Personality and Social Psychology Bulletin

http://psp.sagepub.com/

Stereotype Threat Undermines Academic Learning

Valerie Jones Taylor and Gregory M. Walton

Pers Soc Psychol Bull 2011 37: 1055 originally published online 9 May 2011
DOI: 10.1177/0146167211406506

Personality and Social
Psychology Bulletin
37(8) 1055–1067
© 2011 by the Society for
Personality and Social Psychology, Inc
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146167211406506
http://pspb.sagepub.com

Stereotype Threat Undermines Academic Learning

Valerie Jones Taylor[1] and Gregory M. Walton[2]

Abstract

Two experiments tested whether stereotype threat can undermine the acquisition of academic knowledge and thus harm performance even in nonthreatening settings. In Experiment 1, Black and White students studied rare words in either nonthreatening or threatening conditions. One to two weeks later, participants recalled word definitions, half in a nonthreatening "warm-up" and half in a threatening "test." Replicating past research, Black students performed worse on the test than on the warm-up. But importantly, Black students who had studied in the threatening rather than nonthreatening environment performed worse even on the warm-up. White students were unaffected. In Experiment 2, a value affirmation eliminated the learning-threat effect and provided evidence of psychological process. The results suggest that stereotype threat causes a form of "double jeopardy" whereby threat can undermine both learning and performance. The discussion addresses implications for the interpretation of group differences and for understanding how brief threat-reducing interventions can produce long-lasting benefits.

Keywords

stereotype threat, learning, African Americans, race, value affirmation

Received May 26, 2010; revision accepted February 17, 2011

Imagine a Black student in a large lecture hall eager to learn on the first day of class at a predominately White university. Suppose the student is aware that the class will be intellectually demanding and that performance in it will be used to determine who can handle a difficult area of study and who cannot. In this environment, the student may be keenly aware that her racial group is stereotyped as unintelligent. When the professor lectures, she may worry that any confusion she feels or any question she asks could seem to confirm these negative stereotypes. Thus, stereotype threat—a disruptive apprehension about the possibility that one might inadvertently confirm a negative stereotype about one's group (Steele, Spencer, & Aronson, 2002)—could interfere with how well the student learns the course material. She might, as a result, underperform in the class despite a high level of intellectual potential. And should she take a follow-up course the next term—even one with a more welcoming environment—she might again perform poorly if she did not fully acquire the intellectual building blocks needed to perform well in the previous class.

Hundreds of experiments show that stereotype threat undermines intellectual performance directly by causing stereotyped students to perform below their capabilities (Nguyen & Ryan, 2008; Steele et al., 2002). As a consequence, grades and test scores systematically underestimate stereotyped students' intellectual ability (Walton & Spencer, 2009). The current research investigates a broader and potentially more insidious possibility: Does stereotype threat interfere with *learning* itself? If so, stereotype threat may contribute to group differences in *developed* intellectual ability—not just intellectual performance—by preventing stereotyped students from fully acquiring academic knowledge and skills despite their level of intellectual potential. If this were the case, group differences in academic performance that emerge even in psychologically safe environments could be in part the result of stereotype threat in prior learning environments. Moreover, if stereotype threat pervades *both* learning and performance environments, it could cause cumulative performance deficits, which might together account for more variance in group differences in academic performance than now understood (see Walton & Spencer, 2009).

Although relatively little research has investigated whether stereotype threat can undermine the acquisition of academic knowledge, some work is consistent with it. One study found

[1] Princeton University, Princeton, NJ, USA
[2] Stanford University, Stanford, CA, USA

Corresponding Author:
Valerie Jones Taylor, Princeton University, Department of Psychology,
Green Hall, Princeton, NJ 08540
Email: vjones@princeton.edu

that both men and women who studied academic material as gender solos (one woman with three men or one man with three women) subsequently performed worse on a public oral exam of that material than students who studied in same-sex groups. This occurred even when solo learners performed under nonthreatening conditions, as nonsolos (Sekaquaptewa & Thompson, 2002). This study suggests that numeric underrepresentation can interfere with learning. However, this study did not examine a process rooted in negative intellectual stereotypes; for instance, the material students studied was irrelevant to gender stereotypes (it concerned facts about primates). When academic material is relevant to negative intellectual stereotypes, we suggest that only targets of the stereotype may be negatively affected (Walton & Cohen, 2003) and, furthermore, that in these circumstances even relatively subtle cues like routine evaluative instructions can induce threat and harm learning among people targeted by the stereotype.

More directly relevant is research that examines learning from a computer tutor after failure feedback (Mangels, Good, Whiteman, Maniscalco, & Dweck, 2011). In this study, after women had attempted problems described either as evaluative or as nonevaluative of math ability (threat vs. no-threat condition), they were given the opportunity to use a computer tutor for problems they had answered incorrectly. Replicating the classic stereotype threat effect, women performed worse in the evaluative condition than in the nonevaluative condition. When given a test of similar problems the next day, women in the nonevaluative condition showed a positive correlation between use of the tutor and follow-up performance, whereas women in the evaluative condition showed, if anything, a slight reversal (even though tutor use was equal across conditions). One interpretation of this result is that stereotype threat interfered with how well women learned from the tutor. Extending this research, the present studies test the effect of stereotype threat on learning directly by manipulating the presence or the absence of stereotype threat in a learning environment and then assessing performance in a subsequent nonthreatening environment.

Finally, two recent lines of research investigate the effects of stereotype threat on perceptual and rule learning. Rydell, Shiffrin, Boucher, Van Loo, and Rydell (2010) found that women in a stereotype-threat condition (e.g., reminded of gender differences in cognitive performance) showed no reduction across a series of trials in the search time needed to distinguish target Chinese characters from nontarget characters. By contrast, women in a no-threat condition showed a reduction in search time over trials, implying that they learned a more efficient search process. In addition, in research primarily investigating the role of regulatory focus in stereotype threat, Grimm, Markman, Maddox, and Baldwin (2009, Studies 2a and 2b) found that both men and women told that their gender group performed less well on a line classification task in which good performance would be rewarded (i.e., a gain-focused environment) switched more slowly than nonthreatened peers over a series of trials to a more accurate but more complex classification strategy.

Taking this past research as our starting point, we test the effects of stereotype threat on academic learning. We do so, moreover, using a two-session procedure in which students study academic material in an initial session and then are tested on that material in a subsequent session. Critically, this procedure allows us to manipulate threat in the learning and performance environments separately. In the first experimental session, students studied novel academic material in an evaluative (threatening) or a nonevaluative (nonthreatening) learning environment (a between-subjects manipulation). In the second experimental session, students' retention of half the studied academic material was assessed in a nonevaluative, nonthreatening "warm-up" and half in an evaluative, threatening "test" (a within-subjects manipulation of threat in the performance environment).

Relative to past research, this procedure offers four specific advantages. First, it clearly distinguishes between the effects of stereotype threat on learning versus on performance. Although past studies have used established indices of learning and rigorous controls, their design raises an important ambiguity. The measures of learning (e.g., speed identifying target Chinese characters, accuracy in a line classification task) are also measures of performance, and furthermore, the manipulation of stereotype threat directly targets this performance. It is possible that deficits in performance result from the direct effect of stereotype threat on performance, with no effect on learning. Second, unlike past research, the two-session procedure tests the effects of stereotype threat on retention, a core index of learning. Third, this procedure tests the cumulative effects of stereotype threat in learning and performance environments. Fourth, this approach is relatively ecologically valid, both because it assesses academic learning and because the two-session procedure more closely mimics learning in real-world academic environments where students study academic material and are later tested on their retention of that material.

