

Understanding the Stereotype Threat Effect With “Culture-Free” Tests: An Examination of its Mediators and Measurement

David M. Mayer and Paul J. Hanges
Department of Psychology
University of Maryland

This experiment examined the mediation of the stereotype threat effect and explored the relationship of 2 different types of stereotype threat: *stereotype threat-specific* (threat that results specifically from the testing environment) and *stereotype threat-general* (a global sense of threat that individuals carry across situations) on Raven Advanced Progressive Matrices test performance. As predicted, there was a significant interaction between race (African American or White) and test diagnosticity (intelligence or perceptual ability), such that African Americans experienced more stereotype threat-specific when told the test measured intelligence. Further, both stereotype threat-specific and stereotype threat-general negatively related to test performance. Finally, 3 structural equation models (1 for stereotype threat-specific, 1 for stereotype threat-general, and 1 examining the joint effect of these 2 constructs) simultaneously integrating the proposed mediators failed to find evidence of mediation.

In their 1984 meta-analysis, Hunter and Hunter documented the robustness of the predictive power of cognitive ability tests across a variety of jobs and criteria. Unfortunately, although these tests show substantial criterion-related validity for predicting job performance, they also consistently show substantial mean test score

differences between African Americans and Whites (Hunter & Hunter, 1984). This testing gap increases the likelihood that adverse impact will result when these tests are used in applied settings. Clearly, the use of cognitive ability tests creates dilemmas for organizations needing to assess the cognitive capabilities of their applicants but wanting to maintain or enhance the diversity of their workforce.

Steele and Aronson (1995) recently proposed a self-evaluative explanation for this average test score gap. In particular, these researchers argued that African Americans experience “stereotype threat” when taking cognitive ability tests. *Stereotype threat* is defined as the decrement in test performance that results when members of some group fear that their test performance will confirm a negative stereotype of their group. It is believed that group members will experience stereotype threat if three conditions are met. First, the test has to be relevant to the feared stereotype (e.g., cognitive ability test for African Americans). Second, the construct measured by the test has to be an important aspect of group members’ self-concept. Finally, group members have to identify with their group. If these three conditions are met, then taking a test is believed to activate feelings of stereotype threat.

Steele and Aronson (1995) found that they could induce a test performance gap among African Americans by changing what participants thought the test was assessing. Specifically, half of their participants were told that the test measured mental ability (i.e., commonly referred to as the diagnostic condition) and the other half were told that the test measured some non-stereotype relevant construct (i.e., commonly referred to as the non-diagnostic condition). As predicted, the test score gap between African Americans and Whites appeared in the diagnostic condition but not in the non-diagnostic condition.

Subsequent studies have shown that the stereotype threat effect can be produced with other groups. For example, women show similar performance decrements compared to men when a test is said to measure math ability (Brown & Josephs, 1999; Quinn & Spencer, 2001; Spencer, Steele, & Quinn, 1999). Indeed, the stereotype threat effect has been produced in groups as diverse as Asians, low socioeconomic status individuals, and the elderly which were all negatively influenced by stereotype threat (Croizet & Claire, 1998; Lee & Ottati, 1995; Levy, 1996), and even White males (Aronson et al., 1999; Stone, Lynch, Sjomeling, & Darley, 1999)—a group not typically thought to be stigmatized by stereotypes.

The most common interpretation of these results, at least with regard to the persistent average test score gap for African Americans and Whites, is that cognitive ability tests may be activating the stereotype threat effect in African Americans and thus reducing their average test score (Steele & Aronson, 1995). However, it is not clear whether this is the most appropriate interpretation. Indeed, Sackett, Schmitt, Ellingson, and Kabin (2001) pointed out that the stereotype threat effect is found only after participants are statistically equalized by covarying prior cog-



tive ability test scores. Thus, Sackett et al. argued that the correct interpretation of the stereotype threat studies is that these studies demonstrate that the typical 1 *SD* test score differences can be increased. Regardless of how the stereotype threat results are interpreted, the effect appears to be quite robust in laboratory settings, as evidenced by the findings of a recent meta-analysis (e.g., Jones & Stangor, 2002). Thus, it appears reasonable to explore the mechanism through which the stereotype threat effect operates as well as to assess the boundary conditions of this effect. This study was designed to address these issues.

