samples, our study is the largest meta-analysis on the topic by a considerable margin.

Oren R. Shewach, Paul R. Sackett, and Sander Quint, Department of Psychology, University of Minnesota – Twin Cities.

Portions of this article were presented at the 2016 and 2018 Annual Meetings of the Society for Industrial and Organizational Psychology in Anaheim, California, and Chicago, Illinois. We thank Brenton Wiernik and Jeff Dahlke for their helpful input on earlier versions of this article.

Correspondence concerning this article should be addressed to Oren R. Shewach, who is now at Human Resources Research Organization (HumRRO), 100 Washington Avenue South, Suite 1660, Minneapolis, MN 55401. E-mail: oshewach@humrro.org

Table 1
Stereotype Threat Meta-Analyses' Evaluation of Publication Bias

Meta-analysis	Nature of sample, criterion, and research question	Inclusion criteria	*k*	Substantive findings and conclusions
Present study	Cognitive ability tests in adults, focusing on features relevant to operational testing scenarios	Experiments evaluating within-group performance, with a mean sample age 18 or older containing a stereotype activation and either control or removal group, and assessing performance on a stereotype-relevant cognitive ability test	212	The focal sample with operational test-relevant conditions displayed a threat effect of $d = -.14$. Small subsets of studies in operational test settings and utilizing motivational incentives ranged from $d = .00$ to $-.14$. Tests indicated non-trivial evidence for publication bias in the database (e.g., trim-and-fill analyses reduced the focal analysis $d = -.14$ estimate to $-.09$).
Zigerell (2017)	Subjected Nguyen and Ryan (2008) sample to publication bias tests	Identical to Nguyen and Ryan (2008)	115	"Four methods to adjust the meta-analysis effect size for potential publication bias produced divergent estimates, from essentially no change, to a 50% decrease, to a reduction of the estimated effect size to near zero" (p. 1159).
Flore and Wicherts (2015)	Child and adolescent girls on tests of math, science, and spatial skills	Experiments using girls with mean sample age under 18, manipulating a gender stereotype as a between-subjects factor, with a dependent variable of math, science, or spatial skills	47	"The estimated mean effect size [*d*] equaled $-.22$ and significantly differed from 0. None of the moderator variables was significant; however, there were several signs for the presence of publication bias. We conclude that publication bias might seriously distort the literature on the effects of stereotype threat among schoolgirls" (p. 25).
Picho et al. (2013)	Moderating role of context for females on tests of quantitative ability	Experiments or quasi-experiments with control and experimental groups, evaluating female performance on a quantitative test	103	"Findings revealed that, on average, females in ST conditions performed less well on mathematics tests than their control counterparts ($d = \|.24\|$)" (p. 299).
Stoet and Geary (2012)	Direct replications of the original stereotype threat study (Spencer, Steele, & Quinn, 1999)	Experiments that were direct replications of Spencer et al.'s study, testing between-group threat effects (male vs. female) evaluating math performance using two different threat conditions	19	"Only 55% of the articles with experimental designs that could have replicated the original results did so. But half of these were confounded by statistical adjustment of preexisting mathematics exam scores. Of the unconfounded experiments, only 30% replicated the original" (p. 93).
Walton and Spencer (2009)	Test performance controlling for prior ability	Between-group studies including participants negatively stereotyped and non-stereotyped, manipulating stereotype threat, and assessing performance in a subsequent stereotype-relevant domain while also assessing performance in the same domain in a real-world context outside the testing session	39	"At the mean level of prior performance, stereotyped students performed better than non-stereotyped students, $Z = 3.15$, $p = .002$, $d = .18$" (p. 1135). "Both meta-analyses found that, under conditions that reduce psychological threat, stereotyped students performed better than non-stereotyped studies at the same level of past performance" (p. 1132).
Nguyen and Ryan (2008)	Women and minorities on tests of cognitive ability	Experiments designed to test Steele and Aronson's (1995) within-subgroup performance interference hypothesis regarding stereotyped minorities' or women's cognitive ability test performance	116	"A meta-analysis of stereotype threat effects was conducted and an overall mean effect size of |.26| was found, but true moderator effects existed" (p. 1314).
Walton and Cohen (2003)	*Stereotype lift,* occurring when downward comparisons are made with a denigrated group	Experiments assessing test performance of a non-stereotyped group (e.g., men and Whites), using a stereotype-relevant and stereotype-irrelevant condition, and using difficult tests that were linked to the negative stereotype more in the stereotype-relevant than in the stereotype-irrelevant condition	43	"In a meta-analytic review, members of nonstereotyped groups were found to perform better when a negative stereotype about an outgroup was linked to an intellectual test than when it was not ($d = .24$, $p < .0001$)" (p. 456).

The Current Study: Focus on Estimating Effects in High Stakes Settings

Stereotype threat frequently gets cited as partially or fully explaining group differences in standardized tests. For example, "it is now well established that the threat of confirming a negative stereotype about women's math ability harms their performance on standardized math tests" (Smeding, Dumas, Loose, & Régner, 2013). Sackett, Hardison, and Cullen (2004) document many cases in which stereotype threat evidence is incorrectly interpreted as fully explaining Black–White differences in standardized tests. Note that 208 of 212 studies in this analysis were conducted in lab settings and have not addressed generalizability to operational testing settings. We define operational testing settings (also known as high stakes testing) as tests that have significant real-world consequences for test-takers, primarily in college admissions and personnel selection tests. Although the lab studies cannot per se answer the question of the existence and magnitude of threat effects in operational testing settings, the position we develop here is that lab studies differ in the degree to which they include features likely to be present in operational settings, and that insight can be gained by comparing studies with and without features found in operational settings. This leads to the central research question in this study:

> *Central research question:* What is the magnitude of the stereotype threat effect on stereotyped groups' (i.e., females, ethnic minorities) cognitive ability test performance, in experimental conditions with features likely to be encountered in operational testing settings?

To address this overarching research question, this meta-analysis of stereotype threat pursues five major lines of investigation: (a) examination of four methodological and analytic moderators pertinent to operational testing settings, (b) examination of these four moderators in conjunction, producing a focal sample of studies that displays a greater degree of similarity to operational test-like conditions, (c) examination of a small subset of samples actually conducted in operational contexts, (d) examination of samples that use motivational incentives (i.e., financial or other incentives to top performers on the ability test) to better approximate motivation levels in operational testing settings, and (e) examination of the degree to which stereotype threat is inflated by publication bias, to evaluate the sensitivity of conclusions on the existence of stereotype threat to the presence of unpublished studies and selective reporting in some studies. Each of these five lines of investigation addresses different aspects of generalizability of stereotype threat to high stakes testing settings.

The first line of investigation identifies four moderators relevant to operational testing settings. Fundamentally, we remove conditions that we identify as not plausible in operational testing. The first is use of a covariate to adjust scores on the outcome variable. Some studies adjust cognitive ability outcome scores using estimates of prior ability (e.g., ACT/SAT scores), which we will illustrate below to be in error when making within-group comparisons. The second is metric of cognitive ability score. Some stereotype threat studies score ability tests as accuracy scores (total correct/number attempted), which would not be used in operational settings, as high scores can be obtained by simply skipping questions for which one is not confident of the answer. The third is type of threat comparison group. Some studies use a comparison between a stereotype threat activation group and a stereotype threat removal group; we argue that the standard experimental-control comparison (i.e., stereotype activation vs. no activation) provides clearer interpretation of stereotype threat. Finally, the fourth is stereotype activation strength. Some studies invoke stereotype threat quite explicitly (e.g., the test you are about to take is one which favors men over women"), whereas test-takers in operational settings are much more likely to experience subtle invocation of threat such as priming of gender by asking test-takers to report gender before the test. Prior meta-analyses (i.e., Flore & Wicherts, 2015; Nguyen & Ryan, 2008) have examined comparison group and activation strength as individual moderators, whereas outcome metric and use of a covariate have not been examined. The second line of investigation involves analyzing all four of the above features in conjunction, creating a focal subset of studies that display conditions most similar to operational testing settings.

The third and fourth lines of investigation are of operational samples and studies with motivational incentives, addressing the question of the degree of stereotype threat in actual high stakes settings and settings with increased motivation, respectively. The operational sample consists of large sample studies of college placement exams. However, with a *k* of only four, we turn to other sources of evidence regarding increased-motivation scenarios. Specifically, we present a novel examination of motivational incentives as moderators, which has not been examined in prior stereotype threat meta-analyses. These studies include differential incentives based on test performance (e.g., a $10–20 reward for top performers). By virtue of studying the operational and incentive samples, we examine whether motivation is a potential contributor to generalizability of stereotype threat outside of the lab.

Finally, the fifth line of investigation asks whether estimates of stereotype threat are inflated by publication bias in which null-result studies are suppressed from publication. We evaluate sensitivity to publication bias and the effect of small-sample studies using the test of excess significance (Ioannidis & Trikalinos, 2007), examination of funnel plot asymmetry and the trim and fill technique (Duval & Tweedie, 2000), and cumulative meta-analysis (Borenstein, Hedges, Higgins, & Rothstein, 2009). We also conduct a test of the influence of selective reporting (only reporting significant results, collecting data until a specified significance level is reached, etc.). Zigerell (2017) applied publication bias tests to a 2008 database; we conduct these analyses on a much larger, updated database.

Research Questions 1A–1D address the individual effects of each of the four moderators identified that are relevant to operational testing conditions. Below we outline each of these four moderators.

Addressing a Previously Unrecognized Methodological Error

We take into account a previously unrecognized methodological error present in about 15% of studies. These studies control for scores on a prior cognitive ability test, which can result in a biased estimate of stereotype threat. In experimental studies, looking at test performance within a subgroup (as is the case in all studies here, where subgroup members' scores in a threat condition and a nonthreat condition are compared), a result of random assignment to condition is that controlling for prior ability does not change the expected value of the mean for either the experimental or the control group. The purpose of a control variable in an experimental study is to reduce the standard error, thus increasing the statistical power of the study. This is advantageous for conducting tests of statistical significance, but estimating the effect size (i.e., *d*, the standardized mean difference between groups) still needs to be done using unadjusted scores. However, studies using adjusted scores commonly report means and standard deviations only for adjusted scores, and prior studies have computed *d* using these adjusted standard deviations. Given the smaller standard deviation of adjusted scores, the observed mean difference between groups is divided by a smaller *SD* when computing *d*, thus resulting in an upwardly biased estimate of *d*. As we will show, this biasing effect is large, with mean *d* 67% larger in studies that control for a prior ability test and use an adjusted standard deviation. This error affected prior meta-analyses. We note that this is a methodological error with respect to within-group comparisons (e.g., control vs. threat group for females); the same logic does not apply for between group comparisons (e.g., females vs. males). After documenting the effect of using adjusted *SD*s, we remove studies using adjusted *SD*s from all substantive analyses.

> *Research Question 1A*: What is the magnitude of stereotype threat after removing studies which control for a prior ability test?

Metric of Cognitive Ability Score

Threat studies score tests in different ways. The most common is a straightforward tally of the number of items correct. However, about 10% of studies instead compute the proportion correct among items attempted. Authors of these studies argue that accuracy is a more meaningful metric than total correct because accuracy accounts for more information; namely, number of questions correct and number attempted as opposed to just number correct (Schmader & Johns, 2003; Shih, Pittinsky, & Ambady, 1999). As we will show, such studies produce larger average effects. Such scoring would not be used in operational settings, as it is readily coached to yield highly inaccurate estimates of test-taker ability. Test-takers would be instructed to only answer questions for which they are highly confident in their answers, and leave all other items blank, resulting in a high proportion correct score regardless of true ability. Thus, we exclude studies using proportion correct from our focal sample. This moderator has not been examined in prior meta-analyses.

> *Research Question 1B*: What is the effect of the metric of cognitive ability test score on the magnitude of stereotype threat?

Comparison Group

About 60% of threat studies compare a threatened group with a control group; the rest contrast a threatened group with a threat removal group. Control groups are typically groups in which no information is given about group differences in test performance. A common threat removal manipulation is to tell participants that there are no mean group differences on the test they are about to take, though there are, in fact, mean differences. Such instructions would not likely be possible in operational settings, as presenting inaccurate information about a test would be unethical. Thus, we exclude studies using threat removal groups from our focal sample. Flore and Wicherts (2015) found that comparison group type displayed a small, nonsignificant influence on effect size, whereas Nguyen and Ryan (2008) do not examine this moderator.

> *Research Question 1C*: What is the effect of using a control versus a removal group on the magnitude of stereotype threat?

Stereotype Threat Activation Strength

Threat studies vary in the strength of the threat manipulation, from subtle manipulations (such as priming race/ethnicity or gender by seeking demographic information just prior to taking a test) to blatant manipulations (such as telling test takers that the test they are about to take is one in which members of a particular group tend to obtain lower scores). Blatant manipulations would not be present in any professional testing setting due to ethical concerns (Ryan & Sackett, 2013), because these strategies directly state that there are group differences on the test at hand (or ability type in question) just before taking the test. Thus, our focal analyses examine studies using subtle manipulations that could plausibly be present in operational testing settings. Nguyen and Ryan (2008) found inconclusive results regarding this moderator, with subtle activation strategies producing the largest effect for females and moderately explicit strategies producing the largest effect for racial minorities. We reexamine this moderator with a larger database, while focusing on subtle activation strategies attributable to their potential relevance to operational testing scenarios.