Why would stereotype threat interfere with learning? Past research suggests that the psychological consequences of stereotype threat—such as maladaptive levels of arousal, negative emotion regulation, cognitive depletion, and a prevention focus (see Schmader, Johns, & Forbes, 2008)—can undermine learning as well as performance (e.g., excessive arousal: Hasher & Zacks, 1979; decreased working memory capacity: Rosen & Engle, 1997). Like the effects of stereotype threat on performance, we assume that multiple processes contribute to the negative effects of stereotype threat on learning and that these processes interact in complex ways (see Schmader et al., 2008). Here we examine mechanisms that may *improve* learning in otherwise threatening environments. Experiment 2 tests whether change in two processes linked to stereotype threat—a decrease in stereotype suppression (Logel, Iserman, Davies, Quinn, & Spencer,

2009) and an increase in promotion focus (Grimm et al., 2009; Seibt & Förster, 2004)—facilitate learning when personally important aspects of the self have been affirmed in an otherwise threatening environment (cf. Sherman & Cohen, 2006).

In investigating the effect of stereotype threat on learning, the present research addresses two additional important questions. First, how does stereotype threat in learning and performance environments relate? Although a psychologically safe learning or performance environment might compensate for threat in the other environment, we suggest that threat in either environment will lead to poor performance. If so, stereotyped students may experience a form of "double jeopardy" whereby threat interferes with both the acquisition of academic knowledge (in learning environments) and the retrieval of acquired knowledge (in performance environments).

Second, a consideration of the effect of stereotype threat on learning may shed light on a pressing question in contemporary stereotype threat research: How do brief interventions to reduce stereotype threat raise students' academic performance months and years later? For instance, Cohen and colleagues found that a brief value-affirmation intervention designed to reduce the threat associated with being negatively stereotyped in school improved the classroom performance of Black seventh graders over the next 2 years of middle school (Cohen, Garcia, Apfel, & Master, 2006; Cohen, Garcia, Purdie-Vaughns, Apfel, & Brzustoski, 2009). Walton and Cohen (2007, 2011) found that a brief intervention to buttress Black students' sense of social belonging in college in the face of negative stereotypes raised their GPA from sophomore through senior year, cutting the racial achievement gap in half. The long-term gains produced by both interventions seem to involve recursive processes, whereby better school performance at one time point facilitates better performance at the next time point (Cohen et al., 2009; Walton & Cohen, 2011).

A critical question involves the nature of these recursive processes. Past research emphasizes social and psychological processes; for instance, students' sense of social and academic fit in school may improve over time in a self-reinforcing manner after an early intervention, changing the trajectory of their academic achievement (Cohen et al., 2009; Walton & Cohen, 2007; also see Mendoza-Denton, Downey, Purdie, Davis, & Pietrzak, 2002). We suggest that recursive intellectual processes also contribute to long-lasting gains. A threat-reducing intervention delivered early in an academic environment could help stereotyped students acquire academic skills and knowledge needed to perform well later (see Rydell et al., 2010). To explore this possibility, Experiment 2 tests whether a value affirmation improves learning in the face of stereotype threat. In addition, as noted, this study investigates psychological processes by which a value affirmation may facilitate learning in an otherwise threatening academic environment.

Overview of Studies

Two experiments tested the learning-threat hypothesis. Experiment 1 tested whether stereotype threat impaired Black students' but not White students' learning of new academic material. Experiment 2 tested whether a value affirmation would eliminate the effect of stereotype threat on learning among Black students and examined mediating processes.

Both studies adopt a classic learning–recall paradigm (Anderson & Bower, 1972), where participants study novel material and later recall and then indicate their recognition for that material. In the present research, participants studied the definitions of rare English-language words in either a threatening (i.e., evaluative) or a nonthreatening (i.e., nonevaluative) learning environment (Experiment 1) or in a threatening learning environment following either a value-affirmation or a control exercise (Experiment 2). Then, in a second session 1 to 2 weeks later, participants' memory for the word definitions studied was assessed. First, in a nonthreatening, nonevaluative task called a "warm-up," participants recalled the definitions of half the words (similar to a cued recall task) and then matched these words to their definitions (similar to a source recognition task). Second, in a threatening, ostensibly evaluative "test," participants completed the recall and matching tasks for the remaining words. This design varies some from classic learning studies in which participants typically study a list of words and are tested for their recall and recognition of those words in a single session. This two-session procedure allows us to manipulate threat in the learning and performance environments separately and better captures students' experience in real-world school settings studying and later being tested on academic material.

The recall task was expected to be especially difficult as students had to retrieve word definitions from memory a week after studying them (cf. Anderson & Bower, 1972). As stereotype threat undermines performance most on challenging tasks (Spencer, Steele, & Quinn, 1999), the primary test of the learning-threat hypothesis involved the effect of learning-threat on participants' recall performance on the nonthreatening "warm-up." If threat in the learning environment produced worse performance among Black students in this nonthreatening environment, it would reflect nonoptimal learning. The "test" was included to provide a conceptual replication of past research examining the effects of stereotype threat in performance environments (e.g., Steele & Aronson, 1995) and to examine the cumulative effects of stereotype threat in learning and performance environments. Insofar as the matching task was also challenging, patterns of performance on this task were expected to be in the same direction although weaker. Importantly, the recall task was designed not to be so difficult as to produce floor effects; if it did, the matching task would provide a better test of the hypothesis.

Experiment 1

Experiment 1 tested whether the presence of stereotype threat in a learning environment would impair how well Black students but not White students learned new academic material.

Method

Design and participants. The experiment used a 2 × 2 × 2 mixed-model design. The factors were participant race (Black, White), learning-threat condition (learning-no-threat, learning-threat), and performance-threat condition (performance-no-threat, performance-threat) with the final factor within subjects. In exchange for $10 after each session, 32 Black and 44 White students (46 women) participated in two sessions. Data from one participant (Black female) were excluded as she failed to follow the instructions.

Materials. A total of 24 rare English-language words were selected from the 2003 Scripps Howard National Spelling Bee List and from the *Oxford English Dictionary* (*OED*). A group of 20 undergraduates was asked to define each word. Words that were defined correctly by 20% or more of respondents were dropped and replaced with additional words from the same sources. A second group of 15 undergraduates was asked to define each word in the revised set. No word was defined correctly by more than 20% of respondents. The final set of 24 words was divided into two lists, which, in the second experimental session, were presented separately to participants in the "warm-up" and "test." Word List A contained *canton, ephrasy, glabella, gladiolus, insouciant, ofclepe, prosody, rood, schappe, succedaneum, usufruct*, and *viscid*. Word List B contained *albumen, bedizen, canard, diplacusis, ecumene, foulard, hyan, inspeximus, panegyry, serried, soubrette*, and *syllepsis*. Word definitions were adopted from the *OED*.