Specifically, we attempted to accomplish several goals in this study. In particular, we assessed four potential mediators of the stereotype threat effect. These variables have been identified as potential mediators in the stereotype threat literature, but prior empirical studies have failed to find support for their hypothesized role as mediators. We hypothesize that this lack of support may be due to studying these mediators in isolation and thus, we test a path model that incorporates these four mediators into one model to assess their joint role in mediating the stereotype threat effect.

In addition to simultaneously testing several mediators, we also attempted to directly assess the extent to which participants experienced stereotype threat, to determine the extent to which a culture-free measure of cognitive ability can activate these feelings, and to provide a “realistic” context by using a subtle manipulation of stereotype threat and by presenting the test as one commonly used by organizations for personnel selection. We begin our discussion with an exploration of the potential mediators of the stereotype threat effect.

PROPOSED MEDIATORS OF THE STEREOTYPE THREAT EFFECT

As indicated previously, the Jones and Stangor (2002) meta-analysis has demonstrated the robustness of the stereotype threat effect in laboratory settings. Unfortunately, although this meta-analysis has shown that the stereotype threat effect can be generated in laboratory settings, it is unclear why or how the effect occurs. A number of possible mechanisms have been suggested. For example, it is possible that activating the fear of confirming a stereotype diverts attention to task-irrelevant worries (i.e., cognitive interference) and increases anxiety (Steele & Aronson, 1995). A few studies have attempted to verify these mediating factors. Consistent with the anxiety explanation, Blascovich, Spencer, Quinn, and Steele (2001) demonstrated that African Americans have higher blood pressure when under stereotype threat manipulation. Although suggestive, it should be noted that Blascovich et al. (2001) did not perform a formal test of mediation. Other researchers have more directly tested the hypothesized mediational role of task-irrelevant thoughts and anxiety (Oswald & Harvey, 2000; Schmader, 2002; Spencer, Steele, & Quinn,

1999; Stone et al., 1999). Unfortunately, very limited support has been found for these two potential mediators.

Another possible mediator, originally proposed by Steele and Aronson (1995), is that the stereotype threat manipulation affects participants' self-efficacy. As suggested by the self-efficacy literature (Bandura, 1986), it is possible that factors in the testing environment (e.g., construct believed to be measured by the test) can decrease performance expectations and subsequently hurt actual test performance. Stangor, Carr, and Kiang (1998) found that stereotype threat had a negative effect on the test expectations of targeted group members. Unfortunately, studies that have tested the mediating role of expectancies or self-efficacy (Oswald & Harvey, 2000; Spencer, Steele, & Quinn, 1999; Stone et al., 1999) for the stereotype threat effect have not been supportive.

Finally, Steele and Aronson (1995) suggested that the stereotype threat manipulation might increase the evaluation apprehension of targeted group members. When an individual is put in a position to potentially confirm a negative stereotype, she or he will likely be concerned with how his or her performance will be perceived by others. Thus, the fear of being evaluated negatively could lead to performance decrements of individuals under stereotype threat.

Although some of these variables have been tested in other studies, they have been examined, for the most part, separately. It is possible that stereotype threat influences test scores through several mediators with only a slight affect on any one mediator. It is also possible that these variables influence one another making simple mediation models inappropriate. Thus, the prior support for the mediators may be weak because the mediators were tested separately. One of the major purposes of this study was to simultaneously test these multiple mediators in one study. In this way, this study attempts to provide a comprehensive assessment of the proposed mediators of the stereotype threat effect by testing theoretical models integrating these variables.

