A JOURNAL OF THE ASSOCIATION FOR PSYCHOLOGICAL SCIENCE
aps

Research Article

Latent Ability

Grades and Test Scores Systematically Underestimate the Intellectual Ability of Negatively Stereotyped Students

Gregory M. Walton[1] **and Steven J. Spencer**[2]

[1]*Stanford University and* [2]*University of Waterloo*

ABSTRACT—*Past research has assumed that group differences in academic performance entirely reflect genuine differences in ability. In contrast, extending research on stereotype threat, we suggest that standard measures of academic performance are biased against non-Asian ethnic minorities and against women in quantitative fields. This bias results not from the content of performance measures, but from the context in which they are assessed—from psychological threats in common academic environments, which depress the performances of people targeted by negative intellectual stereotypes. Like the time of a track star running into a stiff headwind, such performances underestimate the true ability of stereotyped students. Two meta-analyses, combining data from 18,976 students in five countries, tested this* latent-ability *hypothesis. Both meta-analyses found that, under conditions that reduce psychological threat, stereotyped students performed better than nonstereotyped students at the same level of past performance. We discuss implications for the interpretation of and remedies for achievement gaps.*

A great and persistent problem in education involves the existence of large group differences in academic performance, such as between non-Asian ethnic minority and nonminority students (Jencks & Phillips, 1998) and between women and men in quantitative fields (Hyde, Fennema, & Lamon, 1990). Standard accounts assume that group differences entirely reflect genuine differences in ability, and point to long-standing causes like poverty (Fryer & Levitt, 2004), gender roles (Eccles, 1987), or hypothesized innate differences (Benbow & Stanley, 1980; Herrnstein & Murray, 1994). In contrast, extending research on stereotype threat (Steele, Spencer, & Aronson, 2002), we argue that at least a portion of group differences is illusory—that this portion results from pervasive psychological threats in academic environments, which undermine the performances of ethnic minority students and of women. If so, such measures are biased: They underestimate the true ability of ethnic minorities and of women relative to nonminorities and to men.

The hypothesized bias may occur even when measures of academic performance are equally predictive for different groups (Jensen, 1980). Instead, the bias takes the form of a difference in the relative level of performance (in regression terms, the bias involves the intercept, not the slope). This analysis yields a specific prediction: If psychological threat is removed, ethnic minority students and women should perform better than nonminority students and men at the same level of prior performance. This hypothesis has never been directly tested. We test it systematically in two meta-analyses, which summarize across diverse testing conditions and provide a high level of statistical power. If our argument is correct, group differences result in part from a failure to validly measure the ability of ethnic minority students and women.

Much past research has examined the impact of psychological threat on intellectual performance in laboratory circumstances (Steele et al., 2002). This psychological threat may be present even in generally egalitarian settings where students of different groups are treated well and where test questions are culturally nonbiased. It arises, instead, from individuals' awareness of widely known negative stereotypes and the possibility that they could be seen in light of them. When ethnic minority students perform in school, or when women perform in quantitative fields, they are often aware of stereotypes that impugn the ability of their ethnic or gender group. They may worry that a poor performance could lend credence to the stereotype. Hundreds of laboratory experiments demonstrate that this experience, termed *stereotype threat*, undermines intellectual performance (Steele et al., 2002). In a seminal series of studies, African American students performed less well than European American students on a Graduate Record Exam (GRE) test described

Address correspondence to Gregory M. Walton, Department of Psychology, Stanford University, 450 Serra Mall, Stanford, CA 94305, e-mail: gwalton@stanford.edu.

Psychological Science
RESEARCH, THEORY, & APPLICATION IN PSYCHOLOGY AND RELATED SCIENCES

Copyright © 2009 Association for Psychological Science

as evaluative of verbal ability, an arena in which African Americans are negatively stereotyped. However, when the same test was described as nonevaluative, African Americans performed as well as European Americans (controlling for SAT score; Steele & Aronson, 1995). Stereotype threat undermines performance by creating distraction and taking up needed executive resources (Schmader, Johns, & Forbes, 2008). It also induces people to adopt a *prevention focus*, or to be overly concerned with not confirming negative stereotypes (Seibt & Förster, 2004).

Despite important exceptions, past research has not systematically tested whether or to what extent stereotype threat undermines performance in real-world settings (cf. Cohen, Garcia, Apfel, & Master, 2006; Danaher & Crandall, 2008; Massey & Fischer, 2005; Steele, 1997; Walton & Cohen, 2007). We suggest that it does, and provide a critical test of this hypothesis. If stereotype threat undermines stereotyped students' real-world performance, then their performance will underestimate their true intellectual ability, and their potential in environments without threat. Their performance would be like the time of a track star running into a stiff headwind: It underestimates her time without the headwind. If this is the case, then, in environments in which threat has been reduced, stereotyped students should perform better than nonstereotyped students at the same level of past performance. In part, the ability of stereotyped students would be *latent*—underestimated by their level of prior performance (for our definition of "latent," see Supporting Details in the Supporting Information available on-line; see p. 1139; cf. Wicherts, Dolan, & Hessen, 2005). If, instead, the prior measure was nonbiased, then reducing threat should result in the same level of performance on the part of stereotyped and nonstereotyped students at the same level of past performance.

The latent-ability hypothesis directly addresses an important question about stereotype threat raised by Sackett, Hardison, and Cullen (2004; see also Sackett, Borneman, & Connelly, 2008). These scholars suggest that real-world measures are nonbiased and that stereotype-threat effects are restricted to the laboratory. According to this hypothesis, in threat conditions, stereotyped students should perform worse than would be expected on the basis of their prior performance, and, in nonthreat conditions, they should not perform better. The critical test Sackett and colleagues propose to distinguish this hypothesis from the hypothesis that stereotype threat undermines real-world performance is exactly the test of latent ability described above: In their terminology, does the real-world academic performance of stereotyped students "underpredict" their performance in a less threatening environment?

Past research has not tested the latent-ability hypothesis. Much past research is correlational, comparing students' level of performance on two measures (e.g., SAT scores and college grades; Cullen, Hardison, & Sackett, 2004). Because such work does not remove psychological threat, it tests only whether, in control conditions, one measure is completed in more threatening circumstances and is therefore more biased than the other. Other past research constitutes individual experiments, which emphasize stereotyped students' level of performance in different conditions (e.g., Spencer, Steele, & Quinn, 1999; Steele & Aronson, 1995). Individually, such studies lack statistical power to test the hypothesized group difference. By meta-analyzing across many such studies, we test the hypothesized bias with considerable power and, as noted, in diverse testing conditions.

A consideration of the relative level of psychological threat in different environments may also shed light on the converse effect—*underperformance*. Often ethnic minority students perform worse than nonminority students at the same level of prior performance (Jensen, 1980). The cause of this effect is not fully understood. One possibility is that, absent intervention, the level of psychological threat increases at each rung of the educational ladder, for instance as students become more anonymous (Walton & Cohen, 2007) and as they reach the edge of their abilities (Steele, 1997). If so, underperformance should be evident only when threat has not been removed (i.e., in "threat" or "control" conditions).

META-ANALYSIS I: STEREOTYPE-THREAT EXPERIMENTS

As an initial test of the latent-ability and underperformance hypotheses, we conducted a meta-analysis of stereotype-threat experiments. Each study manipulated the presence or the absence of stereotype threat ("threat condition" and "safe condition," respectively), assessed the performance of stereotyped and nonstereotyped students on a test relevant to the stereotype, and assessed students' level of performance in the same domain along an independent real-world measure. By assessing performance in situations designed to heighten and to reduce threat, the studies test for both underperformance and latent ability.