Procedure. Students participated individually in two sessions 6 to 13 days apart. In Session 1, students studied the definitions of 24 rare words under either learning-threat or learning-no-threat conditions. In the learning-threat condition, the word-learning task was described so as to be relevant to negative intellectual stereotypes about African Americans, a portrayal that triggers stereotype threat (Steele & Aronson, 1995). Students were told that the study investigated "how well people from different backgrounds learn," that "different people learn differently and we are interested in how well you learn and retain new information," and that the task would provide "a genuine assessment" of students' "learning abilities and limitations." By contrast, in the learning-no-threat condition, the word-learning task was described so as to be irrelevant to intellectual stereotypes, a portrayal shown to not trigger stereotype threat. Students were told that the study investigated "psychological factors that contribute to different learning styles," that "different people learn differently, even though people typically retain the same amount of information," and that the task would provide "a genuine sense" of students' "personal learning style."

In both conditions, students were told the task would be difficult and were urged "to give a strong effort." Students were then given 10 minutes to study the definitions of the 24 words. After doing so, students indicated which words if any they had known prior to the study. Finally, they were thanked and dismissed.

One to two weeks later, students returned to the same laboratory room and their recall of the word definitions was assessed.[1] First, students completed two "warm-up" tasks recalling and then matching the definitions of 12 of the 24 words (performance-no-threat condition). To create a nonthreatening setting, the "warm-up" tasks were said to "help familiarize you with the recall tasks you will complete later" and that their purpose was just to get students "warmed up." Students were given 7 minutes to free recall the definitions of 12 words and a second 7 minutes to match each word to its definition. The matching exercise was made more difficult with the inclusion of 6 filler definitions of different words from the same sources described above. Students were urged "to give a strong effort" ostensibly to help the researchers understand their "learning style."

Next students completed two "tests" (performance-threat condition). The tests were said to include "carefully selected words" that would "evaluate your ability to learn verbal information and your performance on problems requiring verbal reasoning ability" and to "provide a genuine test of your verbal abilities and limitations." These instructions were designed to elicit stereotype threat by making negative intellectual stereotypes about African Americans seem relevant. Students were given 7 minutes to free recall the definitions of the remaining 12 words and a second 7 minutes to match each word to its definition (6 new filler definitions were again included). Again students were urged "to give a strong effort," but here this was ostensibly to help the researchers assess their "verbal ability."[2] Which word list (A or B) was included in the "warm-up" versus "test" was counterbalanced across learning-threat condition.

Finally, students reported demographic information for use as potential covariates: their SAT Verbal score, family socioeconomic status (SES; 1 = *working class*, 5 = *upper class*), year in school, and gender. Students were then debriefed and dismissed.

Measures of recall and matching performance. Recall performance was independently graded by two coders blind to students' race and condition. Definitions were scored as correct if the student's definition reflected the basic definition presented in Session 1, even if minor details were omitted or the specific wording differed. No partial credit was given. Interrater reliability was high, Cohen's Kappa = .90. Disagreements were resolved through discussion. The number of words participants accurately recalled (which they did not indicate they knew previously) ranged from 0 to 6.

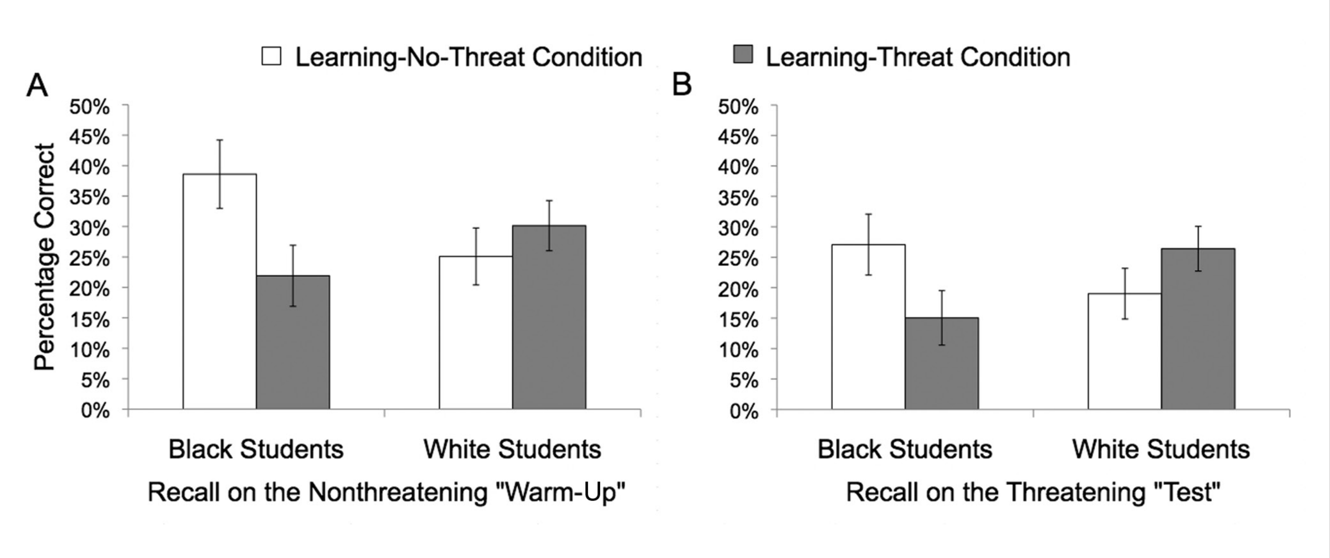

Figure 1. Percentage of rare words defined correctly by participant race, learning-threat condition, and performance-threat condition. for (A) the nonthreatening "warm-up" and (B) the threatening "test" (Experiment 1)
Means are adjusted for participants' family socioeconomic status. Error bars represent ±1 standard error.

To account for words students had known prior to the study, we removed words both that students indicated they knew previously (in Session 1) and that they defined correctly (in Session 2). Recall performance was calculated as the proportion of the remaining words students defined correctly in each performance-threat condition. Matching performance was calculated in the same manner (i.e., removing words both that students indicated they knew previously and that they matched correctly). Analyzing the number of definitions recalled and matched correctly statistically adjusting for the number of words known previously and which students defined or matched correctly yields nearly identical results.

Results

Preliminary analyses. Recall and matching performance were analyzed in a series of analyses of covariance (ANCOVAs). SAT Verbal score, family SES, year in school, the number of days between Sessions 1 and 2, and gender were tested as covariates. In all analyses, only family SES proved to be significant and so was retained in analysis. Students from low-SES backgrounds underperform relative to peers from higher SES backgrounds (National Center for Education Statistics, 1999); thus, SES was considered relevant. The results were not moderated by gender, so analyses collapse across this variable.

Performance did not differ between the two sets of words (Word Lists A and B) and in general word list did not moderate the results. Specifically, analyses of (a) mean recall and matching performance within performance-threat condition by word list and (b) interactions between word list and participant race, learning-threat condition, performance-threat condition, and higher order interactions yielded only one marginal effect.[3] Thus, analyses collapse across word list.

The number of words participants indicated that they had known previously was low (M_{grand} = 1.56 out of 24) and did not vary by participant race, $F < 2.54$, $p > .10$; learning-threat condition, $F < 1$; or their interaction, $F < 1$.