> *Research Question 1D*: What are the effects of subtle, moderately explicit, and blatant stereotype threat activation strategies on the magnitude of stereotype threat?

We note that our strategy of excluding threat removal strategies and blatant manipulations from our focal sample is not intended as criticism of studies that use these approaches. It is perfectly reasonable to use these approaches for a broad understanding of the stereotype threat phenomenon. Such research asks questions regarding the magnitude of effects that *can* occur in settings of interest to the researcher. We argue here that care must be taken in the leap from what *can* occur to what *does* occur in operational settings.

As mentioned previously, we also consider all four of these moderators in conjunction, isolating only features which display similarity to the methodological and analytic decisions that would be seen in operational testing conditions.

> *Research Question 2*: What is the magnitude of stereotype threat for samples which use total correct as the cognitive ability metric, use a control group as a comparison to the stereotype activation group, use a subtle stereotype threat activation manipulation, and do not control for prior ability scores?

Considering the Influence of Motivation

Research Questions 3 and 4 involve subsets of studies that are likely to contain test-takers that are more motivated than is typical. Most studies utilize college students participating in research studies for course credit. Because performance on tests in these experimental settings generally does not affect participants beyond the experiment itself, high test-taker motivation cannot be assumed. In fact, a plausible interpretation of the stereotype threat effect in lab settings can be attributed to motivation. If individuals in a stereotyped group are told or infer that their group performs poorly on the test at hand, participants may not be motivated to invest high levels of effort, and consequently devote reduced effort to completing the test. This scenario is very different than high stakes tests, where consequences of test performance incentivize test-takers to exert high effort and remain focused. Thus, we first focus on a small subset of large-sample experiments using operational college placement exams.

> *Research Question 3*: What is the magnitude of stereotype threat in operational testing settings, and how does this effect compare to studies in lab settings?

To provide insight into the role of motivation on test performance in low stakes settings, Duckworth, Quinn, Lynam, Loeber, and Stouthamer-Loeber (2011) present a meta-analysis showing that financial incentives affect performance on ability tests. Incentive effects were consistently found, with large effects found with larger incentives and with lower-ability test takers. Here we focus on a subset of laboratory studies in which incentives are present that serve to increase test-taker motivation. Such incentives are found in the form of differential incentives (either money or course extra credit) offered to the top performers on ability tests. Another manipulation with potential to raise motivation is the experimenter telling participants to imagine they are applicants for a desirable job before taking the test. It is clear these manipulations and incentives are of a lower magnitude than an operational test with real consequences. Yet, we view these studies with motivational features as useful for shedding light on motivation's role in the external validity of stereotype threat lab studies.

> *Research Question 4*: What is the magnitude of stereotype threat in settings that invoke testing conditions intended to increase test-taker motivation?

Other Moderators of the Stereotype Threat Effect

Research Questions 5A and 5B address two other moderators that are not differentially relevant to operational test settings to provide a comprehensive meta-analysis of stereotype threat, in addition to the focal sample with explicit focus on threat in operational settings.

Domain Identification

A proposed boundary condition for stereotype threat is that the threatened individual must identify within the domain being tested (i.e., identify in math for a math test; Steele, 2010; Steele & Aronson, 1995) to be concerned about negative stereotypes in the domain. In their meta-analytic sample, Nguyen and Ryan (2008) found that those who identified with the domain being tested a medium amount experienced a slightly larger stereotype threat effect than those who identified highly within the domain being tested. However, the *k*s were small for each analysis (12 and 9, respectively). We broaden the examination of domain identification to compare those who have been preselected as identified within a domain versus those who are not, to gather a larger sample of studies for this moderator.

> *Research Question 5A*: Does whether test-takers are identified within the domain of the test influence the magnitude of stereotype threat?

Test Difficulty

Another proposed boundary condition is that the test must be sufficiently difficult to observe the stereotype threat effect (Steele, 2010). As difficult tests demand more cognitive resources than easy tests, it is proposed that on these difficult tests stereotype threat can interfere with cognitive capacity for the threatened groups (Steele, Spencer, & Aronson, 2002). Nguyen and Ryan (2008) found a clear pattern of moderation, with the stereotype threat effect increasing as test difficulty increased. Rather than conceptualizing test difficulty as a boundary condition in which a certain difficulty threshold must be present for threat to occur, these results suggest a linear effect, with greater test difficulty producing a larger effect. We revisit this moderator at different levels of test difficulty with a larger sample.

> *Research Question 5B*: Does cognitive ability test difficulty level alter the magnitude of stereotype threat?

Publication Bias in Stereotype Threat Literature

Although the general conclusion has been that stereotype threat exists and is robust, two recent meta-analyses have raised question about the nature of this effect (Flore & Wicherts, 2015; Stoet & Geary, 2012). Publication bias occurs when research appearing in a published literature is systematically unrepresentative of the population of studies on the topic (Rothstein, Sutton, & Borenstein, 2006). Flore and Wicherts (2015) are thorough in their evaluation of publication bias, finding an overall effect size of $d = -.22$, which is decreased substantially after conducting adjustments for publication bias. *p*-hacking occurs when scientists exploit ambiguity in analyses to obtain statistically significant results (Simonsohn, Nelson, & Simmons, 2014a). Examples include choice of which measures to analyze, what unit of dependent variable to use, and which covariates to use in analyses. In the only test for questionable research practices (QRPs) in the stereotype threat domain to date, Flore and Wicherts (2015) did not find evidence of significant QRPs in their database.

In a recent commentary on Nguyen and Ryan's (2008) meta-analysis, Zigerell (2017) used Nguyen and Ryan's meta-analytic dataset to conduct multiple tests that adjust meta-analytic stereotype threat estimates for publication bias. Depending on the adjustment method, Zigerell found the stereotype threat effect either (a) reduced to a null effect or (b) reduced in magnitude but still a nonzero effect. In a reply, Ryan and Nguyen (2017) conclude that "at this point . . . we do not see conclusive evidence of publication bias so strong as to conclude there is no such thing as stereotype threat" (p. 1174). They also note that "hundreds more studies have been conducted on stereotype threat effects in the past decade, and thus an updated meta-analysis on this topic is warranted" (p. 1175). The present study, well underway when the Zigerell commentary and Ryan and Nguyen reply appeared, represents just such an update, with 88% more samples, representing the largest meta-analysis on stereotype threat to date.

> *Research Question 6*: Does potential existence of publication bias and selective reporting practices affect conclusions sur-

rounding the existence and magnitude of stereotype threat on cognitive ability tests?

Method

Selection of Studies

We used three strategies for locating stereotype threat studies. First, we searched for all published and unpublished studies used in analyses from the most comprehensive meta-analysis on threat in adults to date (Nguyen & Ryan, 2008). We did not use Zigerell (2017) because that study used the Nguyen and Ryan database. In total, 76 articles were identified from the Nguyen and Ryan meta-analysis and 65 were located. After applying inclusion criteria stated below, 48 of these were included. Second, we conducted a computerized literature search for primary stereotype threat studies through the online databases of PsycINFO, PsycARTICLES, Google Scholar, ProQuest Digital Dissertations, and Social Sciences Citation Index from the year 1995 (the year the seminal stereotype threat article was published; Steele & Aronson, 1995) to April 2017. The keywords *stereotype* and *threat* were used as search parameters. This identified 178 published articles with potential for inclusion, and 59 were included. Third, we contacted all first authors of studies included in our database and asked about shareable unpublished articles or conference presentations. Between online literature searches for unpublished studies (i.e., dissertations, theses) and contacting authors, we located 58 unpublished studies, 31 of which were included. Our sample is quite distinct from the Nguyen and Ryan (2008) database: 62% (132 samples) of our samples are nonoverlapping with Nguyen and Ryan.

Inclusion Criteria

We included studies based on four criteria. First, samples had to have a mean age of 18 years old or older. However, we excluded elderly samples examining age-based stereotypes, as Lamont, Swift, and Abrams (2015) address age-based stereotype threat in a recent meta-analysis. Second, studies were only included if they used an experimental design and contained a stereotype threat manipulation. The experimental design needed to contain a stereotype threat activation group and either a control group or a stereotype removal group from the same threatened group (e.g., females). We note that four field studies met this inclusion criterion because they also contained a stereotype threat activation manipulation with activation and control groups. We evaluate stereotype threat within stereotyped groups rather than comparing across threatened and nonthreatened groups. This is because we view the baseline requirement of the threat effect is that members of a threatened group perform worse in conditions hypothesized as more threatening than members of the same group in less threatening conditions (Sackett & Ryan, 2012).

Third, a study needed to contain a test of cognitive ability as a dependent variable that was not a highly speeded test. We excluded basic math operations or arithmetic highly speeded tests (e.g., Krendl, Richeson, Kelley, & Heatherton, 2008; Seitchik, 2013) because these display more limited generalizability to operational test-taking scenarios. For example, Krendl and colleagues (2008) administered a test in which participants were given five seconds each to solve basic operations math problems (i.e., "Is 5 × 2 – 3 = 7?). We excluded these tests when they were highly speeded because we view this as a measure of processing speed.[1] Although it is a speeded test, we include samples using the Wonderlic because it is used in employment testing settings and correlates highly with standardized tests (e.g., the SAT and ACT; Coyle, 2007). We also excluded group tasks as outcomes, which consisted of more than one individual working on the same test. Finally, we exclude learning tasks such as the one found in Taylor and Walton (2011), in which participants studied rare words in threatening and nonthreatening conditions and were tested on recall one week later. In terms of determining what constituted a cognitive ability test, we referred to the specific abilities covered in Hough, Oswald, and Ployhart (2001): verbal ability, quantitative ability, spatial ability, memory, and mental processing speed. As mentioned above, we exclude the processing speed measures (consisting of basic math operations) because of limited generalizability to high stakes testing settings.

Fourth, studies needed to include statistics that were convertible to a weighted effect size of Cohen's *d*. In a small number of cases, (e.g., Lee & Nass, 2012; Van Loo & Rydell, 2013) although the needed statistics were not reported directly, results were reported in bar graphs (i.e., *M*s, *SD*s or *SE*s) that allowed for calculation of an effect size using a plot digitizer (http://arohatgi.info/WebPlotDigitizer/).[2] We also excluded three studies because of retractions from a research fraud investigation involving Diederik Stapel (Marx & Stapel, 2006a, 2006b; Marx, Stapel, & Muller, 2005).

Coding Procedure

Coding was done separately by the first and third authors, with interrater agreement for all variables above 90%. Absent agreement, differences were discussed and resolved. We coded for sample type, threat group, sample demographics, manipulation description, type of nonthreat group, relevant statistics to extract an effect size, reported *d* and *N* in prior meta-analysis, and moderators included.[3] If we could not extract an effect from the study itself, we contacted study authors through e-mail. There were a small number of studies where gender was nested within race (e.g., Schmader & Johns, 2003; Stricker & Ward, 2004[4]). For these studies, we categorized the study based on whether a race or

[1] We note that some tests of a similar format to Krendl et al. (2008) are included. The tests that were included used more involved and/or complex arithmetic and were not highly speeded.

[2] To evaluate the accuracy of the plot digitizer, we found six samples that reported means and standard deviations (or standard errors) both via statistics and bar graphs, calculating *d* values for the reported statistics and using the plot digitizer. The *d* values calculated via statistics versus bar graphs in these six samples differed by an average of $|d| = .01$, and no difference was greater than $|d| = .02$. Thus, the plot digitizer represented an accurate method to extract statistics.

[3] The coding spreadsheet with all relevant variables for each sample included is available upon request.

[4] For Stricker and Ward (2004), there was a threat based on race and gender. To avoid dependence of samples, we calculated two separate *d* values: one for Black participants (male and female) and one for female subjects excluding Black females. We calculated *d*s separately for each test and then averaged them, using only stereotype-relevant tests for each sample. For the female sample, we used the two math-relevant tests, whereas for the Black sample we used all four tests of math and verbal abilities.

gender stereotype was activated (i.e., if the study used a race-based threat, it was categorized as a race study). As such, for these studies we only extracted statistics for samples that were relevant to the stereotype that was activated.

Ability covariate. We coded whether scores on the cognitive ability test were adjusted based on a measure of ability. We extracted statistics that did not adjust for previous levels of ability when possible. For all analyses except an initial documentation of the overall threat effect, we remove studies where only statistics that adjust for previous ability were available, due to the upward bias in these effect size estimates.