Method

Retrieval of Studies and Inclusion Criteria

To retrieve studies, we searched the PsychINFO and ERIC databases (keyword: "stereotype threat"; date: November 2007), e-mailed professional listservs (SPSP, SPSSI, and SESP), and e-mailed the first author of each study obtained through the first two methods.

Studies that met four criteria were included. First, they had to include both participants whose group was negatively stereotyped in the domain at hand and participants whose group was not stereotyped in the domain.

Second, they had to manipulate stereotype threat. The manipulation could be accomplished through explicit instructions that implied the validity or relevance (or invalidity or irrelevance) of a negative stereotype (e.g., Spencer et al., 1999; Studies 1 and 2 in Steele & Aronson, 1995) or through features of the test-taking environment that could increase (or decrease) the perceived relevance of a stereotype (e.g., Study 4 in Steele & Aronson, 1995). Additionally, because detecting latent ability requires removing threat from the postmanipulation measure of performance to a degree lower than that on the premanipulation measure, we excluded studies that did not adequately remove threat in the safe condition. Specifically, people link evaluative tests to negative stereotypes automatically (Steele & Aronson, 1995; Walton & Cohen, 2003). Therefore, we excluded studies whose safe condition simply portrayed the test as evaluative of the stereotyped ability (see Supporting Details in the Supporting Information available on-line). In contrast, the safe conditions in included studies either refuted the validity of the stereotype (portrayed the test as yielding no group differences), severed its relevance to the test (portrayed the test as nonevaluative of the stereotyped ability), or provided participants an identity-relevant antidote to stereotype threat (e.g., a value affirmation).

Third, studies had to assess the subsequent performance of participants on a test relevant to the stereotyped domain.

Fourth, studies had to assess performance in the same domain in a real-world context outside the testing session (e.g., prior grades or test scores). This criterion excluded studies that assessed the premanipulation measure as a pretest in the same testing session as the postmanipulation measure. We wanted to assess latent ability relative to students' real-world classroom or test performance. Additionally, studies that use a pretest design were a priori expected not to yield a latent-ability effect. If studies assessed baseline performance in safe conditions and then manipulated threat during a subsequent performance, no latent-ability effect would occur, because performance on the premanipulation measure would not have been depressed by threat (even if the introduced threat undermined performance). Alternately, if studies assessed baseline performance in threat conditions and then attempted to reduce threat during a subsequent performance, no latent-ability effect (and no stereotype-threat effect) would occur, as removing threat once it has been activated within a testing session may be difficult or impossible (see Supporting Details in the Supporting Information available on-line).

Where studies included several levels of a well-understood moderator of stereotype threat, we restricted analysis to that level where stereotype-threat effects are greatest (see Table S1 in the Supporting Information available on-line; see p. 1139). Where studies assessed several premanipulation measures of performance, we used the measure that was most relevant to the stereotype at hand and which best predicted postmanipulation performance.

Data Analysis

For each study, we calculated the mean level of performance among stereotyped and nonstereotyped students in each condition at various levels of the premanipulation measure. Specifically, in each cell of the experimental design, we conducted a linear regression with the premanipulation measure of performance predicting the postmanipulation measure of performance. Then, using the unstandardized regression coefficient (b), which indexes the slope in the relationship between the two performance measures, and the mean level of each performance measure in each cell, we used the regression equation to calculate the mean level of the postmanipulation measure in each cell at the grand mean on the premanipulation measure and at 1 standard deviation below and above the grand mean (see Supporting Details in the Supporting Information available on-line). Finally, we calculated the effect size for the difference between each of the four cells at each level of past performance. In doing so, we used the residual standard deviations from the regressions pooled across the cells being compared.[1]

Original data were unavailable in several studies, so we could not conduct the regressions. However, these studies either recruited participants within a narrow range on the premanipulation measure or reported performance levels statistically adjusted for the premanipulation measure (see Table S1 in the Supporting Information available on-line). As such, the mean levels of postmanipulation performance were at approximately the same point on the premanipulation measure for each cell. These studies were included in analyses of effects at the mean level of the premanipulation measure, but not for analyses at low and high levels of the premanipulation measure.

Meta-analytic statistics were calculated following procedures outlined by Hedges and Olkin (1985). At each level of past performance, we compared stereotyped students in threat conditions to nonstereotyped students in safe conditions (test of underperformance) and stereotyped students in safe conditions to nonstereotyped students in safe conditions (test of latent ability). There were two reasons we used nonstereotyped students in safe conditions as the comparison group (rather than those in threat conditions). First, nonstereotyped students experience a performance boost called stereotype lift when they know that an out-group is negatively stereotyped in a performance domain (Walton & Cohen, 2003). As such, threat conditions do not provide a neutral baseline. Second, in testing for latent ability we wanted to compare stereotyped and nonstereotyped students in the same condition. In real-world settings, students from both groups perform in the same circumstances.

Results and Discussion

Overview of Sample

A total of 39 independent samples, including 3,180 participants residing in five countries (Canada, France, Germany, Sweden, and the United States), met the inclusion criteria and provided usable data. The studies included a wide range of participant

[1]For implications of this approach for concerns about the use of premanipulation measures of performance as covariates in stereotype-threat research (Wicherts, 2005), see Supporting Details in the Supporting Information available on-line.

ages (kindergarten through college), stereotyped groups (e.g., African Americans, Hispanic Americans, Turkish Germans, women), manipulations of stereotype threat (see Table S1 in the Supporting Information available on-line), and measures of postmanipulation performance (diverse intellectual tests). They also included diverse measures of premanipulation performance, including classroom grades (25.64% of studies), SAT scores (64.10% of studies), state-mandated standardized test scores (5.13% of studies), ACT scores (2.56% of studies), and IQ scores (2.56% of studies). Table S1 summarizes the methodological characteristics of each included study.

Test of Predictive Validity

Meta-analyzing across studies, we first tested for a bias in predictive validity (Jensen, 1980). As predicted, there was none. The slopes indexing the degree to which the premanipulation measures of performance predicted postmanipulation performance did not differ by cell, Zs < 1, ds $\leq |0.04|$.

Tests of Underperformance and Latent Ability

Next we tested for underperformance. In threat conditions, did stereotyped students perform worse than nonstereotyped students at the same level of past performance? They did. The effect was significant at the mean level of prior performance, $Z = 9.55$, $p < 1 \times 10^{-16}$, $d = 0.48$, and at low, $Z = 7.81$, $p = 6 \times 10^{-15}$, $d = 0.44$, and high, $Z = 6.10$, $p = 1 \times 10^{-9}$, $d = 0.32$, levels (i.e., 1 standard deviation below and above the mean).

Finally, we tested the critical question of latent ability. In safe conditions that reduce threat, did the achievement gap reverse? It did. At the mean level of prior performance, stereotyped students performed better than nonstereotyped students, $Z = 3.15$, $p = .002$, $d = 0.18$. The effect proved invariant across stereotyped group (i.e., ethnic minorities vs. women), participant age (K–12 vs. college students), participant nationality (American vs. non-American), and type of premanipulation measure (classroom grades vs. test scores), $\chi^2(1, N\text{s} = 38–39)$s < 1. The effect was also significant both among students scoring low on the premanipulation measure, $Z = 2.59$, $p = .009$, $d = 0.14$, and among students scoring high on the premanipulation measure, $Z = 3.33$, $p = 9 \times 10^{-4}$, $d = 0.22$. Figure 1 depicts the degree of underperformance and of latent ability at each level of past performance.