Recall performance on the nonthreatening "warm-up." The key test of the learning-threat hypothesis involved recall performance on the "warm-up." In this nonthreatening performance setting, did Black students in the learning-threat condition perform worse than Black students in the learning-no-threat condition? They did. Analysis yielded only the predicted Race × Learning-Threat Condition interaction, $F(1, 70) = 4.97$, $p = .029$, $\eta_p^2 = .07$. As displayed in Figure 1a, Black students defined approximately half as many words correctly in the learning-threat condition as in the learning-no-threat condition, $t(70) = 2.32$, $p = .023$, $d = 0.83$. White students showed no condition effect, $t < 1$. Cross-race comparisons found that Black students, if anything, defined fewer words correctly than did White students in the learning-threat condition, $t(70) = 1.27$, $p = .21$, but more words correctly in the learning-no-threat condition, $t(70) = 1.99$, $p = .052$, $d = 0.68$.[4]

Matching performance on the nonthreatening "warm-up." Analysis of matching performance on the "warm-up" yielded only the Race × Learning-Threat Condition interaction, $F(1, 70) = 4.56$, $p = .036$, $\eta_p^2 = .06$. There was a trend for Black students to match fewer words correctly in the learning-threat condition (M_{adj} = 0.60, *SD* = 0.23) than in the learning-no-threat condition (M_{adj} = 0.72, *SD* = 0.25), $t(70)$ = 1.37, *p* = .18. White students showed the opposite pattern (Learning-No-Threat: M_{adj} = 0.58, *SD* = 0.24; Learning-Threat: M_{adj} = 0.71, *SD* = 0.24), $t(70)$ = 1.81, *p* = .074, *d* = 0.55. In the learning-threat condition Black students tended to perform worse than Whites, $t(70)$ = 1.45, *p* = .15, but in the learning-no-threat condition they tended to perform better than Whites, $t(70)$ = 1.68, *p* = .097, *d* = 0.57.

Recall performance on the threatening "test." Analysis of the proportion of words defined correctly on the "test" yielded only the same Race × Learning-Threat Condition interaction, $F(1, 70)$ = 4.99, *p* = .029, η_p^2 = .07. As displayed in Figure 1b, Black students defined marginally fewer words correctly in the learning-threat condition than in the learning-no-threat condition, $t(70)$ = 1.88, *p* = .064, *d* = 0.68. White students showed the opposite pattern, $t(70)$ = 1.37, *p* = .18. Cross-race comparisons found that Black students defined fewer words correctly than did White students in the learning-threat condition, $t(70)$ = 1.97, *p* = .052, *d* = 0.64, but, if anything, more words correctly in the learning-no-threat condition, $t(70)$ = 1.32, *p* = .19.

Matching performance on the threatening "test." Analysis of matching performance on the "test" yielded only a marginal Race × Learning-Threat Condition interaction, $F(1, 70)$ = 3.58, *p* = .063, η_p^2 = .05. The condition difference for Black students was not significant (Learning-Threat: M_{adj} = 0.55, *SD* = 0.22; Learning-No-Threat: M_{adj} = 0.64, *SD* = 0.24), $t <$ 1.15. White students performed marginally better in the learning-threat condition (M_{adj} = 0.69, *SD* = 0.23) than in the learning-no-threat condition (M_{adj} = 0.57, *SD* = 0.24), $t(70)$ = 1.70, *p* = .093, *d* = 0.51. Cross-race comparisons found only that, in the learning-threat condition, Black students performed marginally worse than White students (M_{adj} = 0.55, *SD* = 0.22 vs. M_{adj} = 0.69, *SD* = 0.23), $t(70)$ = 1.88, *p* = .064, *d* = 0.61.

The cumulative effects of stereotype threat in learning and performance environments. As shown in Figure 1, Black students' recall performance dropped by 60% from when they both learned and performed in nonthreatening conditions to when they both learned and performed in threatening conditions.

To test the effect of performance-threat on Black students' recall performance within each learning-threat condition, we conducted a three-way mixed-model ANOVA involving race (between subjects), learning-threat condition (between subjects), and performance-threat condition (within subjects) on residual recall performance adjusted for family SES. We then calculated a priori specified contrasts.[5] In the learning-no-threat condition, Black students recalled 29% fewer definitions on the "test" than on the "warm-up," $t(71)$ = 2.66, *p* = .010, *d* = 0.97. Even in the learning-threat condition where Black students' performance on the "warm-up" had already been depressed, they recalled 31% fewer definitions on the "test" than on the "warm-up," $t(71)$ = 1.73, *p* = .087, *d* = 0.61.

Discussion

Experiment 1 provides direct evidence that stereotype threat can undermine academic learning. Black and White students studied rare words in an ostensibly evaluative, threat-inducing learning environment or in a nonevaluative learning environment. One to two weeks later, Black students who had studied in the evaluative environment defined approximately half as many words correctly in a nonthreatening performance setting. White students showed no such decrement. Notably, the Black students in this study had considerable intellectual potential. When they both learned and performed in nonthreatening environments, they were the best performing group. But when they both learned and performed in threatening conditions, they were the worst performing group. The results illustrate the large cumulative impact stereotype threat can have when it pervades both learning and performance environments. Furthermore, the results suggest that Black students experience a form of "double jeopardy." When they experienced stereotype threat in *either* the learning or the performance environment their recall performance suffered. Threat can cause poor performance by interfering with both the acquisition of academic knowledge and its retrieval. Finally, consistent with research on threat effects on performance (Spencer et al., 1999), effects were most evident on the difficult recall task. The matching task yielded similar but weaker effects.

The results of Experiment 1 suggest that nonthreatening academic environments may facilitate the acquisition of new academic knowledge among stereotyped students. If so, one way interventions that reduce stereotype threat—such as value-affirmation and social-belonging interventions (Cohen et al., 2009; Walton & Cohen, 2007)—improve academic outcomes over long periods of time may be by helping stereotyped students acquire academic skills and knowledge needed to perform well later in school more effectively. To test this hypothesis, Experiment 2 manipulated whether Black students completed a value affirmation in a threatening learning environment and assessed whether this would improve their learning.

Experiment 2

Experiment 2 had two primary purposes. The first was to test whether a value affirmation would eliminate the negative effect of stereotype threat on learning. Past research shows that when people reflect on personally important values they experience less psychological threat and stress (Sherman & Cohen, 2006). If so, completing a value affirmation in a threatening learning environment might reduce the experience of threat among stereotyped students and help them acquire new information more effectively (cf. Cohen et al., 2006, 2009; Martens, Johns, Greenberg, & Schimel, 2006). Laboratory and field-experimental research suggests that timing can be important in the delivery of value affirmations. Affirmations may be more effective when delivered proximal to but before a threatening experience (Cohen et al., 2009; Critcher, Dunning, & Armor, 2010; cf. Rydell et al., 2010). Therefore, in the current study, participants completed a value affirmation prior to the induction of threat in the learning environment.

Second, Experiment 2 explored psychological processes that contribute to the effect of stereotype threat on learning and, in particular, that may mediate the effect of the value affirmation. We examined two potential mediators. First was stereotype suppression (Logel et al., 2009). Stereotype threat may cause people to effortfully suppress thoughts of the stereotype, which can consume cognitive resources and contribute to performance decrements. Such suppression could also interfere with learning. In reducing stress and threat, a value affirmation could reduce the need to suppress the stereotype and improve learning. Second, was regulatory focus (Higgins, 1998). Research shows that stereotype threat can induce a prevention focus where people vigilantly strive to prevent negative outcomes rather than to promote positive outcomes. This prevention focus can undermine performance in gain-focused testing environments where students aim to maximize performance (Grimm et al., 2009; also see Seibt & Förster, 2004). In learning contexts where students are trying to acquire knowledge, a prevention focus may be especially maladaptive. Insofar as a value affirmation helps people cope with threat, it may allow students to adopt more of a promotion focus and so learn more effectively.