In summary, we explored four potential mediators of the stereotype threat effect: increased anxiety, increased cognitive interference, reductions in self-efficacy, and increased evaluation apprehension. It was hypothesized that prior studies failed to find support for these mediators because they studied the mediators separately. It is possible that a more complex mediator model is needed to describe how stereotype threat works.

Stereotype Threat and Culture-Free Cognitive Ability Tests

In a recent study, McKay, Doverspike, Brown-Hilton, and Martin (2002) sought to understand how the characteristics of the test itself affected the robustness of the stereotype threat manipulation. Whereas researchers have previously used verbal questions that are typical of standardized exams (e.g., Steele & Aronson, 1995), McKay et al. (2002) used the Raven Advanced Progressive Matrices (APM) test.

This test was developed in the late 1930s as an attempt to develop a “culture-free” or more accurately a “culture-reduced” measure of intelligence (Jensen, 1980). Culture-reduced tests are carefully constructed to minimize the influence of irrelevant factors that create between-group (i.e., cultural) differences on test scores. Researchers primarily have eliminated the verbal content of their tests to reduce the cultural-loading of their tests (Jensen, 1980). The Raven APM test is a nonverbal test of cognitive ability that requires test-takers to accurately choose which one of eight alternative matrix pieces accurately completes an incomplete matrix diagram. There are a total of 36 problems that are moderately challenging and steadily increase in difficulty. Research has shown that the Raven APM test is a construct valid measure of general cognitive ability (Raven, Court, & Raven, 1985).

In addition to using the Raven APM test, McKay et al. (2002) also directly measured whether participants were experiencing stereotype threat feelings during the experiment. Indeed, measurement of stereotype threat feelings is important since without such a measure, it is impossible to know whether the experimental manipulation was construct valid. That is, the experimental manipulation could be causing test score decrements because it was activating feelings of stereotype threat in certain participants (as believed) or because it was activating some other, unexpected, process. Thus, direct measurement of stereotype threat feelings allows construct validation of the experimental manipulation. McKay et al. (2002) developed a 3-item scale to directly measure stereotype threat.

McKay et al. (2002) found that, when the Raven APM test was described as a cognitive ability measure, African Americans had stronger feelings of stereotype threat than did White participants. Despite the fact that McKay et al. (2002) only found a marginally significant decrement in test performance, their results tentatively suggest that even when “culture-free” tests are used such as the Raven APM test, it is possible to witness a decrease in performance due to stereotype threat. Thus, consistent with hypotheses, it is possible that the nature of the test can diminish or enhance the extent to which the test activates stereotype threat fears.

This study expands on the work of McKay et al. (2002). In particular, although the McKay et al. study provides some evidence that the stereotype threat manipulation does affect stereotype threat perceptions, it did not assess any of the proposed mediator variables. This study expands on this work by assessing the role of multiple mediators when the Raven APM test is used.

Second, this study uses a comprehensive perception measure of stereotype threat. Recent research by Ployhart, Ziegert, and McFarland (2003) indicates that there are two types of stereotype threat. The first type is situation-specific and is activated by characteristics of the test or the testing situation. It is this situation-specific version of the threat that the experimental manipulations discussed previously should affect. We will refer to this form as *stereotype threat-specific*. The second form is a more general perception (*stereotype threat-general*) that generalizes across situations (Ployhart et al., 2003). Stereotype threat-general re-

fers to a constant vigilance against confirming negative stereotypes about one's group across multiple contexts.¹

Although the McKay et al. (2002) study showed that participants had stereotype threat perceptions, it is important to note that their three-item measure did not distinguish between the situation-specific and general variety. It is very possible that some of the participants in the McKay et al. study exhibited stereotype threat-general feelings. Indeed, McKay et al. deliberately sampled their participants from two different academic institutions. One of the universities was a historically African American college and nearly two thirds of the African American participants came from this school. It is possible that participants sampled from this university had stronger ethnic identity feelings and were more intent on not confirming negative stereotypes about their group. Thus, it is unclear whether the stereotype threat measure used in the McKay et al. study reflected situation-specific or more general perceptions. It is also unclear whether only one or both of these stereotype threat forms affect test scores. In this study, we used the multi-dimensional measure of stereotype threat developed by Ployhart et al. (2003).