Comparison group. We operationalized a control group as a condition that received no manipulation designed to influence stereotype threat. A stereotype threat removal group (STR) was defined as a condition that received information intended to reduce the effect of stereotype threat. Some studies compared a threat group, or stereotype activation group (STA), with a control group; others with an STR. In cases of two experimental conditions, the study contributed one effect size, calculated as $d_{\text{STA-Control}}$ or $d_{\text{STA-STR}}$. Effect sizes were calculated such that negative effect sizes indicate the presence of stereotype threat.[5] When studies contained three groups (Control, STR, and STA), we chose the control group and STA to comprise the effect size and excluded the removal group. This is in contrast to Nguyen and Ryan (2008), who calculated effects based on the STR and STA groups, excluding the control group. As removal strategies are implausible in operational test-taking scenarios because of ethical concerns of distributing inaccurate information to test-takers, the comparison of control and STA will yield a more accurate representation of the stereotype threat effect as it may exist outside of the lab setting.

When studies had multiple conditions within the same category (e.g., two STA groups, two control groups), we collapsed *M*s and *SD*s across these conditions, unit-weighting each condition. When collapsing these conditions within the same category, we took the average *N* for conditions in that category so as to not overweight that category. For example, if a study had two stereotype activation conditions with $N = 30$ and 34, the collapsed activation condition counted for $N = 32$. In a small number of cases, multiple stereotype-relevant cognitive ability tests were administered. In these cases, we first calculated *d* values for each test and then created a composite, unit-weighted *d* value across all relevant tests. If a test had multiple cognitive ability tests and only one was stereotype-relevant, only the stereotype-relevant test was used.

Stereotype threat activation strength. We use Nguyen and Ryan's (2008) typology to categorize and examine stereotype threat activation cues based on activation strength. Their typology of stereotype threat activation cues consists of subtle, moderately explicit, and blatant cues. Subtle activation cues occur when "the message of subgroup differences in cognitive ability is not directly conveyed" (Nguyen & Ryan, 2008, p. 1316). Moderately explicit activation cues occur when, "the message of subgroup differences in cognitive ability and/or ability performance is conveyed directly . . . but the direction of these group differences is left open for test takers' interpretation" (Nguyen & Ryan, 2008, p. 1316). Finally, blatant activation cues occur when "the message involving a stereotype about a subgroup's inferiority in cognitive ability and/or ability performance is explicitly conveyed" (Nguyen & Ryan, 2008, p. 1316).

Ability test metric. Many authors reported the cognitive ability test score as an accuracy score (No. correct/No. attempted) rather than a total score. We examined ability test metric (total correct vs. accuracy) as a moderator. When a total correct score was available, it was used over an accuracy score because of the coachability of the accuracy test metric and the unlikelihood of observing accuracy test scores in operational testing scenarios.

Motivational incentives/features. A small subset of studies offered differential incentives to participants based upon performance on the cognitive ability test. These studies primarily offered a monetary bonus ($10–$20) for top performers. One study offered extra credit to top performers, and another offered a gift to top performers. One other study feature that is intended to increase motivation uses an applicant prime; the experimenter instructs participants to imagine that they are applicants for a desirable job when taking the test. All studies involving an applicant prime also used financial incentives. These motivational analyses were run on a nonoverlapping sample as the operational test sample.

Treatment of Studies With Moderators

For studies that split experimental conditions into multiple groups across moderators between-subjects (e.g., low and high domain identification groups), when possible we extracted statistics separately across the different levels of moderators as defined by the researchers. Thus, each subsample contributed an independent effect size. If a moderator was within subjects, we collapsed across levels of the moderator to preserve independence of samples. However, when statistics reported did not allow extraction of independent effect sizes, we took the statistics reported collapsing across the moderator, as reported in the study.

Analyses

We used random effects meta-analysis and the Hunter and Schmidt procedure (Hunter & Schmidt, 2004) to estimate all meta-analytic effect sizes. We used the "psychmeta" package (Dahlke & Wiernik, 2018) in the statistical program R for all meta-analytic statistics and cumulative meta-analyses. We conduct moderator analyses across the moderators used in the focal sample as well as other moderators, and discuss hierarchical moderator analyses when conceptually appropriate. We present one focal analysis that highlights the potential effect of stereotype threat in conditions most similar to operational testing scenarios. Specifically, the focal analysis includes all studies that report number correct for the outcome, do not control for prior ability, use a control group rather than a removal group, and use subtle activation cues. We also present a supplementary analysis that evaluates whether the number of operational test-like conditions a sample contains (i.e., 0–4) affects the magnitude of stereotype threat.

For studies where test performance means and standard deviations were available, we computed Cohen's *d* (Cohen, 1988) as the measure of effect size. If a study did not provide condition sample size but reported overall sample size across both conditions, we

[5] All effect size measures are reported such that negative values represent threat decreasing test performance and positive values indicate threat increasing test performance (i.e., negative indicates presence of stereotype threat).

assumed an even split between the two conditions for the purposes of calculating the pooled *SD*[6]. To calculate sampling error variance of Cohen's *d*, we used formula 7.38 from Hunter and Schmidt (2004), computing sampling error based on mean effect size. We did not account for unequal group sizes in sampling error variance calculations because in some cases (e.g., *t* or *F*-statistic conversions, extracting statistics from plots), sample size per condition was not provided. We use a statistical significance approach to test for differences across moderators, found in Neter, Wasserman, and Whitmore (1988, p. 402). All *d* values were converted to correlations, and *t* statistics were computed based on the difference between correlations, using true correlation variance and moderator *k* as inputs. We use an alpha level of .05 for all moderator tests.

Cohen's (1988) guidelines for the interpretation of *d* values are as follows: small effect- .2, moderate effect- .5, and large effect- .8 or above. However, given the consequences for presence of stereotype threat on high stakes tests, we utilize a slightly more stringent interpretation guideline: .1 to .2 we also interpret as a small effect (in contrast to Cohen's guidelines which would consider this magnitude trivial). We note that reporting of reliability on cognitive ability tests was extremely sparse; we were only able to locate reliability values for 13 out of 212 (6%) samples and the studies that did report reliability were split between internal consistency and test–retest reliability. As such, we did not correct *d* values for unreliability because of instability of the artifact distribution. We also note cognitive ability tests inherently have less than perfect reliability and we are interested in estimating the effect of stereotype threat under these conditions rather than with a hypothetical perfect-reliability test.

We take a multifaceted approach to testing for publication bias because modern publication bias detection methods have their strengths and weaknesses (Inzlicht, Gervias, & Berkman, 2015). Although we document a number of publication bias tests, we present minimal discussion of these tests because Zigerell (2017) recently conducted many of these analyses on Nguyen and Ryan's (2008) stereotype threat database. These analyses are still conducted here because the stereotype threat database has increased since 2008. Publication bias tests are conducted on both the full and focal samples to discern publication bias in the full literature and the more focused studies relevant to operational testing scenarios. The first test we conduct is the exploratory test for excess significance (Ioannidis & Trikalinos, 2007), which tests whether there are more significant findings in the database than would be expected based on all samples' cumulative power. If there is more significance observed than expected, this could be attributable to publication bias. We also present funnel plots with study sample size (i.e., study precision) on the *y* axis, in addition to contour-enhanced funnel plots. The trim and fill method (Duval & Tweedie, 2000) is utilized with the metafor package in R (Viechtbauer, 2017). This method imputes values to create a symmetric funnel plot, portraying an estimate of what the effect size distribution (and its mean) may have been if publication bias were not present.[7] The trim and fill has been criticized on the grounds that effect size estimates can be inaccurate in the presence of between-study heterogeneity (Terrin, Schmid, Lau, & Olkin, 2003). Additionally, Simonsohn, Nelson, and Simmons (2014b) find via simulation that trim and fill undercorrects when publication bias is substantial, thus overestimating true effect sizes. As such, we view the trim and fill as a sensitivity test to the robustness of stereotype threat to publication bias rather than providing exact point estimates of the size of the effect with no publication bias (Banks, Kepes, & McDaniel, 2015).[8]

We also present a cumulative meta-analysis, in which an iterative meta-analysis is conducted on the most precise study (i.e., largest *N*), followed by the second most precise, and so on (Borenstein et al., 2009). To test for presence of QRPs in the stereotype threat literature, we used *p*-curve analyses (*p*-curve 4.0; Simonsohn, Simmons, & Nelson, 2015). The *p*-curve tests for whether a set of studies contains *evidential value*, defined as when selective reporting is ruled out as the sole explanation for a significant set of findings (Simonsohn et al., 2014a). To compute test statistics for the online *p*-curve 4.0 app (*p*-curve.com), we converted all *d* values to *t* test statistics and calculated accompanying degrees of freedom to obtain *p* values.

Results

Table 2 presents the overall meta-analytic estimate of stereotype threat in adults on tests of cognitive ability, as well as moderator analyses. A table detailing all samples included is available in the online supplementary material. Research Question 1A addressed the magnitude of stereotype threat after excluding studies that used an ability covariate. The overall standardized mean difference in this sample was $d = -.31$ ($k = 181$, $N = 10{,}436$, $SD_{\delta} = .38$), and would be slightly larger if ability control studies were incorrectly included ($d = -.33$, $k = 212$, $N = 11{,}521$, $SD_{\delta} = .40$). When looking at these studies in isolation, those with an ability covariate had a larger effect size than the overall effect estimate ($d = -.51$, $k = 31$, $N = 1{,}085$, $SD_{\delta} = .54$), although the covariate–no covariate comparison was not significant, $t(210) = 1.56$, $p = .06$. Going forward, we only present results after excluding studies that used an ability covariate, because of the artificial inflation that occurs in this set of studies. Further, we caution against interpretation of the overall mean effect because attending to study methodological choices and threat activation cues yields more accurate estimates of stereotype threat. This mean estimate is slightly stronger than that of Nguyen and Ryan (2008), who found an overall effect of −.26. The 90% credibility interval of our estimate ranges from −.93 to .32, indicating there is great variability and that the true effect is not always negative on stereotyped test takers. As such, we proceed to discuss the focal analyses and then moderators.

[6] To address the concern that Cohen's *d* uses a pooled *SD* (assuming that population *SD*s in both the treatment and control groups are equivalent), we conducted supplemental analyses with Glass' Δ. These results converged to support equivalence between Cohen's *d* and Glass' Δ, and are available from the first author upon request.

[7] Trim and fill analyses are conducted within all moderator categories because trim and fill assumes sampling error is the only source of variance across the dataset (Duval, 2005; Inzlicht, Gervias, & Berkman, 2015). We use the standard L0 estimator rather than the R0 estimator because the L0 is a more robust estimator, particularly under conditions of small *k* (Duval, 2005; Moreno et al., 2009). The function used does not assume studies are missing from the left or the right side; it imputes studies on the left or right side of the funnel plot based on the results of Egger's regression test.

[8] We chose not to conduct PET-PEESE analyses because we view its flaws to be severe. Specifically, the finding that PET-PEESE has consistent negative bias under conditions of publication bias (i.e., over-correction for publication bias; Gervias, 2015) and its frequent finding of a null effect even when there is a true non-zero effect (Gervias, 2015, 2016).

Table 2
Meta-Analytic Statistics for Overall Threat Effect and Moderator Analyses