Consideration of Alternative Explanations

Some studies assessed prior performance using self-report measures (e.g., self-reported SAT scores). Perhaps stereotyped students underreported their scores on the premanipulation measure relative to nonstereotyped students, overestimating the latent-ability effect. There was no evidence for this contention—the latent-ability effect did not vary by whether the premanipulation measure was self-reported ($n = 29$, $d = 0.16$) or not ($n = 10$, $d = 0.21$), $\chi^2(1, N = 39) < 1$. (Additionally, the intervention experiments in the second meta-analysis all assessed prior performance using official school records.)

Perhaps the results obtained in part from the "file-drawer" problem, whereby studies that show condition differences are more likely to be published and included in meta-analyses than studies that, by chance, show no condition difference. This, too,

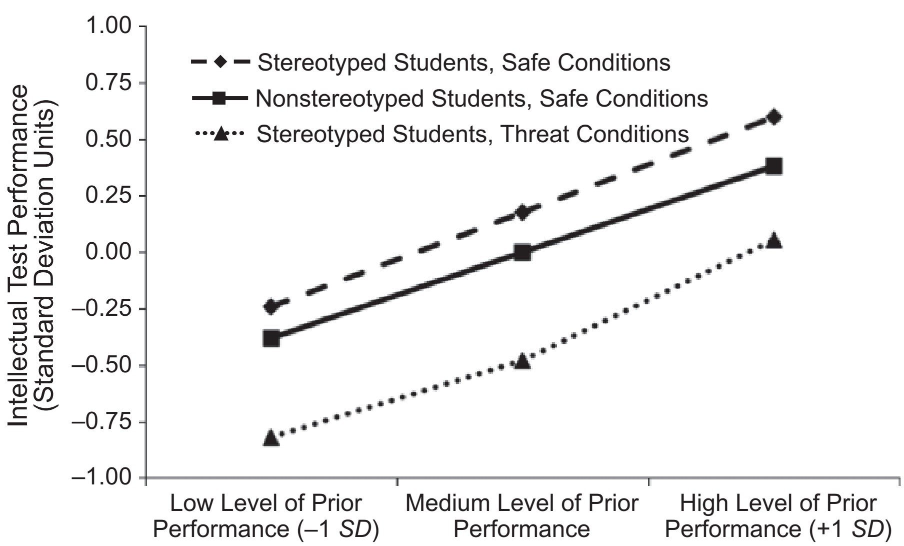

Fig. 1. Intellectual test performance of stereotyped and nonstereotyped students as a function of stereotype-threat condition and prior performance. The figure depicts the precise size of the underperformance effect (i.e., the difference between stereotyped students in threat conditions and nonstereotyped students in safe conditions) and of the latent-ability effect (i.e., the difference between stereotyped students and nonstereotyped students in safe conditions) at each level of past performance. However, the figure only closely approximates other comparisons. For details as to how the figure was created, see Supporting Details in the Supporting Information available on-line (see p. 1139).

TABLE 1
Interventions Included in the Meta-Analysis of Intervention Field Experiments

		Measure of performance	
Intervention	Population	Premanipulation	Postmanipulation
Writing-based value-affirmation exercise to reduce stereotype-related threat (Cohen, Garcia, Apfel, & Master, 2006)	African American and European American 7th-grade students	GPA in 6th grade combined with pretreatment GPA in 7th grade	Course grade over the remainder of the term
Honorific residential dormitory program to mitigate stereotype threat (e.g., students were told of the high expectations the university had for them; Steele, 1997; Steele et al., 1998)	African American and European American 1st-year college students	High school GPA combined with ACT or SAT score(s)	GPA in the first semester of college
Exercise to buffer students against adverse social events that, in light of the stereotype, could undermine their sense of social belonging in school (Walton & Cohen, 2007)	African American and European American 1st-year college students	GPA in the first semester of college	GPA from sophomore to senior year of college

Note. GPA = grade point average.

is unlikely. The included studies were designed to test stereotype threat, not latent ability. The fail-safe *n*—the number of studies, each yielding a null effect, required to render the latent-ability effect nonsignificant—was 105. Finally, the latent-ability effect was the same size in published ($n = 30$, $d = 0.16$) and unpublished studies ($n = 9$, $d = 0.25$), $\chi^2(1, N = 39) < 1.05$.[2]

META-ANALYSIS II: INTERVENTION FIELD EXPERIMENTS

The results of the stereotype-threat meta-analysis suggest that standard measures of academic performance underestimate the ability and potential of ethnic minority students and of women in quantitative fields. At each level of prior real-world performance, ethnic minorities and women scored better on intellectual tests than nonminorities and men in conditions that mitigate threat. An important question remains: Do stereotyped students show the same superior performance when threat is reduced and performance is assessed in real-world school environments? We tested this question in a second meta-analysis. This meta-analysis focused on interventions to reduce stereotype-related threat in real-world school environments.

Method

We retrieved intervention experiments following the same procedures described for the first meta-analysis. The inclusion criteria were identical with the exception of two changes. First, the manipulation had to be an intervention aimed at reducing stereotype-related threat in a real-world classroom environment. This intervention had to be narrowly targeted at reducing stereotype-related threat; if instead it targeted processes that boost achievement among students from all social groups it was excluded (e.g., Aronson, Fried, & Good, 2002; Wilson, Damiani, & Shelton, 2002). Second, the measure of posttreatment performance had to assess real-world classroom performance rather than a score on a test. Data were analyzed in the same manner as in the stereotype-threat meta-analysis.

Results and Discussion

Overview of Sample

We obtained three intervention experiments including data from 15,796 students that met the inclusion criteria. Each intervention aimed to reduce stereotype-related threat among African American students. They employed diverse techniques to reduce threat. These are the only randomized field experiments we know of targeted specifically at reducing stereotype-related threat in a classroom environment. Table 1 summarizes the methodology of each study (for details, see the publications).

Preliminary Analyses

Preliminary analyses tested the appropriateness of combining theoretically similar cells to maximize statistical power in primary analyses. First, we tested whether the treatments affected European American students' performance (controlling for prior performance). No treatment did, $ts < 1$. Therefore, in each study, we combined European American students who did and did not receive the treatment.

Second, in two of the interventions, the original researchers obtained the academic records of all students in the same class

[2]A third possibility involves stereotype lift. Perhaps nonstereotyped students' performance on the premanipulation measure was enhanced by stereotype lift and so overestimated their ability in an environment without threat. The present data do not address this question directly. However, because stereotype threat is twice the magnitude of stereotype lift (Walton & Cohen, 2003), stereotype threat likely explains the bulk of the effect. Furthermore, in the intervention meta-analysis, the performance of nonstereotyped students was the same in treatment and control conditions, so the latent-ability effect observed there clearly reflects the underestimation of stereotyped students' ability on the premanipulation measure.

year as participants who had not participated in the study (Steele, 1997; Walton & Cohen, 2007). These students formed a second control group. In general, they performed at the same level as nontreated, same-race participating students, $ts < 1.35$. There was one exception: In the Steele (1997) intervention, nonparticipating European American students performed slightly worse (adjusted mean = 2.91) than participating European American students (adjusted mean = 3.04), $t(14719) = 4.27$, $p = 2 \times 10^{-5}$, $p_{rep} = .998$. This pattern likely reflects a selection effect (Steele et al., 1998). Whereas African American participants were actively recruited for the study from a random sample of African American students campus-wide, European American participants volunteered. Consequently, it is inappropriate to compare African and European American participants (Steele et al., 1998); doing so would confound race and recruitment procedure (see Supporting Details in the Supporting Information available on-line). So, in both studies, we combined nonparticipating and participating students. The European American group comprised European American students in treatment and control conditions and, if available, nonparticipating European American students. The African American control condition comprised African American students in the control condition and, if available, nonparticipating African American students.