Measures of stereotype accessibility (to index stereotype suppression) and regulatory focus were assessed immediately after participants studied word definitions in the first session. Because stereotype accessibility was assessed before any test of the studied material, we expected to observe suppression effects in the no-affirmation condition (slower identification of stereotype-relevant words in a lexical decision task) rather than a rebound or stereotype activation effects (faster identification of stereotype-relevant words), which can occur after people complete threatening tests (Logel et al., 2009). Notably, in examining these processes in the context of a value affirmation, Experiment 2 investigates processes that may *improve* learning and performance among stereotyped students in the face of threat rather than processes that contribute to underperformance. Although these processes are mutually informative and both important, processes that improve stereotyped students' outcomes despite threat are less studied in past research and less well understood.

Method

Design and participants. Experiment 2 used a 2 × 2 mixed-model design with affirmation condition (affirmation, no affirmation) as the between-subjects factor and performance-threat condition (performance-threat, performance-no-threat) as the within-subjects factor. All participants were Black, and all studied the definitions of rare words in the learning-threat condition described in Experiment 1.

Participants were 36 Black undergraduates who received $10 after each session. Data from 7 participants were excluded: One expressed suspicion of the learning-threat instructions during a systematic funnel-debriefing procedure after Session 2, and 6 did not provide demographic information (i.e., SAT scores) needed to analyze data (2 affirmation condition, 4 no-affirmation condition). Analyses are thus based on 29 participants (16 women).

Procedure. Experiment 2 used the same two-session procedure described in Experiment 1, with several additions to the first session. First, before studying word definitions, students were given 10 minutes to complete a writing exercise that served as the value-affirmation manipulation (Sherman & Cohen, 2006). Students either circled their most important value from a brief list of values and wrote about why that value was important to them (affirmation condition) or circled their least important value and wrote about why it might matter to someone else (no-affirmation condition). In all, 12 values were listed (e.g., relations with friends and family, religious values, sports ability).

The word-learning task was described to students as in the learning-threat condition in Experiment 1—that is, as evaluative of learning ability—thus representing this task as relevant to negative intellectual stereotypes about African Americans. Students then had 10 minutes to study the definitions of the same 24 rare words used in Experiment 1.

Next students completed a lexical decision task to assess stereotype suppression and a questionnaire assessing regulatory focus. Students then indicated which words if any they had known prior to the study. Finally, they were thanked and dismissed.

Session 2 was identical to this session in Experiment 1. From 4 to 9 days after Session 1, students returned to the same laboratory room, defined and then matched 12 of the words in two 7-minute "warm-ups," and then defined and matched the remaining 12 words in two 7-minute "tests" said to "evaluate verbal learning ability." The words included in the "warm-up" versus "test" were counterbalanced across affirmation condition. Finally, students were asked to provide the same demographic information as in Experiment 1 and were debriefed and dismissed.

Measures of recall and matching performance. As in Experiment 1, recall performance was graded by two independent coders blind to participants' condition. Interrater reliability was high, Cohen's Kappa = .93. Disagreements were resolved through discussion. The number of words participants accurately recalled (which they did not know previously) ranged from 0 to 9. The measures of recall and matching performance were the same as in Experiment 1 (i.e., the proportion of words participants had not known prior to the study which they recalled and matched correctly).

Stereotype suppression. A lexical decision task was used to assess suppression of negative racial stereotypes. Students indicated whether letter strings presented on a computer screen were words or nonwords as quickly and accurately as possible (Logel et al., 2009). In all, 38 words and 19 nonwords were presented. Of these, 9 words were relevant to negative stereotypes about African Americans (i.e., *aggressive, bias, class, dumb, inferior, lazy, minority, poor, welfare*; see Steele & Aronson, 1995). Following research showing that the activation of negative racial stereotypes and of positive or neutral thoughts about race can be separable (Walton & Cohen, 2007), we also included words relevant to positive or neutral representations of African Americans (i.e., *black, color, music, race, soul*). In addition, we included failure-related words (i.e., *lose, flunk, fail, shame, weak*). An additional 19 words were neutral in valence and irrelevant to race and to racial stereotypes (e.g., *blind*). Irrelevant words were matched to target words in length and frequency of use (Kucera & Francis, 1967). Nonwords were matched in length and in first letter to a real presented word. In the task, a fixation cross appeared for 500 ms followed by the word or nonword. Each letter string remained on the screen until participants indicated whether it was a word or nonword. The order of presentation was randomized for each participant. Incorrect responses and outliers (<5% of trials) were removed following standard procedures (Van Selst & Jolicoeur, 1994). The mean response time to each category of word was then calculated.

Regulatory focus. Students completed an 18-item questionnaire that assessed level of promotion focus (10 items; e.g., "Right now, I am thinking about how I will achieve academic success") and level of prevention focus (8 items; e.g., "Right now, I am focused on preventing negative events in my life"; 1 = *not at all true of* me, 9 = *very true of me*; Lockwood, Jordan, & Kunda, 2002). Both subscales were examined using factor analyses. One item on each subscale did not load on the first factor (loadings <.45) and so was dropped. The remaining items formed a reliable composite for promotion focus, $\alpha = .87$, and for prevention focus, $\alpha = .75$, and so were averaged.

On the lexical decision task, data from three participants were not recorded because of a computer error and were thus lost. On the measure of regulatory focus two participants had mean scores of 1 on both subscales, meaning that they provided the lowest possible response to every item. Under the assumption that these participants did not complete the post-studying materials seriously, they were excluded from analyses of the lexical decision and regulatory focus data.

Results

Preliminary analyses. The same covariates tested in Experiment 1 were tested in Experiment 2 in analyses of recall and matching performance on the "warm-up" and "test." In all analyses, only SAT Verbal score and the number of days between Sessions 1 and 2 were significant and so were retained in analyses.[6] The results were not moderated by gender, so analyses collapse across this variable. As in Experiment 1, performance on the two sets of words (Word Lists A and B) was similar and there was no consistent pattern of moderation, so analyses collapse across word list. Finally, the number of words known previously did not vary by affirmation condition, $F < 1$.

Recall performance on the nonthreatening "warm-up." The key outcome was recall performance on the nonthreatening "warm-up." Did Black students who had been affirmed in a threatening learning environment perform better than Black students who had not been affirmed? They did. The effect of Affirmation Condition was significant, $F(1, 25) = 4.77$, $p = .039$, $d = 0.83$. As displayed in Figure 2a, Black students defined more words correctly on the "warm-up" in the affirmation condition than in the no-affirmation condition.

Matching performance on the nonthreatening "warm-up." Analysis of matching performance on the "warm-up" yielded the same effect of Affirmation Condition (Affirmation: $M_{adj} = 0.56$, $SD = 0.18$; No Affirmation: $M_{adj} = 0.41$, $SD = 0.18$), $F(1, 25) = 5.24$, $p = .031$, $d = 0.87$.