	Meta-analytic statistics								Publication bias statistics			Significant differences between moderators (p values)	
Variable	k	N	Mean d	SD_d	SE_d	SD_δ	90% CrI	95% CI	TF d	95% CI	Estimated no. missing studies	Mod 1	Mod 2
Overall (removing ability control)	181	10,436	−.31	.46	.03	.38	−.93, .32	[−.38, −.24]	−.20	−.27, −.13	35		
Focal sample[a]	45	3,532	−.14	.38	.06	.31	−.66, .38	[−.26, −.03]	−.09	−.21, .03	8	<.01	
Control group type[b]													
Control group	114	7,611	−.28	.42	.04	.33	−.83, .28	[−.35, −.20]	−.17	−.25, −.08	26		
Removal group	67	2,825	−.39	.56	.07	.47	−1.17, .39	[−.53, −.26]	−.39	−.53, −.26	0	.04	
Threat strength[b]													
Subtle	63	4,149	−.17	.39	.05	.30	−.67, .34	[−.26, −.07]	−.10	−.20, .00	12		
Moderately explicit	51	2,766	−.38	.56	.08	.48	−1.19, .43	[−.54, −.23]	−.23	−.40, −.06	11	<.01	
Blatant	62	3,371	−.41	.42	.05	.31	−.93, .11	[−.52, −.30]	−.41	−.52, −.30	0	<.01	.73
DV metric[b]													
Total correct	169	9,929	−.30	.46	.04	.38	−.93, .33	[−.37, −.23]	−.19	−.26, −.11	35		
Accuracy	12	507	−.50	.36	.11	.18	−.82, −.18	[−.73, −.33]	−.61	−.86, −.35	3	.02	
Threat group type													
Ethnic minority	35	2,432	−.26	.40	.07	.32	−.81, .28	[−.40, −.12]	−.18	−.33, −.03	7		
Female	136	7,442	−.33	.47	.04	.38	−.97, .30	[−.41, −.25]	−.21	−.29, −.12	27	.12	
Test difficulty													
Easy	39	3,323	−.24	.41	.07	.34	−.82, .34	[−.37, −.11]	−.11	−.26, .03	12		
Moderate	78	3,958	−.30	.49	.06	.40	−.97, .37	[−.41, −.19]	−.18	−.31, −.06	14	.23	
Difficult	17	999	−.40	.34	.08	.21	−.76, −.04	[−.57, −.22]	−.37	−.55, −.20	1	.04	.09
Domain identification-preselected													
No	149	9,266	−.30	.46	.04	.38	−.93, .33	[−.37, −.22]	−.19	−.27, −.11	29		
Yes	32	1,170	−.40	.48	.09	.34	−.98, .18	[−.58, −.23]	−.31	−.49, −.13	5	.10	
Publication status													
Published	132	7,469	−.37	.51	.04	.43	−1.08, .34	[−.46, −.28]	−.23	−.32, −.14	28		
Unpublished	49	2,967	−.17	.32	.05	.19	−.48, .14	[−.27, −.08]	−.13	−.23, −.04	6	<.01	
Sample size													
Under 50	125	4,005	−.39	.64	.06	.52	−1.26, .47	[−.51, −.28]	−.39	−.51, −.28	0		
50–99	34	2,314	−.52	.31	.05	.18	−.82, −.22	[−.63, −.41]	−.64	−.76, −.53	9	.01	
100 and over	22	4,117	−.12	.22	.05	.16	−.39, .16	[−.21, −.02]	−.08	−.18, .03	3	<.01	<.01
Sample setting type													
Operational samples[c]	4	1,670	−.01	.11	.06	.06	−.15, .13	[−.19, .17]	—[e]	—	—		
Lab samples	177	8,766	−.36	.48	.04	.39	−1.00, .27	[−.44, −.29]	−.27	−.35, −.19	25	<.01	
Motivational incentives													
Yes	11	697	−.14	.45	.14	.38	−.82, .54	[−.45, .16]	−.10	−.40, .20	1		
No	137	6,690	−.41	.49	.04	.40	−1.07, .25	[−.50, −.33]	−.31	−.40, −.22	20	.03	
Motivational incentives													
Financial incentives	9	526	.00	.43	.14	.34	−.63, .62	[−.33, .33]	—[e]	—	—		
Financial & applicant[d]	6	443	−.03	.36	.15	.27	−.57, .51	[−.40, .35]	—[e]	—	—	—[f]	
Other	2	171	−.57	.20	.14	.00	−.57, −.57	[−.2.38, 1.24]	—[e]	—	—	—[f]	—[f]

Note. k = number of samples; N = number of subjects; Mean d = sample-size weighted Cohen's d; SD_d = sample-size weighted standard deviation of d values; SE_d = standard error of mean d; SD_δ = true standard deviation of d values; 90% CrI = 90% credibility interval; 95% CI = 95% confidence interval; TF d = Trim and fill d value imputing missing studies based on funnel plot asymmetry; 95% CI (in publication bias columns) = 95% confidence interval around trim and fill-adjusted d; Estimated no. of missing studies = the number of studies trim and fill estimates is missing and imputes (on either the left or right side of the distribution) based on funnel plot asymmetry.

[a] Focal sample contains studies that did not use an ability control, used total number correct on the ability test, used a subtle threat manipulation, and had a control group. [b] Denotes moderator that was used to subset studies for the focal sample. [c] The operational samples use only samples that were taken from real-world testing scenarios. All samples utilized for this analysis were college placement examinations. [d] The "Financial & Applicant" subset is nested within the "Financial Incentives" sample. [e] We do not conduct trim and fill analyses on any moderators with k less than 10 because of instability in estimates. [f] Because of nesting of samples and small k, we do not present significance tests on the comparison of moderators within the motivational incentives category.

Focal Sample Moderators

Research Question 1B addressed the effect of ability test metric on stereotype threat. Studies using total number correct found a threat effect of $d = -.30$ ($k = 169$, $N = 9{,}929$, $SD_{\delta} = .38$), whereas studies using test accuracy found a stronger threat effect of $d = -.50$ ($k = 12$, $N = 507$, $SD_{\delta} = .18$), which was significant at an alpha of .05, $t(179) = 2.09$, $p = .02$. Thus, ability test metric was a notable moderator and studies using the less common accuracy metric produced a systematically larger threat effect than studies using traditional total scores.

Research Question 1C addressed the effect of using a control group versus a removal group when comparing test scores to a threat activation group. Samples that used a control group had a significantly weaker threat effect at an alpha of .05 ($d = -.28$, $k = 114$, $N = 7{,}611$, $SD_{\delta} = .33$) than samples using a removal group ($d = -.39$, $k = 67$, $N = 2{,}825$, $SD_{\delta} = .47$), $t(179) = 1.78$, $p = .04$. Thus, comparison group type was also a significant moderator; studies using a removal group produce a larger effect on average than the typical experimental paradigm comparing an experimental group and a control group.

Research Question 1D addressed the effects of subtle, moderately explicit, and blatant stereotype threat activation strategies on the magnitude of threat. Nguyen and Ryan (2008) found inconclusive patterns regarding stereotype activation strength as a moderator. With our more robust sample of studies, we find that stereotype threat becomes stronger when shifting from not mentioning group differences (i.e., subtle activation) to explicitly mentioning group differences (i.e., moderately explicit and blatant activation). In the overall sample, subtle cues produced the weakest threat effect ($d = -.17$, $k = 63$, $N = 4{,}149$, $SD_{\delta} = .30$), followed by moderately explicit cues ($d = -.38$, $k = 51$, $N = 2{,}766$, $SD_{\delta} = .48$), with blatant cues producing a comparable threat effect as moderately explicit cues ($d = -.41$, $k = 62$, $N = 3{,}371$, $SD_{\delta} = .31$). Both moderately explicit and blatant cues had a significantly stronger threat effect than subtle cues, whereas differences were nonsignificant and trivial when comparing the two stronger threat activation cues. This pattern was also found for females and racial minorities, with subtle cues producing the weakest effects (ds $= -.19$ and $-.14$, respectively), followed by moderately explicit (ds $= -.37$ and $-.51$, respectively). Blatant cues produced the strongest effect in females ($d = -.45$), although this could not be calculated for racial minorities because of a k of only two. Overall, all moderators examined in the focal sample influenced stereotype threat's effect size.

Focal, Operational, and Motivational Analyses

Research Question 2 addressed the magnitude of stereotype threat in the focal sample (i.e., studies that display similarity to operational testing conditions by using subtle activation cues, total correct ability scores, control groups as comparison groups, and making no adjustments for prior ability). Critically, in this focal sample a mean threat effect of $d = -.14$ ($k = 45$, $N = 3{,}532$, $SD_{\delta} = .31$) was found. Thus, our estimate of stereotype threat in studies with features most similar to operational testing conditions is small in magnitude. The true variation in this estimate is large, suggesting that even in this relatively narrow subset of studies the true effect varies considerably.

We conduct supplementary analyses to assess the influence of similarity to operational test-like conditions on stereotype threat. Samples were coded as 0 (absent) or 1 (present) on each of the four operational sample moderators and values were summed, such that higher scores indicate greater similarity to operational testing conditions. Samples with one, two, three, and four operational test-like conditions produced effects of $d = -.41$, $-.47$, $-.38$, and $-.14$, respectively. Thus, the reduction in threat comes from using all four operational test conditions rather than subsets of these conditions, which is consistent with the findings that each of these conditions moderates the effect size.

Research Question 3 addressed the magnitude of the stereotype threat effect for all available samples conducted in operational settings and how this compared with studies in lab settings. There were four such samples in our database and all used college placement exams.[9] The estimate of stereotype threat in operational contexts was $d = -.01$ ($k = 4$, $N = 1{,}670$, $SD_{\delta} = .06$). Studies in lab contexts yielded a much stronger effect of $d = -.36$ ($k = 177$, $N = 8{,}766$, $SD_{\delta} = .36$). The difference between lab and operational samples was significant $t(179) = -5.73$, $p < .01$.

Research Question 4 addressed the effect of motivational incentives on the magnitude of stereotype threat. When any form of motivational incentive was present, the magnitude of stereotype threat was $d = -.14$ ($k = 11$, $N = 697$, $SD_{\delta} = .45$), which was significantly weaker than $d = -.41$ ($k = 137$, $N = 6{,}690$, $SD_{\delta} = .49$) when no motivational incentives were present, $t(146) = 1.85$, $p = .03$. This motivational incentive subset was further broken down and examined by type of motivational incentive. Studies utilizing monetary incentives yielded a null effect ($d = .00$, $k = 9$, $N = 526$, $SD_{\delta} = .43$), and the subset of monetary incentive studies also using an applicant prime yielded $d = -.03$ ($k = 6$, $N = 443$, $SD_{\delta} = .36$). However, the applicant prime studies were fully nested within the financial incentive studies, making interpretation of the applicant-prime effect challenging. Overall, studies involving forms of motivational incentives served to decrease the stereotype threat effect.

Other Moderators

The effect of stereotype threat was slightly stronger for females ($d = -.33$, $k = 136$, $N = 7{,}442$, $SD_{\delta} = .38$) than for ethnic minorities ($d = -.26$, $k = 35$, $N = 2{,}432$, $SD_{\delta} = .32$), although this difference was nonsignificant, $t(169) = -1.20$, $p = .12$.

Research Question 5A addressed the threat effect magnitude for subjects who identified within the tested domain as compared with those who did not. Larger effects were found for participants who were preselected as identified with the domain of the test ($d = -.40$, $k = 32$, $N = 1{,}170$, $SD_{\delta} = .34$) than for those who were not ($d = -.30$, $k = 149$, $N = 9{,}266$, $SD_{\delta} = .38$). However, this difference was not significant, $t(179) = -1.30$, $p = .10$, and confidence intervals are almost entirely overlapping, thus failing to

[9] We also note that in one of the four operational samples, the threatened group was males (Anderson, 2001) and involved a test of arts and humanities. Gender-primed males were hypothesized to perform worse than nonprimed males because of the stereotype that men perform below women in those domains. This sample accounted for less than 9% of the total operational sample N, so received proportionally low weight in this subset. Removing this one sample from the operational analysis changed d from $-.01$ to $-.04$.

support this feature as necessary to produce stereotype threat. Hierarchical moderator analysis showed this pattern was retained in both females and ethnic minorities.

Research Question 5B addressed the extent to which test difficulty altered the stereotype threat effect. In line with previous research, test difficulty shows a gradient of stronger effect size with increasingly harder tests. Easy tests (50–100% correct) produced an effect size of $d = -.24$ ($k = 39$, $N = 3{,}323$, $SD_\delta = .34$), moderate tests (25–50% correct) produced an effect size of $d = -.30$ ($k = 78$, $N = 3{,}958$, $SD_\delta = .40$), and difficult tests (<25% correct) produced an effect size of $d = -.40$ ($k = 17$, $N = 999$, $SD_\delta = .21$). Difficult tests produced a significantly stronger effect than easy tests, $t(54) = -1.82$, $p = .04$, although easy–moderate and moderate–difficult comparisons were nonsignificant. This pattern of increasing magnitude of stereotype threat with harder tests is also observed in Nguyen and Ryan (2008).

One concern is that because the four operational samples have large *N*s, they carry disproportionate weight in the moderator analyses. To address this issue, we present all moderator analyses conducted removing the four operational samples in the online supplementary materials. All patterns across moderators were replicated in this analysis except for test difficulty: easy tests produced the greatest mean difference ($d = -.42$), followed by difficult tests ($d = -.40$), then moderate ($d = -.33$). This was because three of four operational samples fell into the "Easy" category. One might interpret the fact that ¾ of operational samples are easy tests as evidence against the validity of the null effect in operational samples. However, the mean effect for all other 36 easy samples was substantial, which argues against the notion that easy tests are the reason why operational samples do not find a threat effect.

We also present intercorrelations between moderators in Table 3 to discern the extent to which presence of certain study characteristics covary with other study characteristics. The largest moderator intercorrelation was between accuracy score and test difficulty ($r = .34$), and the mean absolute value of all intercorrelations was $|r| = .12$. As these values are predominantly small in magnitude, we view the moderator results as largely independent.

Publication Bias Analyses

Research Question 6 addressed whether publication bias and selective reporting affect conclusions. We first sought to quantify the degree of publication bias. As an initial indication that publication bias may affect this literature, published studies displayed more than twice as strong of a threat effect ($d = -.37$, $k = 132$, $N = 7{,}469$, $SD_\delta = .43$) as unpublished studies ($d = -.17$, $k = 49$, $N = 2{,}967$, $SD_\delta = .19$), $t(179) = -3.08$, $p < .01$. Study sample size also was linked to differing effect sizes. Studies with both small ($N < 50$) and moderate ($100 > N > 49$) sample sizes had sizable mean effect sizes (*d*s = −.39 and −.52, respectively), whereas studies with large sample sizes yielded a weak mean effect ($d = -.12$). All sample size categories were significantly different from one another, indicating the importance of sampling error when evaluating stereotype threat studies. The correlation between effect size and sample size in the overall sample was $r = .09$,[10] with larger samples producing slightly weaker effect sizes.