Tests of Underperformance and Latent Ability

First, we obtained evidence for underperformance. Replicating past studies (e.g., Jensen, 1980), in control conditions, African American students performed worse than European American students at the mean level of prior performance, $Z = 8.83$, $p < 1 \times 10^{-16}$, $d = 0.27$, at low levels of prior performance, $Z = 11.73$, $p < 1 \times 10^{-16}$, $d = 0.31$, and at high levels of prior performance, $Z = 5.83$, $p = 6 \times 10^{-9}$, $d = 0.23$. As noted, a possible cause of this effect is that, absent intervention, psychological threat may increase as students progress to higher levels of study.

Second, we obtained evidence for latent ability. In treatment conditions, African American students performed better at the mean level of prior performance than European American students, $Z = 2.97$, $p = .003$, $d = 0.17$. The effect was also significant at both low levels of prior performance, $Z = 2.67$, $p = .008$, $d = 0.12$, and high levels of prior performance, $Z = 3.23$, $p = .001$, $d = 0.22$. Figure 2 depicts the degree of underperformance and of latent ability in each study and meta-analytically at each level of past performance.

GENERAL DISCUSSION

Both meta-analyses reveal a bias in standard measures of academic performance—both test scores and classroom grades. The bias results from psychological threat. It causes measures of academic performance to underestimate the true intellectual ability and potential of ethnic minority students and of women in quantitative fields.

The results quantify the bias. It is just under one fifth of a standard deviation: $0.17 \leq d \leq 0.18$. Although the present analyses yielded no significant moderator, the level of bias may differ for different students or in different performance environments. Additionally, the estimate may prove conservative. It reflects only that portion of psychological threat that research has identified and remedied. To the extent that unidentified or unremedied psychological threats further undermine performance, the results underestimate the bias.

Nevertheless, the observed bias is large enough to account for a meaningful proportion of group differences on high-stakes tests. We illustrate the magnitude of the bias in relation to the SAT, because most of the studies in the stereotype-threat meta-analysis used SAT scores as the premanipulation measure of performance (64% did). Although the present sample is not necessarily representative of students who take the SAT, it does reflect the level of bias in a broad range of promising students (for a comparison of these populations, see Supporting Details in the Supporting Information available on-line).

The observed effect sizes suggest that the SAT Math test underestimates the math ability of women like those in the present sample by 19 to 21 points, and that the SAT Math and SAT Reading tests underestimate the intellectual ability of African and Hispanic Americans like those in the present sample by a total of 39 to 41 points for each group. Insofar as the overall gender gap on the SAT Math test is 34 points and as the overall Black-White and Hispanic-White gaps on the SAT (combining math and reading) are 199 and 148 points, respectively (The College Board, 2007), these differences are substantial. The results suggest that, at least for these promising students, the psychological context of common testing environments significantly undermines real-world performance.

The results also show that psychological treatments can recover much of this otherwise lost human potential. Future research should develop ways to "scale up" threat-reducing interventions to improve performance in mass. To close achievement gaps, it is necessary both to eradicate psychological threats embedded in academic environments and to remove other barriers to achievement including objective biases, the effects of poverty, and so forth.

Finally, the results suggest that schools or employers that reduce psychological threat in their internal environment may observe latent ability in measures used to make admissions or hiring decisions. People from stereotyped groups may perform better than people from nonstereotyped groups at the same level of prior performance. This observation suggests that the prior measure, even if predictive of subsequent intellectual or work performance, is biased against certain groups. A critical task for institutions is to determine how to account for this bias so as to make selection decisions that are meritocratic and that do not discriminate against deserving people from stereotyped groups.