Performance on the threatening "test." The patterns of means for recall and matching performance on the threatening "test" were in the same direction as on the "warm-up," but the effect of Affirmation Condition on both outcomes was nonsignificant, Fs < 1.

The cumulative effects of stereotype threat in learning and performance environments. Overall, Black students who had been affirmed in the threatening learning environment and who performed on the nonthreatening "warm-up" performed nearly 70% better on the recall tasks than Black students who had not been affirmed and who performed on the threatening "test." Examined differently, the combination of threat in the learning environment (i.e., no-affirmation condition) and in the performance environment (i.e., on the "test") caused Black students' recall performance to drop by 41%.

As in Experiment 1, we conducted a mixed-model ANOVA involving affirmation condition (between subjects) and performance-threat condition (within subjects) on residual recall performance adjusted for the aforementioned covariates and then calculated a priori specified contrasts.[7] In the affirmation condition, Black students' recall performance dropped 37% from the "warm-up" to the "test," $t(27) = 3.25$, $p = .003$, $d = 1.23$. In the no-affirmation condition, there was no difference in performance between the "warm-up" and the "test," $t < 1$.

Stereotype suppression. We tested the effect of condition on participants' mean response time to (a) negative racial-stereotype-related words, (b) positive or neutral race-related words, and (c) failure-related words in separate ANCOVAs controlling for response time to irrelevant words. Because there was a trend for the covariate to be skewed, it was log transformed prior to analysis.

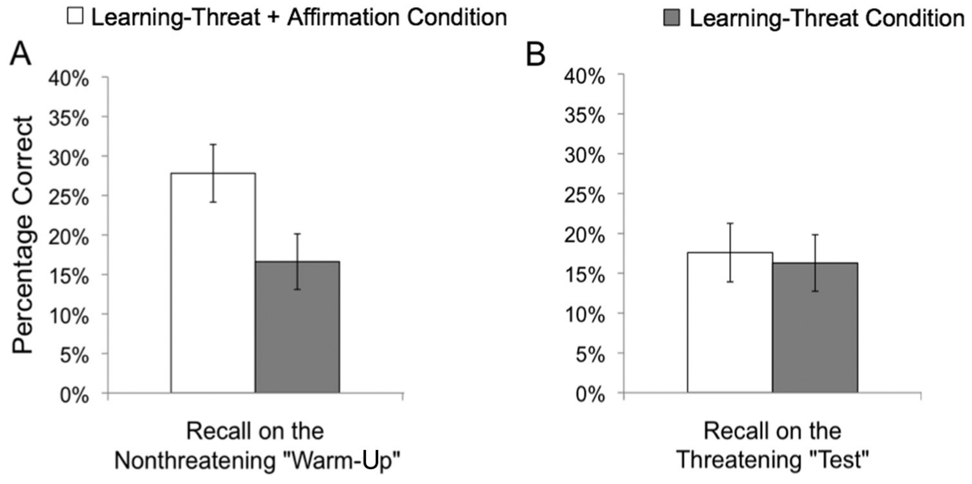

Figure 2. Percentage of rare words Black students defined correctly by affirmation condition and performance-threat condition on (A) the nonthreatening "warm-up" and (B) the threatening "test" (Experiment 2).
Means are adjusted for SAT Verbal score and for the number of days between Session 1 and Session 2. Error bars represent ±1 standard error.

On the key category of negative racial-stereotype-related words, the effect of Affirmation Condition was marginally significant, $F(1, 21) = 3.97$, $p = .059$, $d = 83$. Consistent with the hypothesis that affirmation reduced stereotype suppression, participants identified negative racial stereotype words more quickly in the affirmation condition ($M_{adj} = 621.27$ ms, $SD = 41.93$) than in the no-affirmation condition ($M_{adj} = 655.87$ ms, $SD = 41.92$).

Interestingly, analysis of positive or neutral race-related words also yielded a marginal effect of Affirmation Condition, $F(1, 21) = 3.47$, $p = .077$, $d = 0.77$. Participants identified positive or neutral race-related words more slowly in the affirmation condition ($M_{adj} = 657.77$ ms, $SD = 66.68$) than in the no-affirmation condition ($M_{adj} = 606.36$ ms, $SD = 66.67$). Although not reported previously, this finding suggests that, under threat and in the absence of affirmation, Black students suppressed negative stereotypes about their group and simultaneously activated positive representations.

Analysis of failure-related words yielded no condition effect, $F < 1.40$, *ns*.

Finally, we examined the relationships between response time to each category of race-related words and recall performance on the "warm-up." These analyses use residual response time to target words controlling for response time to irrelevant words and residual "warm-up" recall performance controlling for its relevant covariates. Consistent with a suppression process, slower response time to negative racial-stereotype words predicted worse recall performance, $r(20) = -.53$, $p = .011$. However, response time to positive or neutral race-related words was unrelated to performance, $r(20) = .01$, *ns*.

Regulatory focus. Promotion and prevention focus were analyzed in separate *t*-tests. Participants reported higher levels of promotion focus in the affirmation condition ($M = 7.87$, $SD = 1.07$) than in the no-affirmation condition ($M = 6.70$, $SD = 1.21$), $t(24) = 2.54$, $p = .018$, $d = 0.91$. There was no effect of condition on prevention focus, $t < 1$. Correlational analyses found that promotion focus predicted better recall performance, $r(22) = .51$, $p = .012$, but no relationship between prevention focus and recall performance, $r(22) = .14$, *ns*.

Mediation analyses. Did stereotype suppression and/or promotion focus mediate the effect of Affirmation Condition on participants' recall performance on the nonthreatening "warm-up"? The data suggest they did. We regressed Affirmation Condition and each candidate mediator on recall performance on the "warm-up" in separate analyses (SAT Verbal score and the number of days between Sessions 1 and 2 were also included as covariates). As displayed in Table 1, in both analyses the candidate mediator was significant, Affirmation Condition dropped to nonsignificance, and this reduction was significant. We also conducted a regression including both candidate mediators. Again Affirmation Condition was nonsignificant, but both stereotype suppression and promotion focus were separately predictive. The results suggest that affirmation improved participants' learning both by preventing stereotype suppression and by encouraging a promotion focus.

Table 1. Mediation Analyses in Experiment 2

	Regression 1: Condition effect on “warm-up” recall performance		Regression 2: Mediation by stereotype suppression		Regression 3: Mediation by promotion focus		Regression 4: Mediation by stereotype suppression and promotion focus	
Predictor	β	*t*	β	*t*	β	*t*	β	*t*
Covariates								
SAT Verbal score	.50	3.21*	.58	3.65*	.56	3.58*	.63	4.20*
Days between sessions	–.30	–1.97†	–.37	–2.36*	–.30	–2.01†	–.35	–2.35*
1. Value-affirmation condition	–.34	–2.18*	–.14	–0.86	–.11	–0.66	.04	0.19
2. Response time to negative racial stereotype words			–.38	–2.22*			–.34	–2.08*
3. Promotion focus					.36	2.13*	.35	1.91†
R^2	.43		.56		.54		.64	
Asymmetric distribution of products test 95% confidence interval (MacKinnon, Lockwood, Hoffman, West & Sheets, 2002)	—		–.018 to –.089*		–.023 to –.094*		–.015 to –.080*(2)	
							–.019 to –.093*(3)	

In each analysis, the criterion is “warm-up” recall performance. To control for individual differences in processing speed, response time to negative racial stereotype words is a residual controlling for response time to words irrelevant to race or racial stereotypes. The asymmetric distribution of products test assesses the significance of the reduction in the condition effect. If the 95% confidence interval does not include zero, the effect is significant.
$^{\dagger}p < .10$. $*p < .05$.