We conducted the test of excess significance (Ioannidis & Trikalinos, 2007) to evaluate the presence of statistically significant effects as compared with the cumulative power of samples included, presented in Table 4. Both the overall sample and focal sample displayed substantially more significance effects than expected based on cumulative power, indicating potential existence of publication bias. Figure 1 shows that all funnel plots are asymmetrical, with more studies displaying strong negative effects than strong positive effects. The sample size funnel plots show the majority of studies have relatively small samples, with three large sample studies ($N > 400$) clustering around an effect of approximately zero. The contour-enhanced funnel plot of the overall sample shows many studies were significant, whereas the focal sample shows proportionally fewer significant samples. When correcting for funnel plot asymmetry, the trim and fill adjusted estimates of stereotype threat for the focal sample was $d = -.09$, which was a decrease of 36% from the estimated effect size. Although the unadjusted focal sample estimate was statistically significant at an alpha level of .05 (95% CI [−.26, −.03]), the trim and fill adjusted estimate was not significant (95% CI [−.21, .03]). Trim and fill estimates for the overall threat effect and the focal sample appear to be subject to moderate publication bias (Kepes, Banks, McDaniel, & Whetzel, 2012). Within-moderator publication bias tests address the issue that publication bias adjustments can be problematic in the case of substantial between-study heterogeneity. Across all moderators, the average percent decrease in effect size was 23% with the trim and fill. One caveat to this finding was that the SD_δs within the moderator categories are similar to the SD_δ in the overall sample, suggesting that effect sizes may not be more homogeneous with moderator categories.

Cumulative meta-analysis was used to evaluate the amount of drift in the threat effect when iteratively less precise studies are added to the database. Results from the cumulative meta-analyses on the overall and focal samples are presented in Figure 2. This figure shows substantial negative drift in the overall sample in which the less precise studies make the stereotype threat effect larger (i.e., more negative). In fact, the point estimate of stereotype threat from the 10% most precise studies is $d = -.11$ ($k = 18$, $N = 3{,}700$), whereas samples with an $N \leq 35$ have an effect more than three times as large, $d = -.38$ ($k = 78$, $N = 2{,}076$). We note that the case observed where the most precise studies showed weaker threat effects could be attributable to sampling error in small studies and publication bias, but could also be attributable to unexplored moderators. The cumulative meta-analysis for the focal sample shows substantially less negative drift than the overall cumulative meta-analysis.

To examine potential presence of questionable research practices and selective reporting, we use the *p*-curve on the focal and overall samples. Refer to Figure 3 for the *p*-curve figures and Table 4 for accompanying test statistics. Based on multiple tests, the *p*-curve for the focal sample primarily displays evidential value in three of the four tests conducted. The *p*-curve for the overall

[10] When the three sample size outliers' (Anderson, 2001; Stricker & Ward, 2004) *N* were replaced with the next largest *N* (213), the correlation between sample size and effect size remained comparable at $r = .08$. Furthermore, replacing the three studies whose sample size was substantially larger than all other samples with the next largest sample size ($N = 213$; Anderson, 2001; Stricker & Ward, 2004), the overall effect size increased in absolute magnitude by only .02 to $d = -.33$ in the overall sample and by .03 to $d = -.17$ in the focal sample.

Table 3
Moderator Intercorrelation Matrix

Variable	1	2	3	4	5	6	7	8	9	10
1. Publication status[a]										
2. DI-Preselected[b]	.01									
3. Control group type[c]	.08	−.07								
4. Ability control[d]	−.04	−.13	−.23							
5. Accuracy score[e]	.09	−.12	−.16	.16						
6. Threat group type[f]	.23	−.22	.25	.03	−.14					
7. Test difficulty[g]	−.02	−.26	.23	.00	.34	.05				
8. Threat strength[h]	−.16	−.05	−.05	−.25	.30	−.31	.03			
9. Sample *N*	.03	−.14	.17	−.12	−.06	.08	.11	−.07		
10. *d* value	.16	−.07	.01	−.09	.00	.04	−.06	−.10	.09	

[a] Published coded as 0, unpublished coded as 1. [b] Not preselected on domain identification = 0, Preselected on domain identification = 1. [c] Threat removal group = 0, Control group = 1. [d] No ability control = 0, Ability control = 1. [e] Total correct = 0, Accuracy score = 1. [f] Females = 0, Racial minority = 1. [g] Test difficulty was coded as a continuous variable (% correct). [h] Subtle threat = 1, Moderately explicit threat = 2, Blatant threat = 3. When both variables were dichotomous, a phi correlation was calculated. When one variable was dichotomous and the other was continuous, a point-biserial correlation was calculated. All correlations are unweighted and were calculated on the subset of samples with ability covariate samples removed (k = 181), with the exception of the ability control correlates that were calculated on the full sample (k = 212).

sample displays evidential value for all tests conducted. In sum, tests indicate that it is unlikely the excess of significance found in the stereotype threat database is due to QRPs.

Discussion

We found the magnitude of the overall stereotype threat effect in adults to be comparable with other meta-analytic estimates of stereotype threat in children and adolescents (Flore & Wicherts, 2015), and slightly stronger than Nguyen and Ryan's (2008) original meta-analysis of threat effects in adult and student samples. We discovered a number of moderators relevant to operational testing scenarios that influenced stereotype threat's mean effect size. Our study addressed the previously unrecognized error of miscomputing effect size measures when using an ability covariate for within-group comparisons, finding a larger (but not significantly larger) threat effect in the subset of studies using a covariate. Prior meta-analyses that make comparisons within stereotyped groups (i.e., comparing females in threat and control groups) and do not attend to this issue of a covariate may contain artificially inflated estimates of stereotype threat.

The other three features relevant to operational testing scenarios were stereotype activation strength, test accuracy metric, and comparison group type. The previously inconclusive moderator of stereotype activation strength was revisited with a larger database and a monotonic pattern was found: subtle activation strategies produced a significantly weaker effect than moderate and blatant activation strategies, and blatant activa-

Table 4
Additional Publication Bias Analyses

		Test of excess significance (Ioannidis & Trikalinos, 2007)			
Sample	Total *k*	Expected significant effects[a]	Observed significant effects	$\chi^2(df)$[b]	*p* value
Overall sample	181	55.0	78	14.4 (1)	<.01
Focal sample	45	6.3	14	11.3 (1)	<.01

		p-curve analyses (Simonsohn, Simmons, & Nelson, 2015)			
	# of *p* values[d]	Test of right-skew[c]		Test of flatness[c]	
		Z value	*p* value	*Z* value	*p* value
Overall sample – Full *p-curve*	59	−6.09	<.01	1.34	.91
Overall sample – Half *p*-curve	39	−7.53	<.01	9.36	.99
Focal sample – Full *p*-curve	11	−2.89	<.01	.75	.77
Focal sample – Half *p*-curve	9	−1.45	.07	3.18	.99

[a] Expected significant effects calculated based on the cumulative power of all studies included. [b] The formula for the chi-square statistic for the test of excess significance is: $A = [(O - E)^2/E + (O - E)^2/(n - E)] \sim \chi^2$, where the individual study is i ($i = 1 \ldots n$) across n samples, O is the number of "positive" or significant samples at a specified alpha level (.05 in our case), and E is the number of expected significant samples. $E = \Sigma_{i=1}^{n}(1 - \beta_i)$, which is the sum of all samples' power at the specified alpha level. [c] These two tests evaluate (a) whether the full and half *p*-curves are significantly right-skewed, and (b) whether the set of findings is significantly flatter than the 33% power curve. Evaluating both skew and flatness of the *p*-curve can rule out selective reporting as the sole explanation for significance (i.e., evidential value). To detect evidential value, the test of right-skew should be significant and the test of flatness should be nonsignificant. See Simonsohn, Nelson, and Simmons (2014a) for further description of these tests. [d] Simonsohn and colleagues (2014a) present simulations indicating that the *p*-curve quite accurately reaches conclusions on evidential value with 10 to 20 *p* values.

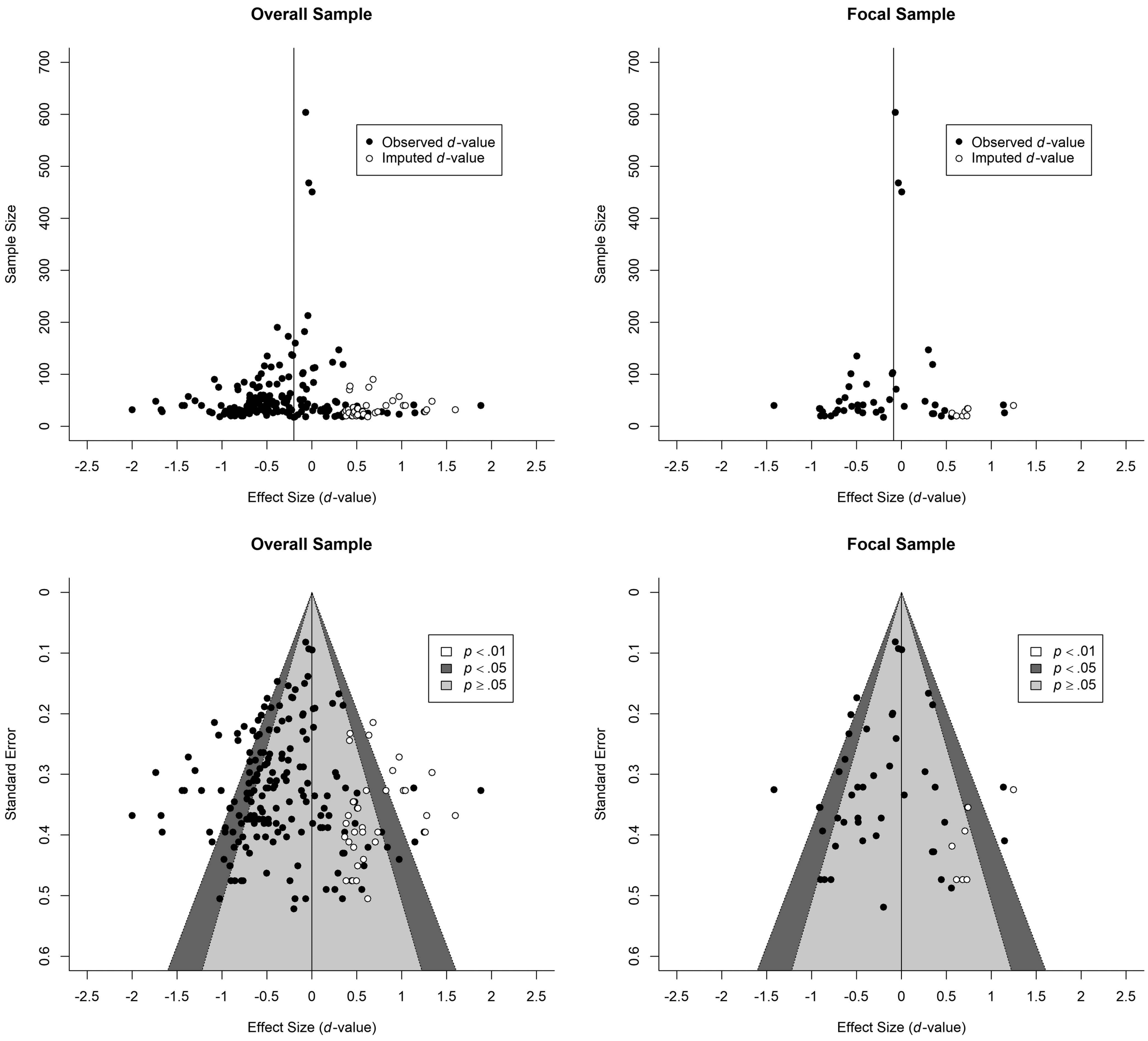

Figure 1. Sample-size (top) and contour-enhanced (bottom) funnel plots of effect sizes with missing studies imputed using trim and fill method. Contour-enhanced funnel plots show two levels of significance for all samples (α = .01 and .05). Sampling error variances for the contour-enhanced funnel plots (and the meta-analyses) were calculated using the mean effect size for all studies. Vertical reference lines for the sample size funnel plots center on the mean effect size, whereas they center on the null hypothesis of $d = 0$ for the contour-enhanced funnel plot (to evaluate statistical significance visually).

tion strategies produced a comparable threat effect as moderate activation strategies, with effects plateauing starting at moderate activation strategies. This finding adds clarity to what had previously been found to be a nonmonotonic effect, with the pattern of effects differing by threat group type (females vs. ethnic minorities). Analyses indicated that studies using the accuracy metric produced a significantly stronger threat effect than studies using the total correct metric. We also found that removal groups produced a significantly stronger threat effect than control groups. These findings highlight the importance of methodological and analytic decisions in shaping the magnitude of the stereotype threat effect. As such, the reader must be attentive to these choices in stereotype threat research.