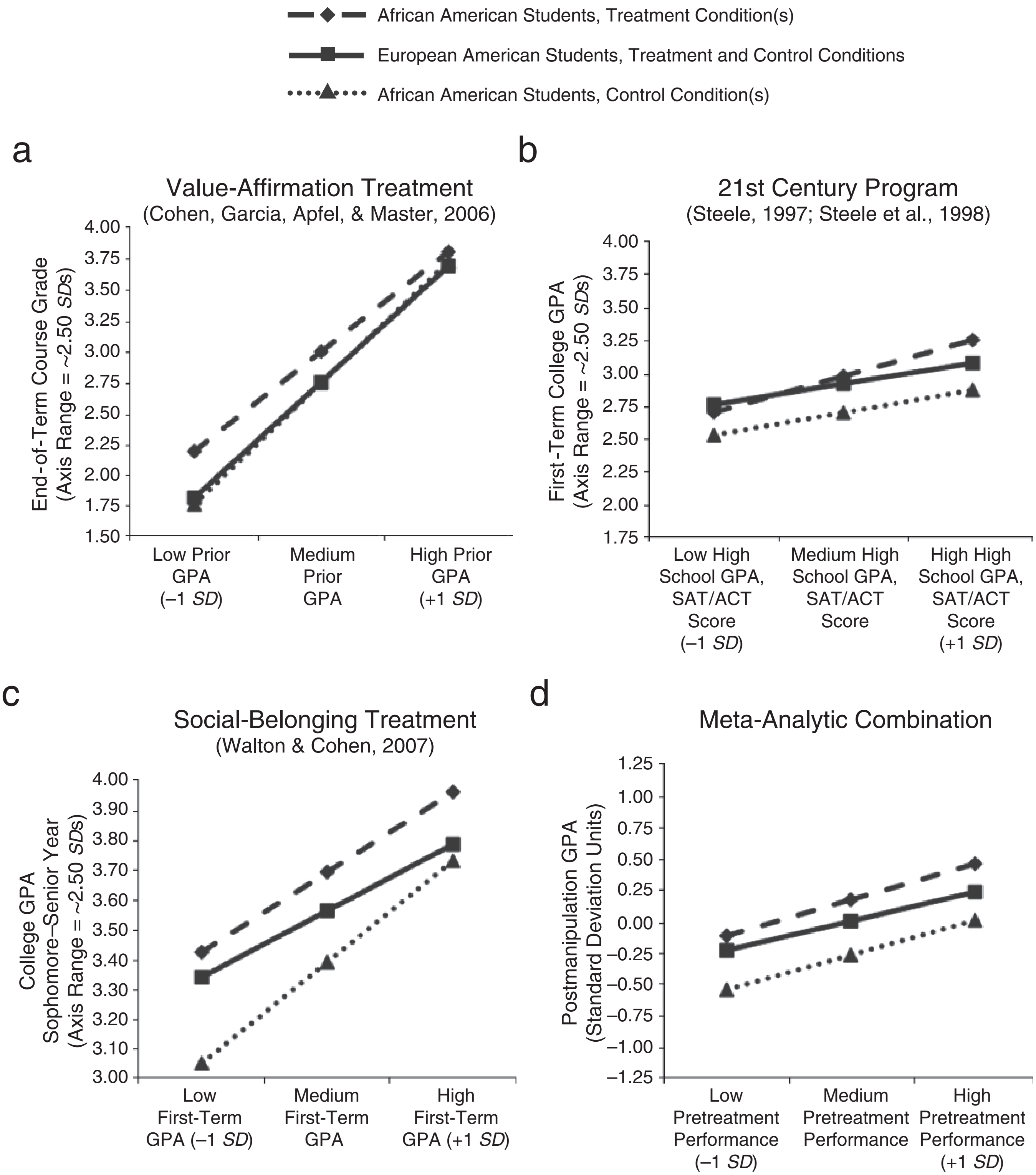

Fig. 2. Classroom performance of African American and European American students in control conditions and in treatment conditions designed to reduce stereotype-related threat, as a function of prior performance. The European American control and treatment conditions are combined because they did not yield differences in performance in any study, $ts < 1$. The figure shows the results of (a) the value-affirmation intervention (Cohen, Garcia, Apfel, & Master, 2006), (b) the 21st Century Program (Steele, 1997; Steele et al., 1998), (c) the social-belonging intervention (Walton & Cohen, 2007), and (d) the meta-analytic combination of the intervention results. The graph in (d) has the same characteristics and was created in the same manner as Figure 1 (see Supporting Details in the Supporting Information available on-line; see p. 1139). GPA = grade point average.

Acknowledgments—We thank the authors of the included studies for generously contributing data; Mahzarin Banaji, Geoffrey Cohen, Carol Dweck, John Haymaker, Scott Klemmer, David Nussbaum, Christine Logel, Pam Salder, Claude Steele, Julia Steinberg, Eric Uhlmann, and Erik Woody for helpful input; Bjanka Pokorny for assistance; and the National Academy of Education/Spencer Foundation for supporting this research through a postdoctoral fellowship awarded to the first author.

REFERENCES

Aronson, J., Fried, C.B., & Good, C. (2002). Reducing the effect of stereotype threat on African American college students by shaping theories of intelligence. *Journal of Experimental Social Psychology*, *38*, 113–125.

Benbow, C.P., & Stanley, J.C. (1980). Sex differences in mathematical ability: Fact or artifact? *Science*, *210*, 1262–1264.

Cohen, G.L., Garcia, J., Apfel, N., & Master, A. (2006). Reducing the racial achievement gap: A social-psychological intervention. *Science*, *313*, 1307–1310.

The College Board. (2007). *2007 college-bound seniors: Total group profile report*. Retrieved April 19, 2008, from http://www.collegeboard.com/prod_downloads/about/news_info/ cbsenior/yr 2007/national-report.pdf

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

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

Eccles, J.S. (1987). Gender roles and women's achievement-related decisions. *Psychology of Women Quarterly*, *11*, 135–172.

Fryer, R.G., Jr., & Levitt, S.D. (2004). Understanding the Black-White test score gap in the first two years of school. *The Review of Economics and Statistics*, *86*, 447–464.

Hedges, L.V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Herrnstein, R.J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life*. New York: Free Press.

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

Jencks, C., & Phillips, M. (Eds.). (1998). *The Black-White test score gap*. Washington DC: Brookings Institution Press.

Jensen, A.R. (1980). *Bias in mental testing*. New York: Free Press.

Massey, D.S., & Fischer, M.J. (2005). Stereotype threat and academic performance: New findings from a racially diverse sample of college freshmen. *Du Bois Review*, *2*, 45–67.

Sackett, P.R., Borneman, M.J., & Connelly, B.S. (2008). High-stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist*, *63*, 215–227.

Sackett, P.R., Hardison, C.M., & Cullen, M.J. (2004). On interpreting research on stereotype threat and test performance. *American Psychologist*, *60*, 271–272.

Schmader, T., Johns, M., & Forbes, C. (2008). An integrated process model of stereotype threat effects on performance. *Psychological Review*, *115*, 336–356.

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.

Spencer, S., Steele, C.M., & Quinn, D. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*, 4–28.

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

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

Steele, C.M., Spencer, S., Nisbett, R., Hummel, M., Harber, K., Schoem, D., & Carter, K. (1998). *African American college achievement: A "wise" intervention*. Unpublished manuscript, Stanford University, Stanford, CA.

Steele, C.M., Spencer, S.J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. In M.P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 37, pp. 397–407). San Diego, CA: Academic Press.

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

Walton, G.M., & Cohen, G.L. (2007). A question of belonging: Race, social fit, and achievement. *Journal of Personality and Social Psychology*, *92*, 82–96.

Wicherts, J.M. (2005). Stereotype threat research and the assumptions underlying analysis of covariance. *American Psychologist*, *60*, 267–269.

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.

Wilson, T.D., Damiani, M., & Shelton, N. (2002). Improving the academic performance of college students with brief attribution interventions. In J. Aronson (Ed.), *Improving academic achievement: Impact of psychological factors on education* (pp. 89–108). Oxford, England: Academic Press.

(RECEIVED 11/10/08; REVISION ACCEPTED 1/25/09)

SUPPORTING INFORMATION

Additional Supporting Information may be found in the on-line version of this article:

Supporting Details
Table S1

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

SUPPORTING DETAILS

I. What We Mean By "Latent"

We use the word "latent" in its common language meaning of "hidden, concealed; present or existing, but not manifest, exhibited, or developed" (Oxford English Dictionary, 2009). The argument that the intellectual ability of stereotyped students is "latent" is thus to say that their ability is (in part) hidden on common measures of academic performance—that these measures underestimate stereotyped students' true intellectual ability relative to nonstereotyped students.

This usage differs from the use of latent in statistical parlance as for instance a "latent variable" (e.g., Wicherts, Dolan, & Hessen, 2005). In the sense of a latent variable, an academic performance may tap a student or group's underlying (or latent) level of intellectual ability. In this sense, all students and all groups have some degree of underlying "latent ability." To denote this underlying level of ability, we reserve the word "true" as in "true intellectual ability."

This distinction explicates the "latent ability" argument. Our claim that stereotyped students' intellectual ability is "latent" (or hidden) on common performance measures (i.e., "the latent ability hypothesis") means that these measures underestimate stereotyped students' true ability (latent variable) such that, at the same level of performance on these measures, stereotyped students are on average better students than nonstereotyped students. This claim should not be confused with a claim about the presence or the absence of group differences in true intellectual ability (in either direction) whatever students' level of performance on common academic measures. The present results could obtain if group differences in true ability exist or not. For instance, we do not claim that stereotyped students have "more latent ability" (in the latent variable sense) than nonstereotyped students—that they are necessarily better students.

II. The Latent Ability Effect by Type of Safe Condition

The meta-analysis of stereotype threat experiments excluded studies whose safe condition portrayed the test simply as evaluative of the stereotyped ability. As people link evaluative tests to negative stereotypes automatically (Walton & Cohen, 2003), such "weak" safe conditions may not remove stereotype threat sufficiently to detect the latent ability effect. In fact, evaluative instructions are a classic means of inducing stereotype threat (Steele & Aronson, 1995).