Discussion

In Experiment 2, a value affirmation eliminated the negative effect of stereotype threat on learning among Black students. Black students who completed a value affirmation before studying rare words in a threatening learning environment defined more words correctly on the nonthreatening “warm-up” a week later than did Black students who had studied in the same setting without having completed the value affirmation. The results suggest that one way value-affirmation interventions may raise stereotyped students’ academic performance is by improving learning in otherwise threatening environments. Furthermore, the results suggest *how* the value affirmation improved learning in the face of threat: It forestalled stereotype suppression and increased promotion focus.

General Discussion

Two experiments provide direct evidence that stereotype threat undermines academic learning. In Experiment 1, Black and White students studied rare words in either a threatening or a nonthreatening learning environment. One to two weeks later, in a nonthreatening performance setting (called a “warm-up”), Black students who had studied under threat defined just half as many words correctly as Black students who had studied without threat. White students showed no such deficit. In Experiment 2, Black students completed a value affirmation designed to reduce stress and threat before studying rare words in the threatening learning environment. This exercise forestalled decrements in learning. Approximately 1 week later, Black students who had completed the value affirmation defined more words correctly on the “warm-up” than peers who had completed a control exercise.

In both studies, when Black students performed in a threatening setting (i.e., on a “test”), even those who had studied without threat performed worse. The results suggest that stereotype threat causes a form of “double jeopardy”—it both interferes with how well stereotyped students learn new material and prevents stereotyped students from performing well on material they have learned well. In both studies, the combination of threat in the learning and performance environments caused large cumulative declines in performance—Black students’ recall performance dropped 60% and 41% in the two studies when they both learned and performed in threatening conditions as compared to nonthreatening conditions.

Comparing the Effect of Stereotype Threat on Learning and on Performance

These results provide direct evidence that stereotype threat undermines academic learning (also see Grimm et al., 2009; Mangels et al., 2011; Rydell et al., 2010; Sekaquaptewa & Thompson, 2002). This finding raises a series of important questions for future research. One set of questions involves the similarities and differences between the effects of stereotype threat on learning and on performance. In several respects, the present results point to broad similarities between the two. First, in both contexts, stereotype threat undermines outcomes only for students targeted by the negative stereotype. In Experiment 1, if anything, White students performed better when they had learned in the threatening environment, a finding consistent with research on stereotype lift (Walton & Cohen, 2003). Second, in both contexts, stereotype threat undermines performance most on challenging academic material (Spencer et al., 1999); here, the learning-threat effect was more evident on the difficult recall task than on the easier matching task. Third, in both environments routine evaluative instructions trigger stereotype threat (Steele & Aronson, 1995) and value affirmations reduce threat (Cohen et al., 2006, 2009; Martens et al., 2006).

A fourth area of similarity involves psychological process. In both contexts, stereotype suppression (Logel et al., 2009) and a change in regulatory focus (Grimm et al., 2009; Seibt & Förster, 2004) seem to contribute to the effects of stereotype threat. Experiment 2 is relatively novel in investigating processes by which a strategy to mitigate threat (i.e., affirmation) improves outcomes in an otherwise threatening environment, namely, by reducing stereotype suppression and increasing promotion focus. This finding suggests the psychological processes by which value-affirmation interventions may improve real-world academic outcomes (e.g., Cohen et al., 2006, 2009). However, an important direction for future research is to test whether these processes mediate the effects of a threatening learning environment itself (i.e., relative to a nonthreatening learning environment), and how other processes linked to stereotype threat contribute to decrements in learning (e.g., excessive levels of arousal, deficits in working memory; Schmader et al., 2008).

Another question involves the triggers of stereotype threat. Research has identified a wide range of individual differences and situational cues that trigger stereotype threat in performance environments, both shedding light on the nature of this threat and suggesting effective remedies. Do similar individual differences and situational cues lead to worse learning? Are students who are personally identified with a performance domain also more vulnerable to stereotype threat in learning environments (Aronson et al., 1999)? Do cues such as the request to report one's group identity (Steele & Aronson, 1995) or being in the numeric minority (Inzlicht & Ben-Zeev, 2000) also elicit stereotype threat in learning environments?

Implications for Group Differences and for Intervention

If stereotype threat undermines learning, it may contribute to group differences in *developed* intellectual ability. Moreover, the finding that the mere evaluative quality of a learning environment—even with no explicit mention of groups or group differences—can impair learning suggests that the effects of stereotype threat on learning may be pervasive. Insofar as an evaluative quality is implied or perceived in many learning environments, stereotype threat may contribute substantially more to group differences in academic performance than is now understood (see Walton & Spencer, 2009). Furthermore, whereas stereotype threat in performance environments can be remedied through simple framing manipulations (e.g., Steele & Aronson, 1995), if stereotype threat interferes with learning, just presenting performance opportunities in less threatening ways will be insufficient. Optimal performance will also require creating nonthreatening learning environments.

In addition, the present research provides insight into how stereotype threat and stigmatization affect students' academic performance over time and how brief interventions to reduce stereotype threat could generate long-lasting improvements in academic achievement (Cohen et al., 2006, 2009; Walton & Cohen, 2007, 2011). In general, past research emphasizes how poor performance and negative social and psychological processes may build over time in recursive cycles (Cohen et al., 2009). For instance, with repeated exposure stereotype threat can cause disidentification, whereby students detach their sense of self-worth from academic tasks and disengage from school (e.g., Osborne, 1997). Students who anxiously anticipate rejection on the basis of their race (Mendoza-Denton et al., 2002) or who feel uncertain of their social belonging in school (Walton & Cohen, 2007) may interpret negative social events in school as evidence of their lack of belonging in general, which may undermine motivation and performance.

Complementing this research, the present findings point to a recursive intellectual process. Stereotype threat may prevent students from acquiring intellectual building blocks they need to perform well later in school (also see Rydell et al., 2010). Moreover, deficits in learning could contribute to recursive social and psychological processes, feeding disidentification or doubts about belonging. How such interactive processes unfold over time, and how they can be interrupted, is an important direction for research. A key implication of these processes is that intervening early in academic environments can forestall these negative cycles and thus generate lasting benefits (Cohen et al., 2006, 2009; Walton & Cohen, 2007, 2011).

An important aspect of the present studies is that they held constant the opportunity to learn. All students studied the presented academic material for the same amount of time; effects emerged because students who experienced stereotype threat were less able to take advantage of this time than others (also see Grimm et al., 2009; Rydell et al., 2010). But in real-world settings, students experiencing stereotype threat may not pursue the same learning opportunities as others. They may dismiss substantive critical feedback if they suspect it results from bias (Cohen, Steele, & Ross, 1999). They may avoid activities that pose a risk of failure and rejection but that facilitate learning, such as seeking help on challenging academic material or taking difficult but educational classes (Mendoza-Denton et al., 2002; Walton & Cohen, 2007). Reducing threat in academic settings may thus make students both more likely to seek out learning opportunities and better able to take advantage of learning opportunities.