Critically, we used the focal subset to estimate the effect of stereotype threat that may occur in high-stakes testing scenarios. After analyzing this more refined sample, we found that the stereotype threat effect was small in magnitude (mean $d = -.14$, vs. mean $d = -.31$ in the overall sample). This is still an estimate of what *can* be experienced, because most studies meta-analyzed are lab studies. However, this estimate contains testing conditions

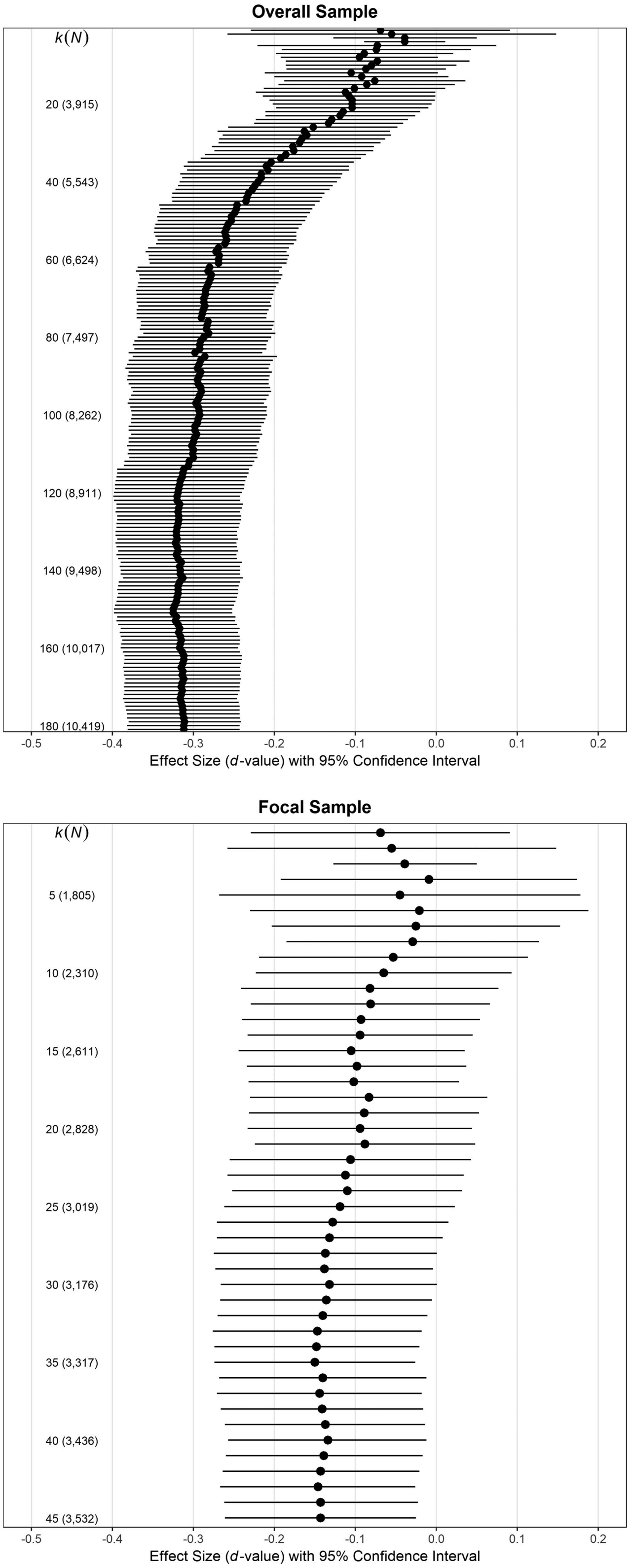

Figure 2. Cumulative meta-analyses of overall and focal samples.

that display considerably more similarity to operational testing conditions and as such, is more relevant to estimating stereotype threat's potential manifestation outside the lab.

Further, the small subset of samples from operational testing programs displayed a null threat effect (mean $d = -.01$).

To further examine the difference in findings between the small but nonzero effect of the focal sample and the zero effect of operational samples, we turned to the motivational incentive sub-

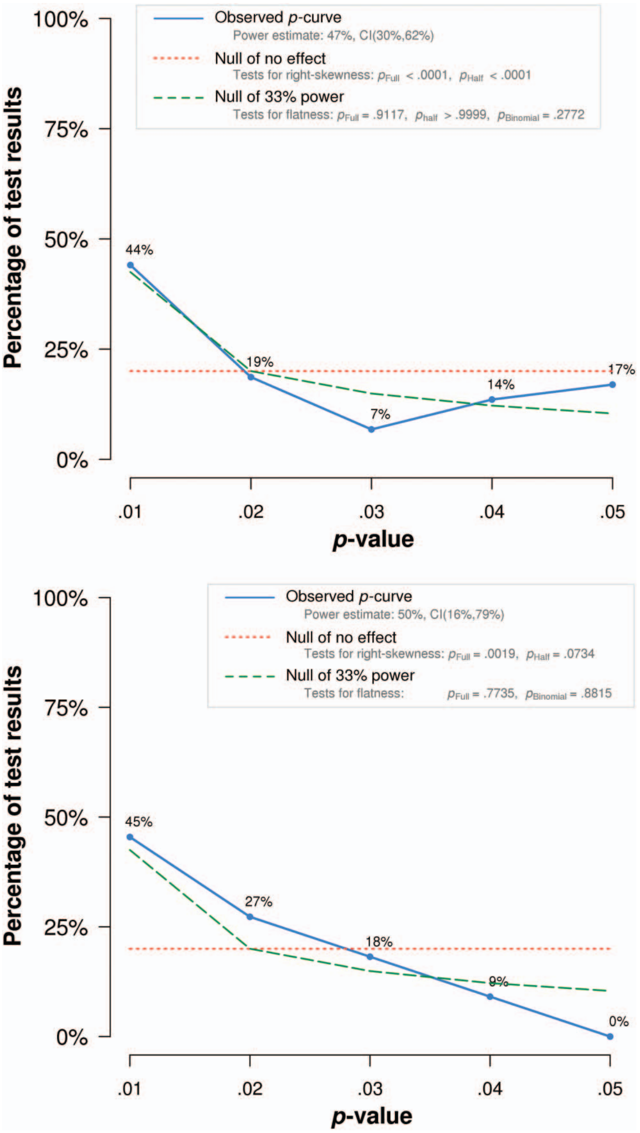

Figure 3. p-curve of overall sample (top) and focal sample (bottom). The overall sample observed p-curve includes 59 statistically significant ($p <$ .05) results, of which 39 are $p < .025$. There were 122 additional results entered but excluded from the overall sample p-curve because they were $p > .05$. The focal sample observed p-curve includes 11 statistically significant ($p < .05$) results, of which nine are $p < .025$. There were 34 additional results entered but excluded from the focal sample p-curve because they were $p > .05$. Focal Sample = Subtle activation cues, total correct (not accuracy), no covariate used, and control group (no removal groups). See the online article for the color version of this figure.

set. In these studies, effort is not necessarily maximized as would generally be expected with operational tests. However, test-taker effort is likely increased beyond a minimal level due to incentives for top performers. We found that this subset overall yielded the same magnitude effect as the focal sample ($d = -.14$), a smaller effect than for the overall mean across studies ($d = -.31$). Studies using monetary incentives displayed a null effect. These findings suggest test-taker motivation is a key mechanism toward understanding whether the threat effect will generalize outside of the laboratory. At a minimum, these findings suggest we cannot assume laboratory samples with no incentives or consequences represent maximum effort exerted analogous to high stakes testing. Overall, the pattern of results converges: removing features not relevant to operational tests, studying operational tests themselves, and providing motivational incentives all yield a reduced magnitude threat effect. The results from the focal, operational, and motivational incentive samples indicate that the size of the stereotype threat effect that can be experienced on cognitive ability tests in operational testing scenarios such as admissions tests and employment testing may range from negligible to small.

All tests of publication bias indicated that the stereotype threat effect is inflated to some degree in both the overall and focal samples. We find similar overall patterns as Flore and Wicherts (2015) grade and high school sample regarding the overall magnitude of threat, existence of publication bias, and lack of evidence for QRPs. Our findings also converge with Zigerell (2017) to suggest nontrivial publication bias is present in this literature. We believe the largest cause of publication bias is null-result-suppression in the form of failure to publish nonsignificant findings, exacerbated by sampling error in small sample studies.

Limitations

With the exception of four samples (college placement exams, Anderson, 2001; Stricker & Ward, 2004), all samples included are experiments in nonhigh stakes or operational testing scenarios. Ryan and Sackett (2013) note that this primarily lab-based literature addresses the question, "*Can this happen*?" whereas the issue of if the effect generalizes in operational settings shifts the question to, "*Does this happen*?" We attend to the shift in focus of whether stereotype threat actually occurs by examining features of stereotype threat relevant to operational testing scenarios and removing features that would not be present in these situations because of legal and ethical concerns (Ryan & Sackett, 2013).

Although we have removed testing features that are not relevant to operational testing, this does not necessarily mean that this subset of studies perfectly mimics operational testing scenarios. In other words, we have removed testing conditions that will not be present in operational testing scenarios, but removing these conditions does not imply that the focal subset is analogous to operational testing scenarios. For example, standardized admissions or employment tests are likely prone to more test-taker stress and anxiety than lab studies because of the stakes of these tests. It could be the case that subtle threat cues may evoke a larger effect when accompanied by greater test-taker anxiety. It may also be the case that subtle threat cues are more salient and more likely to be noticed by test-takers in threatened groups in high stakes settings. For these reasons, we sought to run moderator analyses isolating both operational studies and studies with features likely to motivate participants greater than baseline motivation for an often-voluntary experimental task. In these subsets, the effect was negligible to small, but there were an admittedly small number of samples available for these analyses.

We also sought to explain the difference between the null effect found in operational samples versus the nonzero effect found in lab studies. We accomplished this by examining lab studies that included motivational incentives rewarding participants based on ability test performance, finding the threat effect to be either reduced substantially or to a null effect depending on the subset. Although these findings indicate lack of test-taker motivation may present a major concern when generalizing stereotype threat outside of the lab, we have not explicitly modeled the effect that motivation has on stereotype threat. In relation to these small subsets, we discuss directions for future research below.

The degree of heterogeneity across moderator analyses was substantial. Even when subsetting the focal sample by four moderators, there was sizable heterogeneity. This suggests there is a good deal of variability in the true effect of stereotype threat, and that any firm conclusions regarding the existence of stereotype threat depend on multiple factors.

Future Research Directions

We believe that research on stereotype threat must place a higher priority on external validity with respect to nonlab, high stakes testing settings. Of 200-plus samples meeting inclusion criteria, more than 75% of studies were identified as containing at least one major feature that would not be present in operational testing settings. Fewer than 10% of studies contained a motivational incentive that would serve to increase test-taker motivation beyond a minimal level. Research often utilizes participants who are identified in the domain to ensure that participants are invested in the subject matter. Yet, domain identification produces a nonsignificant effect on threat, and domain identification is not necessarily a proxy for motivation because there are not incentives to create extrinsic motivation to perform on the test. Intrinsic motivation cannot be assumed, even for those who care about the domain. We acknowledge that in some cases experiments on stereotype threat do not seek to generalize to tests such as admissions or employment tests. However, as stereotype threat is often invoked as an explanation for group differences on tests, generalization to these settings is a critical concern. Additionally, only one in 10 studies had a sample size of more than 100. Given recent large-scale failures to replicate numerous effects in social psychology (Open Science Collaboration, 2015), statistical power and transparent research practices (cf., Nosek et al., 2015) are essential when evaluating reproducibility of this effect and combating publication bias.

For these reasons, we believe future research on stereotype threat should (a) place a greater emphasis on modeling and manipulating motivation, (b) when ethical and feasible, examine operational tests in experimental or quasi-experimental settings, and (c) use large samples informed by power analyses and use preregistration databases to allow for more stable estimation of effects and more visible studies. One specific avenue for future research would be to subtly prime stereotype threat across varying levels of motivational incentives (e.g., financial incentives to perform well on an ability test) to observe the impact on test performance. Although we observe the threat effect to be negligible in our subset of financial incentive studies, manipulating motivation within the same experimental context would provide a more direct, head-to-head comparison (i.e., Sackett, Shewach, & Keiser, 2017) of the influence of motivation on stereotype threat. Future research would also benefit from disentangling the effect of financial incentives versus the effect of an applicant prime on stereotype threat, as the applicant prime studies were fully nested within the financial incentives studies. Additionally, gamification represents a future avenue of research for stereotype threat; it is becoming widespread across employee selection contexts (Armstrong, Landers, & Collmus, 2015) and gender stereotypes exist in aspects of gamification.