We tested this hypothesis empirically. This criterion excluded 15 independent samples (Inzlicht & Ben-Zeev, 2000 [Study 2]; Josephs, Newman, Brown, & Beer, 2003 [Study 1]; Keller, 2002; Marx & Goff, 2005; Marx & Roman, 2002 [Studies 1 and 2]; Mayer & Hanges, 2003; Nguyen, O'Neal, & Ryan, 2003; Schmader, 2002; Schultz, Baker, Herrera, & Khazian, 2008 [Studies 1-3]; Tagler, 2003 [Study 2]; Wicherts, Dolan, & Hessen, 2005 [Studies 1 and 3]). We compared the latent ability effect in these studies to the latent ability effect in studies that used "strong" safe conditions (i.e., the effect reported in the main text) at the mean level of past performance. As predicted, the effect differed significantly across the two types of studies, $\chi^2(1, n = 39) = 13.19$, $p = .0003$. Whereas studies that used strong safe conditions yielded a significant latent ability effect, $Z = 3.15$, $p = .002$, $d = .18$, studies that used weak safe conditions yielded, if anything, an effect in the opposite direction, $Z = -1.93$, $p = .053$, $d = -.16$. The results underscore the critical importance of using effective means for reducing stereotype threat to detect latent ability.

III. Stereotype Threat Experiments Excluded For Using a "Pre-Test" Design

The second category of excluded stereotype threat experiments were those that assessed the premanipulation measure as a pre-test in the same testing session in which the postmanipulation measure was assessed. This criterion excluded 5 independent samples (Dinella, 2004; Muzzatti & Agnoli, 2007 [Study 2: 3rd graders, 5th graders, 8th graders]; Salinas & Aronson, 2007). Retaining these samples does not affect the statistical significance of any reported analysis. Other than these studies and the studies cited in the previous section, no study that provided usable data (and which met the basic inclusion criteria) was excluded from the meta-analysis of stereotype threat experiments.

IV. Calculation of the Level of Postmanipulation Performance at Each Level of the Premanipulation measure

For each study in both meta-analyses we calculated the mean level of postmanipulation performance of participants in each cell of the experimental design at the mean level of the premanipulation measure of performance and at 1 standard deviation below and above the mean level. First, in each cell we conducted a linear regression with the premanipulation measure predicting the postmanipulation measure. Second, we used the equations below derived from the regression equation to calculate the level of postmanipulation performance at each level of the premanipulation measure of performance.

Definitions

- b = 'b' unstandardized regression coefficient indexing the slope in the relationship between the pre- and postmanipulation measure in the cell
- mpostc = mean level of the postmanipulation measure of performance in the cell
- mprec = mean level of the premanipulation measure of performance in the cell
- mpreg = grand mean of the premanipulation measure of performance
- sdpreg = standard deviation of the premanipulation measure of performance

Equations

- mean level of the postmanipulation measure of performance in the cell at 1 standard deviation below the grand mean of the premanipulation measure of performance
  = mpostc – b*(mprec – [mpreg – sdpreg])
- mean level of the postmanipulation measure of performance in the cell at the grand mean of the premanipulation measure of performance
  = mpostc – b*(mprec – mpreg)
- mean level of the postmanipulation measure of performance in the cell at 1 standard deviation above the grand mean of the premanipulation measure of performance
  = mpostc – b*(mprec – [mpreg + sdpreg])

V. Implications For the Use of Premanipulation measures of Performance As Covariates in Stereotype Threat Research

Whereas concerns about the robustness of stereotype threat effects have been raised as many stereotype threat studies feature analyses of covariance (ANCOVAs) that may not meet the assumptions of such analyses (Wicherts, 2005), the current analyses, which do not rely on these assumptions, demonstrate that both the basic stereotype threat effect and the latent ability effect are robust even in a more appropriate and conservative test.

VI. Combination of Theoretically Similar Cells in the Meta-Analysis of Intervention Field Experiments

The main text describes procedures undertaken to combine theoretically similar cells in the meta-analysis of intervention field experiments. As just 3 independent samples were available for this meta-analysis, doing so was especially important to maximize statistical power (see Hedges & Pigott, 2001). First, as noted, because no treatment affected European Americans' performance, each study combined control- and treatment-condition European Americans. Doing so does not affect the statistical significance of any reported analysis.

Second, in the Walton and Cohen (2007) and Steele (1997) interventions, we combined students campus-wide who had not participated in the study with non-treated, same-race participating students. While there was no difference in performance between these groups in the Walton and Cohen intervention, in the Steele intervention participating European American students performed slightly better than non-participating European American students (adjusted mean difference = .13). Because non-participating European American students performed slightly worse, their inclusion slightly lowers the performance of the European American comparison group used to compute the latent ability and underperformance effects.

However, combining non-participating and control students not only increases statistical power; in the Steele intervention, it is the appropriate way to analyze the data. As the key analyses compare European and African American students, such students must be comparable. But in the Steele intervention, European American students volunteered for this "honorific" program whereas African American students were actively recruited from a random sample of African American students to participate in it. As a consequence, the original authors treat non-participating European American students as the most appropriate cross-race comparison group (Steele et al., 1998). They contend that it would be inappropriate to compare a (somewhat) select group of European American students to a more or less representative group of African American students; doing so would confound race and recruitment procedure. Following this reasoning, we compare a representative group of African American students to a representative group of European American students (i.e., those participating and those not participating in the study).

Additionally, the latent ability effect in the meta-analysis of intervention field experiments remains significant even entirely excluding the Steele intervention, at low levels of prior performance, Z = 3.85, $p = 1 \times 10^{-4}$, medium levels of prior performance, Z = 3.07, $p = .002$, and at high levels of prior performance, $Z = 2.25$, $p = .024$.

VII. Creation of Meta-Analytic Figures (Figures 1 and 2D)

Figure 1 was created by assigning the performance of nonstereotyped students in safe conditions at the mean level of past performance a value of *0*, and using this value to calculate the relative level of performance at each other point. First, the levels of performance of nonstereotyped students in safe conditions at 1 standard deviation below and above the mean represent the effect sizes indexing the differences in performance between these students and nonstereotyped students in safe conditions at the mean level of prior performance. Second, the levels of performance of stereotyped students in threat and safe conditions represent the effect sizes indexing the differences in performance between these students and nonstereotyped students in safe conditions at each level of prior performance (i.e., underperformance and latent ability, respectively). The figure thus depicts the precise size of the underperformance and latent ability effect at each level of past performance. However, as the studies have slightly different sample sizes for other comparisons (e.g., for the stereotype threat effect—the difference between stereotyped students in safe vs. threat conditions), the figure only closely approximates such effects. Figure 2D was created in the same manner and has the same characteristics.

VIII. Comparison of Participant Sample to Students Who Take the SAT

We compared students who participated in the included stereotype threat studies to students who take the SAT on two dimensions. First, group differences in the two samples were similar in magnitude. The race difference on the premanipulation measures of performance (d = .87; 95% CI: $.73 \leq d \leq 1.02$) was comparable to race differences on the SAT (SAT-Reading and SAT-Math: White/Black ds = .83 and .92, respectively; White/Hispanic ds = .65 and .66, respectively; College Board, 2007). The gender difference on the premanipulation measures of math performance (d = .16; 95% CI: $.07 \leq d \leq .26$) was comparable to the gender difference on the SAT-Math test (d = .30; College Board, 2007).

Second, participating students were somewhat higher performing. Among studies that used SAT scores as the premanipulation measure and weighting by sample size, participating students earned average SAT-Reading scores between 530 and 688 and average SAT-Math scores between 563 and 699 (each at 1 standard deviation below and above the mean), which correspond to 60^{th} to 95^{th} and 66^{th} to 95^{th} percentile on these tests (College Board, 2007).