Often persistent group differences in academic achievement are attributed primarily to poverty and structural factors. Research on stereotype threat demonstrates the importance, in addition, of psychological processes. By investigating learning, the present research shows how psychological threat can affect achievement even in latter environments. Correspondingly, this research suggests how interventions to reduce stereotype threat could yield broad, long-lasting benefits and underscores the urgency of creating psychologically safe real-world school environments. More generally, it illustrates how subtle features of social situations and the psychological processes they trigger can shape important aspects of individuals' lives.

Acknowledgments

Portions of this research were presented at the meeting of the Society for Personality and Social Psychology, Las Vegas, Nevada, January 2010. We thank Gordon Bower, Carol Dweck, Christine Logel, Nicole Shelton, Steve Spencer, Claude Steele, Eric Uhlmann, and Princeton's SSP lab for feedback and Kai Larson and Margaret Tankard for assistance.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interests with respect to the authorship and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Portions of this research were supported by a National Science Foundation Minority Postdoctoral Fellowship (SMA-0905695) awarded to Valerie Jones Taylor.

Notes

1. Although participants could have studied word definitions between sessions, which could pose a threat to internal validity, it appears that they did not. The words chosen were obscure, and participants were not provided a list of words or word definitions after Session 1, making it difficult to study the words even if participants were motivated to do so. Furthermore, systematic funnel-debriefing procedures after Session 2 suggested that no participant reviewed words between sessions.
2. Complete instructions and materials are available from the first author on request.
3. There was a marginal interaction between word list and learning-threat condition on recall "test" performance, $F(1, 69) = 3.65$, $p = .060$. No simple effect approached significance.
4. Standard deviations and effect size estimates were calculated using the mean-square error term from the ANCOVA, which represents the residual within-cell variability.
5. The complete results of this analysis were as follows. The three-way interaction was predicted to be nonsignificant, as similar patterns were predicted on the "warm-up" and on the "test." This was the case, $F < 1$. The only significant two-way interaction was the Race × Learning-Threat Condition interaction, $F(1, 71) = 5.70$, $p = .020$, $\eta_p^2 = .07$. Collapsing across the "warm-up" and "test," Black students who studied under nonthreatening conditions tended to perform better than Black students who studied under threatening conditions, $t(71) = 1.54$, $p = .13$. White students' performance did not differ by Learning-Threat Condition, $t < 1$. The only significant main effect was the effect of Performance-Threat. Students performed better on the "warm-up" than on the "test," $t(71) = 3.91$, $p < .001$, $d = 0.64$. This effect was significant for Black students, $t(71) = 3.12$, $p = .003$, $d = 0.79$, replicating the standard stereotype threat effect. Unexpectedly, it was also significant for White students, $t(71) = 2.27$, $p = .026$, $d = 0.48$.
6. Including all three tested covariates in analyses in both studies yields results similar to those reported here. However, following Darlington (1996) and past practice (e.g., Walton & Cohen, 2007), we retained only those covariates that were significant in each experiment.
7. The complete results of this analysis were as follows. The main effect of Performance-Threat was significant. Black students performed better on the "warm-up" than on the "test," $F(1, 27) = 6.03$, $p = .021$, $\eta_p^2 = .18$. The main effect of Affirmation Condition was a trend, $F(1, 25) = 1.89$, $p = .18$, $\eta_p^2 = .07$. These main effects were qualified by an Affirmation × Performance-Threat interaction, $F(1, 27) = 4.90$, $p = .036$, $\eta_p^2 = .15$.

References

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review, 79*, 97-123.

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.

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

Cohen, G. L., Garcia, J., Purdie-Vaughns, V., Apfel, N., & Brzustoski, P. (2009). Recursive processes in self-affirmation: Intervening to close the minority achievement gap. *Science, 324*, 400-403.

Cohen, G. L., Steele, C. M., & Ross, L. D. (1999). The mentor's dilemma: Providing critical feedback across the racial divide. *Personality and Social Psychology Bulletin, 25*, 1302-1318.

Critcher, C. R., Dunning, D., & Armor, D. A. (2010). When self-affirmations reduce defensiveness: Timing is key. *Personality and Social Psychology Bulletin, 36*, 947-959.

Darlington, R. (1996). *How many covariates to use in randomized experiments?* Retrieved from http://www.psych.cornell.edu/darlington/covarnum.htm

Grimm, L. R., Markman, A. B., Maddox, W. T., & Baldwin, G. C. (2009). Stereotype threat reinterpreted as a regulatory mismatch. *Journal of Personality and Social Psychology, 96*, 288-304.

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General, 108*, 356-388.

Higgins, E. T. (1998). Promotion and prevention: Regulatory focus as a motivational principle. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 30, pp. 1-46). San Diego, CA: Academic Press.

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.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present day American English*. Providence, RI: Brown University Press.

Lockwood, P., Jordan, C. H., & Kunda, K. (2002). Motivation by positive or negative role models: Regulatory focus determines who will best inspire us. *Journal of Personality and Social Psychology, 83*, 854-864.

Logel, C., Iserman, E. C., Davies, P. G., Quinn, D. M., & Spencer, S. J. (2009). The perils of double consciousness: The role of thought suppression in stereotype threat. *Journal of Experimental Social Psychology, 45*, 299-312.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7*, 83-104.

Mangels, J. A., Good, C., Whiteman, R. C., Maniscalco, B., & Dweck, C. S. (2011). Emotion blocks the path to learning under stereotype threat. *Social Cognitive and Affective Neuroscience*. doi:10.1093/scan/nsq100

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.

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

National Center for Education Statistics. (1999). *National assessment of educational progress 1998 reading report card for the nation and the states*. Washington, DC: Author.

Nguyen, H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology, 93*, 1314-1334.

Osborne, J. W. (1997). Race and academic disidentification. *Journal of Educational Psychology, 84*, 728-735.

Rosen, V. M., & Engle, R. W. (1997). The role of working memory capacity in retrieval. *Journal of Experimental Psychology: General, 126*, 211-227.

Rydell, R. J., Shiffrin, R. M., Boucher, K. L., Van Loo, K., & Rydell, M. T. (2010). Stereotype threat prevents perceptual learning. *Proceedings of the National Academy of Sciences, 107*, 14042-14047.

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.

Sekaquaptewa, D., & Thompson, M. (2002). The differential effects of solo status on members of high- and low-status groups. *Personality and Social Psychology Bulletin, 28*, 694-707.

Sherman, D. K., & Cohen, G. L. (2006). The psychology of self-defense: Self-affirmation theory. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 38, pp. 183-242). San Diego, CA: Academic Press.

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

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. 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. 34, pp. 379-440). San Diego, CA: Academic Press.

Van Selst, M., & Jolicoeur, P. (1994). A solution to the effect of sample size on outlier elimination. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 47A*, 631-650.

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.

Walton, G. M., & Cohen, G. L. (2011). A brief social-belonging intervention improves academic and health outcomes of minority students. *Science, 331*, 1447-1451.

Walton, G. M., & Spencer, S. J. (2009). Latent ability: Grades and test scores systematically underestimate the intellectual ability of women and ethnic minority students. *Psychological Science, 20*, 1132-1139.