Implications

We believe that our study has clear implications for researchers and for those attempting to influence public policy. We offer multiple prescriptions. First, although use of a pretest can be useful for increasing statistical power, it is critical to avoid the error we found in multiple prior studies of using a pretest-adjusted standard deviation in the computation of effect size measures to index stereotype threat. Second, if one is interested in the effects of threat on cognitive ability test scores, it is crucial to avoid the use of a "proportion correct among items attempted" scoring method, as such a method is rarely, if ever, to our knowledge, used in operational testing. The use of this scoring method produces larger threat effects than are observed using realistic scoring methods. Third, we recommend the inclusion of a control group in stereotype threat experiments. Researchers may also be interested in a threat removal condition, but including a control group in addition to threat induction and threat removal conditions permits clearer interpretation of findings. Fourth, we suggest attention to the nature of a threat induction mechanism in the design of studies, as our results show markedly smaller threat effects in studies using the types of subtle induction mechanisms that are likely to be present in operational testing settings. Researchers certainly may choose to study more blatant induction mechanisms in research on basic mechanisms underlying threat effects, but research aimed at insight into threat effects in operational settings should use realistic induction mechanisms. Fifth, researchers should aim for larger sample sizes than are typically seen in threat research. In light of the systematic relationship we observed between sample size and effect size in the threat literature, small sample studies merit skepticism. Sixth, threat researchers need to attend to issues of participant motivation. We find much smaller threat effects in studies offering an incentive for devoting effort to the test. Note that the theoretical mechanism behind stereotype threat is that threat detracts from attentional resources, resulting in lower test scores. Absent motivation to devote effort to a test, a reasonable reaction to induced threat is to exert minimal effort to the test and exit the situation, and thus reduced motivation offers an alternate explanation for lower test performance in threatening situations. Seventh, we encourage research in operational testing settings. Only four of the 212 studies we located were in applied settings, and these produced negligible threat effects. To the extent that the field wishes to offer prescriptions about the effects of threat in operational settings, efforts to study operational testing programs are critical.

Based on the results of the focal analysis, operational and motivational subsets, and publication bias analyses, we conclude that the burden of proof shifts back to those that claim that stereotype threat exerts a substantial effect on standardized test-takers. Our best estimate of stereotype threat effects within groups in settings with conditions most similar to operational testing is small and inflated by publication bias. Furthermore, estimates of threat in situations more likely to include motivated participants (i.e., operational test samples; motivational incentive samples) range from negligible to small. We do not discount the notion that small effects, compounded over many individuals and across time, can yield substantial consequences. Yet, the small effect of stereotype threat we observe in this meta-analysis still represents an effect that *can* (but not necessarily *does*) occur in stereotyped test-takers. This evidence indicates that any claims of sizable threat effects on standardized tests must be substantiated in light of these findings and with high stakes test-like conditions.

References

References marked with an asterisk indicate sample contributed an effect size to the meta-analysis.

*Adams, G., Garcia, D. M., Purdie-Vaughns, V., & Steele, C. M. (2006). The detrimental effects of a suggestion of sexism in an instruction situation. *Journal of Experimental Social Psychology, 42,* 602–615. http://dx.doi.org/10.1016/j.jesp.2005.10.004

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

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

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

Armstrong, M. B., Landers, R. N., & Collmus, A. B. (2015). Gamifying recruitment, selection, training, and performance management: Game-thinking in human resource management. In H. Gangadharbatla & D. Z. Davis (Eds.), *Emerging research and trends in gamification* (pp. 140–165). Hershey, PA: IGI Global.

Aronson, J., Dweck, C. S., Erman, S., Good, C., Inzlicht, M., Logel, C., . . . Yeager, D. (2015). *Brief of experimental psychologists as amici curiae in support of respondents*. Fisher v. University of Texas at Austin, 579 US__(2016).

*Aronson, J., Lustina, 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. http://dx.doi.org/10.1006/jesp.1998.1371

*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*. Georgia Institute of Technology. Retrieved from https://smartech.gatech.edu/handle/1853/4942

Banks, G. C., Kepes, S., & McDaniel, M. A. (2015). Publication bias: Understanding the myths concerning threats to the advancement of science. In C. E. Lance & R. J. Vandenberg (Eds.), *More statistical and methodological myths and urban legends* (pp. 36–64). New York, NY: Routledge.

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

*Ben-Zeev, T., Fein, S., & Inzlicht, M. (2005). Arousal and stereotype threat. *Journal of Experimental Social Psychology, 41,* 174–181. http://dx.doi.org/10.1016/j.jesp.2003.11.007

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley LTD. http://dx.doi.org/10.1002/9780470743386

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

*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. http://dx.doi.org/10.1037/0021-9010.91.4.979

*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. http://dx.doi.org/10.1037/0022-3514.76.2.246

*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. http://dx.doi.org/10.1002/ejsp.145

*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. http://dx.doi.org/10.1111/j.0956-7976.2005.01577.x

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

*Chalabaev, A., Radel, R., Masicampo, E. J., & Dru, V. (2016). Reducing stereotype threat with embodied triggers: A case of sensorimotor-mental congruence. *Personality and Social Psychology Bulletin, 42,* 1063–1076. http://dx.doi.org/10.1177/0146167216651407

*Clark, J. K., Eno, C. A., & Guadagno, R. E. (2011). Southern Discomfort: The Effects of Stereotype Threat on the Intellectual Performance of U.S. Southerners. *Self and Identity, 10,* 248–262. http://dx.doi.org/10.1080/15298861003771080

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

*Coleman, J. M., & Harr, V. (2007). Group disidentification in response to stereotype threat: When math is hard, is it hard to identify with women? Unpublished manuscript.

*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, New York, NY.

Coyle, T. R. (2007). Test-retest changes on scholastic aptitude tests are not related to *g*. *Intelligence, 34,* 15–27. http://dx.doi.org/10.1016/j.intell.2005.04.001

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

Dahlke, J. A., & Wiernik, B. M. (2018). psychmeta: An R package for psychometric meta-analysis. *Applied Psychological Measurement*. Advance online publication. http://dx.doi.org/10.1177/0146621618795933

*Dar-Nimrod, I., & Heine, S. J. (2006). Exposure to scientific theories affects women's math performance. *Science, 314,* 435. http://dx.doi.org/10.1126/science.1131100

*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. http://dx.doi.org/10.1177/014616702237644

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences of the United States of America, 108,* 7716–7720. http://dx.doi.org/10.1073/pnas.1018601108

*Duncan, L. (2005, January). *Stereotype threat and women's performance on a mental rotation task*. Paper presented at the 6th annual convention for Personality and Social Psychology, New Orleans, LA.

Duval, S. (2005). The trim and fill method. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments* (pp. 127–144). Chichester, England: Wiley. http://dx.doi.org/10.1002/0470870168.ch8

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56,* 455–463. http://dx.doi.org/10.1111/j.0006-341X.2000.00455.x

*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, TX A&M University, College Station, TX.

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

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

*Finnigan, K. M., & Corker, K. S. (2016). Do performance avoidance goals moderate the effect of different types of stereotype threat on women's math performance? *Journal of Research in Personality, 63,* 36–43. http://dx.doi.org/10.1016/j.jrp.2016.05.009

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

*Fogliati, V. J., & Bussey, K. (2013). Stereotype threat reduces motivation to improve: Effects of stereotype threat and feedback on women's intentions to improve mathematical ability. *Psychology of Women Quarterly, 37,* 310–324. http://dx.doi.org/10.1177/0361684313480045

*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. http://dx.doi.org/10.1177/0146167203262851

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

Gervias, W. (2015). *Putting PET-PEESE to the test* [Web log post]. Retrieved from http://web.archive.org/web/20160120140336/http://willgervais.com/blog/2015/6/25/putting-pet-peese-to-the-test-1

Gervias, W. (2016). *Enough heavy PETting*. Retrieved from http://willgervais.com/blog/2016/3/3/enough-heavy-petting

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

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

*Grand, J. A., Ryan, A. M., Schmitt, N., & Hmurovic, J. (2011). How far does stereotype threat reach? The potential detriment of face validity in cognitive ability testing. *Human Performance, 24,* 1–28. http://dx.doi.org/10.1080/08959285.2010.518184

*Gresky, D. M., Ten Eyck, L. L., Lord, C. G., & McIntyre, R. B. (2005). Effects of salient multiple identities on women's performance under mathematics stereotype threat. *Sex Roles, 53,* 703–716. http://dx.doi.org/10.1007/s11199-005-7735-2

*Guajardo, G. A. (2005). *Modifying stereotype relevance and altering affect attributions to reduce performance suppression on cognitive ability selection tests*. Unpublished master's thesis, Northern Illinois University, DeKalb, IL.

*Halim, M. L., Aronson, J., & Amodio, D. M. (2010, January). *Stereotype threat effects on African Americans' implicit academic self-concepts*. Paper presented at the 11th annual meeting for the Society for Personality and Social Psychology, Las Vegas, NV.

*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.

*Henderson, J. M. (2017). *Black racial identity protecting against stereotype threat on collegiate academic achievement*. New York, NY: Fordham University. Retrieved from http://search.proquest.com/openview/78a4a1888a03ef3d71bf594b5e768f77/1?pq-origsite=gscholar&cbl=18750&diss=y

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment, 9,* 152–194. http://dx.doi.org/10.1111/1468-2389.00171

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis*. Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781412985031

*Inzlicht, M. (2001). *A threatening intellectual environment: When and why females are susceptible to experiencing problem-solving deficits in the presence of males*. Unpublished doctoral dissertation, Brown University, Providence, Rhode Island.

*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. http://dx.doi.org/10.1111/1467-9280.00272

*Inzlicht, M., & Ben-Zeev, T. (2003). Do High-Achieving Female Students Underperform in Private? The Implications of Threatening Environments on Intellectual Processing. *Journal of Educational Psychology, 95,* 796–805. http://dx.doi.org/10.1037/0022-0663.95.4.796

Inzlicht, M., Gervias, W., & Berkman, E. (2015). *Bias-correction techniques alone cannot determine whether ego depletion is different from zero: Commentary on Carter, Kofler, Forster, & McCullough, 2015*. Unpublished manuscript. http://dx.doi.org/10.2139/ssrn.2659409

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

Ioannidis, J. P., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials, 4,* 245–253. http://dx.doi.org/10.1177/1740774507079441

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

*Jamieson, J. P., & Harkins, S. G. (2010). Evaluation is necessary to produce stereotype threat performance effects. *Social Influence, 5,* 75–86. http://dx.doi.org/10.1080/15534510903512409

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

*John-Henderson, N. A., Rheinschmidt, M. L., & Mendoza-Denton, R. (2015). Cytokine responses and math performance: The role of stereotype threat and anxiety reappraisals. *Journal of Experimental Social Psychology, 56,* 203–206. http://dx.doi.org/10.1016/j.jesp.2014.10.002

*Johns, M., Inzlicht, M., & Schmader, T. (2008). Stereotype threat and executive resource depletion: Examining the influence of emotion regulation. *Journal of Experimental Psychology: General, 137,* 691–705. http://dx.doi.org/10.1037/a0013834

*Johnson, H. J., Barnard-Brak, L., Saxon, T. F., & Johnson, M. K. (2012). An experimental study of the effects of stereotype threat and stereotype lift on men and women's performance in mathematics. *Journal of Experimental Education, 80,* 137–149. http://dx.doi.org/10.1080/00220973.2011.567312

*Jones, P. R. (2011). Reducing the impact of stereotype threat on women's math performance: Are two strategies better than one? *Electronic Journal of Research in Educational Psychology, 9,* 587–616.

*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. http://dx.doi.org/10.1111/1467-9280.t01-1-01435

*Joy-Gaba, J. A., Nosek, B. A., & Schmader, T. (2007). *Expectations and test relevance as potential moderators of stereotype threat*. Paper presented at the 8th annual meeting for the Society for Personality and Social Psychology, Memphis, TN.

*Keller, J., & Bless, H. (2008). When positive and negative expectancies disrupt performance: Regulatory focus as a catalyst. *European Journal of Social Psychology, 38,* 187–212. http://dx.doi.org/10.1002/ejsp.452

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

Kepes, S., Banks, G. C., McDaniel, M., & Whetzel, D. L. (2012). Publication bias in the organizational sciences. *Organizational Research Methods, 15,* 624–662. http://dx.doi.org/10.1177/1094428112452760

*Kiefer, A. K., & Sekaquaptewa, D. (2007). Implicit stereotypes and women's math performance: How implicit gender-math stereotypes influence women's susceptibility to stereotype threat. *Journal of Experimental Social Psychology, 43,* 825–832. http://dx.doi.org/10.1016/j.jesp.2006.08.004

*Kirnan, J. P., Alfieri, J. A., Bragger, J. D., & Harris, R. S. (2009). An investigation of stereotype threat in employment tests. *Journal of Applied Social Psychology, 39,* 359–388. http://dx.doi.org/10.1111/j.1559-1816.2008.00442.x

Krendl, A. C., Richeson, J. A., Kelley, W. M., & Heatherton, T. F. (2008). The negative consequences of threat: A functional magnetic resonance imaging investigation of the neural mechanisms underlying women's underperformance in math. *Psychological Science, 19,* 168–175. http://dx.doi.org/10.1111/j.1467-9280.2008.02063.x

Lamont, R. A., Swift, H. J., & Abrams, D. (2015). A review and meta-analysis of age-based stereotype threat: Negative stereotypes, not facts, do the damage. *Psychology and Aging, 30,* 180–193. http://dx.doi.org/10.1037/a0038586

Lee, J. E. R., & Nass, C. (2012). Distinctiveness-based stereotype threat and the moderating role of coaction contexts. *Journal of Experimental Social Psychology, 48,* 192–199. http://dx.doi.org/10.1016/j.jesp.2011.06.018

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

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

*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. http://dx.doi.org/10.1016/j.jesp.2005.04.010

*Martin, D. E. (2003). *Stereotype threat, cognitive aptitude measures, and social identity*. Unpublished doctoral dissertation, Howard University.