Supplementary References

References to studies included in the meta-analysis of stereotype threat experiments are marked with an asterisk ("*").

*Bell, A. E., Spencer, S. J., Iserman, E., & Logel. C. E. R. (2003). The effect of stereotype threat on women's engineering performance. *Journal of Engineering Education, 92*, 307-312.

*Biek, D. M. (2006). Stereotype threat and domain identification. Unpublished doctoral dissertation, Cornell University.

*Broadnax, S., Crocker, J., & Spencer, S. (1997). African-Americans and academic performance: The mediational role of stereotype vulnerability. Unpublished manuscript, State University of New York at Buffalo.

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

*Brown, J. L., Steele, C. M., & Atkins, D. (2001). Performance expectations are not a necessary mediator of stereotype threat in African American verbal test performance. Unpublished manuscript, University of Washington, Seattle.

The College Board. (2007). *2007 college-bound seniors: Total group profile report*. Retrieved April 19, 2008 from http://www.collegeboard.com/prod_downloads/about/news_info/cbsenior/yr2007/national-report.pdf.

*Davies, P. G., Spencer, S. J., Quinn, D. M., & Gerhardstein, R. (2001). Consuming images: How television commercials that elicit stereotype threat can restrain women academically and professionally. *Personality and Social Psychology Bulletin, 28,* 1615-1628.

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

*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.

*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.

*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.

Hedges, L. V. & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods, 6,* 203-217.

*Ho, A. K. & Sidanius. J. (2008). Preserving positive identities: Public and private regard for one's ingroup and susceptibility to stereotype threat. Unpublished manuscript, Harvard University.

*Huguet, P. & Régner, I. (2007). Stereotype threat among schoolgirls in quasi-ordinary classroom circumstances. *Journal of Educational Psychology, 99,* 545-560.

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.

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

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

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

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

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

*Kellow, J. T. & Jones, B. D. (2005). Stereotype threat in African-American high school students: An initial investigation. *Current Issues in Education* [On-line]*, 8(15).* Available: http://cie.ed.asu.edu/volume8/number20/

*Kellow, J. T. & Jones, B. D. (2008). The effects of stereotypes on the achievement gap: Reexamining the academic performance of African American high school students. *Journal of Black Psychology, 34,* 94-120.

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

*Logel, C., Iserman, E. C., Davies, P. G., Quinn, D. M., & Spencer, S. J. (in press). The perils of double consciousness: The role of thought suppression in stereotype threat. *Journal of Experimental Social Psychology*.

*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.

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.

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

Mayer, D. M. & Hanges, P. J. (2003). Understanding the stereotype threat effect with "culture-free" tests: An examination of its mediators and measurement. *Human Performance, 16,* 207-230.

*McKown, C. & Weinstein, R. S. (2003). The development and consequences of stereotype consciousness in middle childhood. *Child Development, 74,* 498-515.

Muzzatti, B. & Agnoli, F. (2007). Gender and mathematics: Attitudes and stereotype threat susceptibility in Italian children. *Developmental Psychology, 43,* 747-759.

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 on cognitive ability test performance. *Human Performance, 16,* 261-293.

Oxford English Dictionary (2009). Retrieved from http://dictionary.oed.com/cgi/entry/50130201?query_type=word&queryword=latent&first=1&max_to_show=10&sort_type=alpha&search_id=bIjv-hGsSIF-6240&result_place=1. Accessed May 20, 2009.

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

*Rudiger, L. P. (1998). Resisting stigma: Stereotype compensation and women's math performance. Unpublished doctoral dissertation, University of Vermont.

Salinas, M. F. & Aronson, J. (2007). Differential stereotype threat mechanisms in high and low stakes situations for Mexican Americans. Manuscript in preparation.

Schmader, T. (2002). Gender identification moderates stereotype threat effects on women's math performance. *Journal of Experimental Social Psychology, 38,* 194-2001.

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

Schultz, P. W., Baker, N., Herrera, E., & Khazian, A. (2008). Stereotype threat among Hispanic-Americans and the moderating role of ethnic identity. Manuscript in preparation, California State University, San Marcos.

*Shih, M., Bonam, C., Sanchez, D., & Peck, C. (2007). The social construction of race: Biracial identity and vulnerability to stereotypes. *Cultural Diversity and Ethnic Minority Psychology, 13,* 125-133.

*Spencer, S. J. (1993). The effect of stereotype vulnerability on women's math performance. Unpublished doctoral dissertation, University of Michigan, Ann Arbor.

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

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

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

Steele, C. M., Spencer, S., Nisbett, R., Hummel, M., Harber, K., Schoem, D., & Carter, K. (1998). African American college achievement: A "wise" intervention. Manuscript in preparation, Stanford University.

*Stünzendörfer, A. (2007). Stereotype threat: Eine Bedrohung für türkische Schüler an deutschen Grundschulen? [Stereotype threat: A threat to Turkish students in German elementary schools?] Unpublished masters thesis, Friedrich-Alexander Universität.

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

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

Wicherts, J. M. (2005). Stereotype threat research and the assumptions underlying analysis of covariance. *American Psychologist, 60,* 267-269.

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.

*Yopyk, D. J. A. & Prentice, D. A. (2005). Am I an athlete or a student? Identity salience and stereotype threat in student-athletes. *Basic and Applied Social Psychology, 27,* 329-336.

Table S1

Methodological characteristics of studies included in the meta-analysis of stereotype threat experiments (Walton & Spencer, 2009)