*Marx, D. M., & Goff, P. A. (2005). Clearing the air: The effect of experimenter race on target's test performance and subjective experience. *British Journal of Social Psychology, 44,* 645–657. http://dx.doi.org/10.1348/014466604X17948

*Marx, D. M., & Ko, S. J. (2012). Superstars "like" me: The effect of role model similarity on performance under threat. *European Journal of Social Psychology, 42,* 807–812. http://dx.doi.org/10.1002/ejsp.1907

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

Marx, D. M., & Stapel, D. A. (2006a). Distinguishing stereotype threat from priming effects: On the role of the social self and threat-based concerns. *Journal of Personality and Social Psychology, 91,* 243–254. http://dx.doi.org/10.1037/0022-3514.91.2.243

Marx, D. M., & Stapel, D. A. (2006b). Retracted: It's all in the timing: Measuring emotional reactions to stereotype threat before and after taking a test. *European Journal of Social Psychology, 36,* 687–698. http://dx.doi.org/10.1002/ejsp.310

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. http://dx.doi.org/10.1037/0022-3514.88.3.432

*McGlone, M. S., & Aronson, J. (2006). Stereotype threat, identity salience, and spatial reasoning. *Journal of Applied Developmental Psychology, 27,* 486–493. http://dx.doi.org/10.1016/j.appdev.2006.06.003

*McIntyre, R. B., Lord, C. G., Gresky, D. M., Ten Eyck, L. L., Frye, G. 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 Social 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. http://dx.doi.org/10.1016/S0022-1031(02)00513-9

*McIntyre, R. B., Paulson, R. M., Taylor, C. A., Morin, A. L., & Lord, C. G. (2011). Effects of role model deservingness on overcoming performance deficits induced by stereotype threat. *European Journal of Social Psychology, 41,* 301–311. http://dx.doi.org/10.1002/ejsp.774

*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, Akron, OH.

Moreno, S. G., Sutton, A. J., Ades, A. E., Stanley, T. D., Abrams, K. R., Peters, J. L., & Cooper, N. J. (2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology, 9,* 2. http://dx.doi.org/10.1186/1471-2288-9-2

*Nadler, D. R., & Komarraju, M. (2016). Negating stereotype threat: Autonomy support and academic identification boost performance of African American college students. *Journal of College Student Development, 57,* 667–679. http://dx.doi.org/10.1353/csd.2016.0039

Neter, J., Wasserman, W., & Whitmore, G. A. (1988). *Applied Statistics* (3rd ed.). Newton, MA: Allyn & Bacon.

*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,* 261–293. http://dx.doi.org/10.1207/S15327043HUP1603_5

Nguyen, H.-H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology, 93,* 1314–1334. http://dx.doi.org/10.1037/a0012702

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., . . . Yarkoni, T. (2015). SCIENTIFIC STANDARDS. Promoting an open research culture. *Science, 348,* 1422–1425. http://dx.doi.org/10.1126/science.aab2374

*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. http://dx.doi.org/10.1177/0146167203029006010

Open Science Collaboration. (2015). PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science, 349,* aac4716. http://dx.doi.org/10.1126/science.aac4716

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

*Pennington, C. R., & Heim, D. (2016). Creating a critical mass eliminates the effects of stereotype threat on women's mathematical performance. *British Journal of Educational Psychology, 86,* 353–368. http://dx.doi.org/10.1111/bjep.12110

*Perry, S. P., & Skitka, L. J. (2009). Making lemonade? Defensive coping style moderates the effect of stereotype threat on women's math test performance. *Journal of Research in Personality, 43,* 918–920. http://dx.doi.org/10.1016/j.jrp.2009.05.013

*Philipp, M. C., & Harton, H. C. (2004, January). *The role of social dominance in stereotype threat effects*. Paper presented at the 5th annual convention for Personality and Social Psychology, Austin, TX.

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

*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. http://dx.doi.org/10.1207/S15327043HUP1603_4

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

*Rosenthal, H. E. S., Crisp, R. J., & Suen, M.-W. (2007). Improving performance expectancies in stereotypic domains: Task relevance and the reduction of stereotype threat. *European Journal of Social Psychology, 37,* 586–597. http://dx.doi.org/10.1002/ejsp.379

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2006). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Hoboken, NJ: Wiley.

*Rucks, L. J. (2005). *The generality of misattributing arousal while under stereotype threat*. Unpublished master's thesis, Ohio State University, Columbus, OH.

*Rucks, L. J. (2008). *Me, women, and math: The role of personal and collective threats in the experience of stereotype threat*. Columbus, OH: The Ohio State University. Retrieved from http://rave.ohiolink.edu/etdc/view?acc_num=osu1204661976

Ryan, A. M., & Nguyen, H. D. (2017). Publication bias and stereotype threat research: A reply to Zigerell. *Journal of Applied Psychology, 102,* 1169–1177. http://dx.doi.org/10.1037/apl0000242

Ryan, A. M., & Sackett, P. R. (2013). Stereotype threat in workplace assessments. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology: Vol. 1. Test theory and testing and assessment in industrial and organizational psychology* (pp. 661–673). Washington, DC: American Psychological Association.

*Rydell, R. J., & Boucher, K. L. (2010). Capitalizing on multiple social identities to prevent stereotype threat: The moderating role of self-esteem. *Personality and Social Psychology Bulletin, 36,* 239–250. http://dx.doi.org/10.1177/0146167209355062

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

*Rydell, R. J., Van Loo, K. J., & Boucher, K. L. (2014). Stereotype threat and executive functions: Which functions mediate different threat-related outcomes? *Personality and Social Psychology Bulletin, 40,* 377–390. http://dx.doi.org/10.1177/0146167213513475

Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2004). On interpreting stereotype threat as accounting for African American-White differences on cognitive tests. *American Psychologist, 59,* 7–13. http://dx.doi.org/10.1037/0003-066X.59.1.7

Sackett, P. R., & Ryan, A. M. (2012). Concerns about generalizing stereotype threat research findings to operational high-stakes testing. In M. Inzlicht & T. Schmader (Eds.), *Stereotype threat: Theory, process, and application* (pp. 249–263). New York, NY: Oxford University Press.

Sackett, P. R., Shewach, O. R., & Keiser, H. N. (2017). Assessment centers versus cognitive ability tests: Challenging the conventional wisdom on criterion-related validity. *Journal of Applied Psychology, 102,* 1435–1447. http://dx.doi.org/10.1037/apl0000236

*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, Austin, TX.

*Saunders, B. A. (2008, February). *Stereotype threat and academic disengagement: The role of self-worth contingencies*. Paper presented at the 9th annual meeting for the Society for Personality and Social Psychology, Albuquerque, NM.

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

*Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology, 85,* 440–452. http://dx.doi.org/10.1037/0022-3514.85.3.440

*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.

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

*Seibt, B., & Förster, J. (2004). Stereotype threat and performance: How self-stereotypes influence processing by inducing regulatory foci. *Journal of Personality and Social Psychology, 87,* 38–56. http://dx.doi.org/10.1037/0022-3514.87.1.38

Seitchik, A. E. (2013). *Stereotype threat, mental arithmetic, and the mere effort account*. Unpublished doctoral dissertation, Northeastern University, Boston, MA.

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

*Shaffer, E. S., Marx, D. M., & Prislin, R. (2013). Mind the gap: Framing of women's success and representation in STEM affects women's math performance under threat. *Sex Roles, 68,* 454–463. http://dx.doi.org/10.1007/s11199-012-0252-1

*Shapiro, J. R., Williams, A. M., & Hambarchyan, M. (2013). Are all interventions created equal? A multi-threat approach to tailoring stereotype threat interventions. *Journal of Personality and Social Psychology, 104,* 277–288. http://dx.doi.org/10.1037/a0030461

Shih, M., Pittinsky, T. L., & Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological Science, 10,* 80–83. http://dx.doi.org/10.1111/1467-9280.00111

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143,* 534–547. http://dx.doi.org/10.1037/a0033242

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). p-Curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science, 9,* 666–681. http://dx.doi.org/10.1177/1745691614553988

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). *P-curve.com*. Retrieved from http://www.p-curve.com/

Smeding, A., Dumas, F., Loose, F., & Régner, I. (2013). Order of administration of math and verbal tests: An ecological intervention to reduce stereotype threat on girls' math performance. *Journal of Educational Psychology, 105,* 850–860. http://dx.doi.org/10.1037/a0032094

*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. *The Western Journal of Black Studies, 28,* 312.

*Smith, J. L., & Chase, J. P. (2012). *Comparing interventions that negate stereotype threat effects on women's mathematical performance and STEM motivation*. Unpublished manuscript.

*Smith, L. G. E., & Postmes, T. (2011). Shaping stereotypical behaviour through the discussion of social stereotypes. *British Journal of Social Psychology, 50,* 74–98. http://dx.doi.org/10.1348/014466610X500340

*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, Brunswick, NJ.

Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype threat. *Annual Review of Psychology, 67,* 415–437. http://dx.doi.org/10.1146/annurev-psych-073115-103235

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

*Ståhl, T., Van Laar, C., & Ellemers, N. (2012). The role of prevention focus under stereotype threat: Initial cognitive mobilization is followed by depletion. *Journal of Personality and Social Psychology, 102,* 1239–1251. http://dx.doi.org/10.1037/a0027678

Steele, C. M. (2010). *Whistling Vivaldi and other clues to how stereotypes affect us*. New York, NY: Norton.

*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. http://dx.doi.org/10.1037/0022-3514.69.5.797

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. *Advances in Experimental Social Psychology, 34,* 379–440. http://dx.doi.org/10.1016/S0065-2601(02)80009-0

*Steinberg, J. R. (2008). *Stereotype threat in persistent women*. Unpublished doctoral dissertation, Arizona State University, Tempe, AZ.

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

*Stricker, L. J., & Ward, W. C. (2004). Stereotype threat, inquiring about test takers' ethnicity and gender, and standardized test performance. *Journal of Applied Social Psychology, 34,* 665–693. http://dx.doi.org/10.1111/j.1559-1816.2004.tb02564.x

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

Taylor, V. J., & Walton, G. M. (2011). Stereotype threat undermines academic learning. *Personality and Social Psychology Bulletin, 37,* 1055–1067. http://dx.doi.org/10.1177/0146167211406506

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine, 22,* 2113–2126. http://dx.doi.org/10.1002/sim.1461

*Van Loo, K. J., & Rydell, R. J. (2013). On the experience of feeling powerful: Perceived power moderates the effect of stereotype threat on women's math performance. *Personality and Social Psychology Bulletin, 39,* 387–400. http://dx.doi.org/10.1177/0146167212475320

Viechtbauer, W. (2017). *Package 'metafor.' R package version 2.0–0.*

*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. http://dx.doi.org/10.1037/0022-3514.89.1.22

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

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

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

*Wen, F., Zuo, B., Wu, Y., Dong, X., & Wang, W. (2016). Reducing the effect of stereotype threat: The role of coaction contexts and regulatory fit. *Social Psychology of Education, 19,* 607–626. http://dx.doi.org/10.1007/s11218-016-9344-z

*Werhun, C. D. (2007). *The limitations of stereotype threat: Not all math and science women are threatened by stereotypes*. Unpublished doctoral dissertation, University of Toronto, Toronto, Ontario, Canada.

*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. http://dx.doi.org/10.1037/0022-3514.89.5.696

*Wout, D. A. (2007). *The effects of stereotypes and individuating information on black students' susceptibility to stereotype threat*. Unpublished doctoral dissertation, University of Michigan, Ann Arbor, MI.

*Yopyk, D. A. (2005). *Social rejection as a mediating variable in the link between stereotype threat and math performance*. Retrieved from http://scholarworks.umass.edu/cgi/viewcontent.cgi?article=3563&context=theses

Zigerell, L. J. (2017). Potential publication bias in the stereotype threat literature: Comment on Nguyen and Ryan (2008). *Journal of Applied Psychology, 102,* 1159–1168. http://dx.doi.org/10.1037/apl0000188

Received February 11, 2018
Revision received February 16, 2019
Accepted April 8, 2019 ■