Study	Population	Stereotyped Group	Non-Stereotyped Group	Manipulation of Stereotype Threat	Pre-Manipulation Measure of Performance	Post-Manipulation Measure of Performance	Notes
Bell et al. (2003)	US college engineering students	Women	Men	Test characterized as ability-diagnostic, as yielding no gender differences, or as diagnostic of ability	GPA in engineering	Fundamentals of Engineering Exam problems	1
Biek (2005), Study 1	US college students	Women	Men	Test characterized either as yielding gender differences or as diagnostic of mathematical ability	SAT-Math score	GRE-Math problems	
Broadnax, Crocker, & Spencer (1997)	US college students	African and Hispanic Americans	European Americans	Test characterized either as ability-diagnostic or as culture fair and as yielding no race differences	SAT-Verbal score	English literature test	
Brown & Day (2006)	US college students	African Americans	European Americans	Test characterized either as ability-diagnostic or as diagnostic of intelligence	ACT score	Raven Advanced Progressive Matrices	
Brown, Steele, & Atkins (2001), Study 1	US college students	African Americans	European Americans	Test characterized either as ability-diagnostic or as racially-fair	SAT-Verbal score	GRE-Verbal problems	2
Brown et al. (2001), Study 2	US college students	African Americans	European Americans	Test characterized either as ability-diagnostic or as racially-fair	SAT-Verbal score	GRE-Verbal problems	2
Davies, Spencer, Quinn, & Gerhardstein (2001), Pilot Study	US college students	Women	Men	Test characterized either as a test or as diagnostic of mathematical ability	SAT-Math score	GRE-Math problems	
Eriksson & Lindholm (2007)	Swedish college students	Women	Men	Diagnostic test said either to assess gender differences or to assess individual differences in math ability and to yield no gender differences	High school math grade	GRE-math problems	
Gonzales, Blanton, & Williams (2002), by gender	US college students	Women	Men	Test characterized either as diagnostic or as diagnostic of spatial and mathematical ability	SAT-Math score	Wonderlic Personnel Test	2, 3
Gonzales et al. (2002), by race	US college students	Latino Americans	European Americans	Test characterized either as diagnostic or as diagnostic of spatial and mathematical ability	SAT-Math score	Wonderlic Personnel Test	2, 3
Good, Aronson, & Harder (2008)	US college calculus students	Women	Men	Test characterized either as diagnostic of math ability or as yielding no gender differences.	Grade in calculus course	Practice calculus test	
Ho & Sidanius (2008)	US college students	African Americans	European Americans	Test characterized either as diagnostic or as diagnostic of verbal ability.	SAT-Verbal score	GRE-Verbal problems	
Huguet & Régner (2007), Study 1	French middle school students	Girls	Boys	Test characterized either as diagnostic of ability in geometry or as a memory game	Middle school math grade	Recall of a complex figure	
Huguet & Régner (2007), Study 2a	French middle school students	Girls	Boys	Mixed-sex groups completed test characterized as diagnostic of ability either in geometry or in drawing	Middle school math grade	Recall of a complex figure	
Huguet & Régner (2007), Study 2b	French middle school students	Girls	Boys	Same-sex groups completed test characterized as diagnostic of ability either in geometry or in drawing	Middle school math grade	Recall of a complex figure	
Johns, Schmader, & Martens (2005)	US college students	Women	Men	Test characterized either as ability-diagnostic and as examining gender difference or as diagnostic	SAT-Math score	GRE-Math problems	4
Keller (2007)	German high school students	Girls	Boys	Test characterized either as yielding gender differences or as yielding no gender differences	High school math grade	TIMSS and math textbook problems	5
Keller & Dauenheimer (2003)	German high school students	Girls	Boys	Test characterized either as yielding gender differences or as yielding no gender differences	High school math grade	TIMSS and math textbook problems	
Kellow & Jones (2005)	US 9th grade students	African Americans	European Americans	Test characterized either as evaluative of math ability or as culture fair and as yielding no group differences	8th grade FCAT math score	APR Spatial Ability Test	
Kellow & Jones (2008)	US 9th grade students	African Americans	European Americans	Test characterized either as evaluative of math ability or as culture fair and as yielding no group differences	8th grade FCAT math score	APR Spatial Ability Test	
Lesko & Corpus (2006)	US college students	Women	Men	Diagnostic math test characterized either as yielding gender differences or as yielding no gender differences	SAT-Math score	GRE-Math problems	

(*table continues*)

Table S1 (*Continued*)

Study	Population	Stereotyped Group	Non-Stereotyped Group	Manipulation of Stereotype Threat	Pre-Manipulation Measure of Performance	Post-Manipulation Measure of Performance	Notes
Logel, Iserman, Davies, Quinn, & Spencer (in press), Study 4	Canadian college students	Women	Men	Before ability-diagnostic test, participants either provided no special instructions or a value-affirmation strategy	High school math grade	GRE-Math problems	
Martens, Johns, Greenberg, Schimel (2006), Study 1	US college students	Women	Men	Test characterized either as ability-diagnostic or as diagnostic, or participants completed a value-affirmation task before the ostensibly diagnostic test	SAT-Math (or equivalent ACT-Math) score	GMAT-Math problems	1
Martens et al. (2006), Study 2	US college students	Women	Men	Before ability-diagnostic test examining gender differences, participants completed either a neutral task or a value-affirmation task	SAT-Math score	Mental Rotation Test	
McKown & Weinstein (2003), Study 2	US children aged 6-10	African and Hispanic Americans	European and Asian Americans	Test characterized either as diagnostic or as diagnostic of ability in school	Vocabulary Subtest of the WISC-III	Backwards recall of the alphabet	6
Quinn & Spencer (2001), Study 2	US college students	Women	Men	Test characterized either as math problems developed for the SAT or as gender-fair and as yielding no gender differences	SAT-Math score	SAT-Math problems	2
Rudiger (1998)	US college students	Women	Men	Test characterized either as evaluative of math ability and as yielding gender differences or as not predictive of ability	SAT-Math score	GRE-Math problems	2, 5, 7
Schmader & Johns (2003), Study 1	US college students	Women	Men	Test characterized as diagnostic of either quantitative capacity and as examining gender differences or of working memory capacity	SAT-Math score	Operation-Span Test	
Schmader & Johns (2003), Study 2	US college students	Hispanic Americans	European Americans	Test of working memory capacity either characterized as predictive of intelligence test performance or not	SAT-Verbal score	Operation-Span Test	
Shih, Bonam, Sanchez, & Peck (2007), Study 3	US college students	African Americans	European Americans	Before test, participants either completed questions related to their ethnicity or not	SAT-Math score	Canadian Math Competition problems	
Spencer (1993), Study 4	US college students	Women	Men	Test characterized either as yielding gender differences or as yielding no gender differences	SAT-Math score	GRE-Math problems	2, 8
Spencer (1993), Study 5	US college students	Women	Men	Test characterized either as a math test, as a math test that might yield gender differences, or as yielding no gender differences	SAT-Math score	GRE-Math problems	2, 9
Spencer, Steele, & Quinn (1999), Study 2	US college students	Women	Men	Test characterized either as yielding gender differences or as yielding no gender difference	SAT-Math score	GRE-Math problems	2
Spencer et al. (1999), Study 3	US college students	Women	Men	Test characterized either as diagnostic of math ability or as yielding no gender differences	SAT-Math score	GMAT-Math problems	
Steele & Aronson (1995), Study 1	US college students	African Americans	European Americans	Test characterized either as diagnostic or as diagnostic of verbal ability	SAT-Verbal score	GRE-Verbal problems	
Steele & Aronson (1995), Study 2	US college students	African Americans	European Americans	Test characterized either as diagnostic or as diagnostic of verbal ability	SAT-Verbal score	GRE-Verbal problems	
Steele & Aronson (1995), Study 4	US college students	African Americans	European Americans	Before diagnostic test, participants either indicated race or not	SAT-Verbal score	GRE-Verbal problems	
Stünzendörfer (2007)	German 4th grade students	Turkish Germans	Non-Minority Germans	Test characterized either as ability-diagnostic and as examining group differences or as diagnostic	Teacher evaluation of German language ability	Raven Standard Progressive Matrices	
Yopyk & Prentice (Study 1, 2005)	US college students	Male athletes	Male singers	Before test, participants wrote either about a recent athletic/singing event or about a recent school experience	SAT score	GRE- and SAT-Math-like problems	2

Notes. 1. Analysis combines two safe conditions.
2. Complete data on the pre-measure were unavailable. However, the study equated students on the pre-measure either by recruiting a narrow range of participants along this measure or by reporting mean performance levels statistically adjusted for the pre-measure. Results were treated as a point estimate at the mean level of the pre-measure. Study was excluded from analyses at low and high levels of the pre-measure.
3. As the same group of participants was included in analyses of both gender and race, both analyses are weighted by .5.
4. Excludes a condition in which the test was portrayed as ability-diagnostic and where participants were taught about stereotype threat.
5. Analysis restricted to participants identified as valuing math.
6. Analysis restricted to participants identified as aware of societal ethnic stereotypes.
7. Analysis restricted to participants under time pressure.
8. Excludes a condition in which participants were told they were expected to perform poorly.
9. Analysis combines two threat conditions.