Third, we attempt to provide a realistic context to the experimental design. Participants were told that they would be taking a test commonly used for personnel selection in fields that would appeal to many college students (e.g., architecture, graphic design, computer jobs) while maintaining face validity given the format of the test, and that experimenters were interested in what they thought and how they felt about the test. Further, unlike some of the previous stereotype threat studies, we did not emphasize participants' race as part of our manipulation nor did we remind participants about a negative stereotype regarding their racial group's test performance because these aspects of the stereotype threat manipulation probably would not occur in real-world contexts. Although race may be salient to actual job applicants, the manipulation of race salience is probably much more subtle in the real world (e.g., single minority applicant taking test with only majority group applicants). It is extremely unlikely that an organizational representative would

¹We attempted to confirm the factor structure of the Ployhart, Ziegert, and McFarland (2003) stereotype threat measure in this study. As indicated earlier, these authors found that a two factor model (i.e., stereotype threat-general and stereotype threat-specific) was appropriate for Black participants, but that a single factor solution was needed for White participants. We found comparable results. For our Black participants, the two factor model fit well, $\chi^2(19) = 13.86, p > .10$, (CFI = 1.00; RMSEA = 0.00) with all the estimated parameters being within acceptable levels. However, similar to Ployhart et al., we found that this two factor model created problems when it was imposed on the White participant data, $\chi^2(19) = 78.46$ (CFI = .97; RMSEA = .10). In particular, we noted that the correlation between the two latent factors exceeded 1.0. Thus, we re-estimated the model for this subgroup by imposing a single factor solution. Although the fit indexes for this model did not substantially change, $\chi^2(20) = 88.90, p < .05$ (CFI = 0.97; RMSEA = 0.10), all the estimated parameters of this model were within acceptable levels. These results are suggestive but not definitive because the number of observations within our two racial subgroups was smaller than desired for a confirmatory factor analysis.

mention race or remind applicants of a racial stereotype right before handing out an exam. In fact, the bluntness of these manipulations probably activates other constructs (e.g., reactance) that can affect participants' test performance.

In summary, this study explored the influence of multiple mediators of the stereotype threat effect. Further, we distinguish between the situation-specific and more general forms of this perception to determine the separate contributions of these variables on test performance. We also used the culture-free Raven APM test as our cognitive ability measure. Finally, we attempted to create a testing context that was as close to the testing environment that would be encountered in an actual organization.

Research Hypotheses

Based on this review of the literature, the following hypotheses were generated.

- H1. African American participants will report significantly greater levels of stereotype threat—general than will White participants.
- H2. African Americans who identify with the stigmatized group and who do not discount the test will experience more stereotype threat—specific than White participants when they are told that the test measures intelligence.
- H3. African Americans who identify with the stigmatized group and who do not discount the test will show greater decrements in test performance than will White participants when they are told that the test measures intelligence.
- H4. There will be a negative relationship between stereotype threat—specific, stereotype threat—general and test performance.
- H5. The relationship between stereotype threat and test performance will be mediated by anxiety, cognitive interference, self-efficacy, and evaluation apprehension.

METHOD

Participants and Design

A total of 62 African American and 90 White undergraduates from a large University in the Mid-Atlantic participated in the experiment. The sample was 64% female and the average age of participants was 19-years-old.

The experimental design was a 2 (race: African American or White) \times 2 (test diagnosticity: intelligence or perceptual ability) between-subjects design. The primary manipulation in the study involved telling participants that the cognitive ability test was either a measure of intelligence (stereotype threat condition) or a mea-

sure of perceptual ability (non-stereotype threat condition). As discussed earlier, we believed this to be a subtle manipulation and could potentially be used in a real world job selection process. In addition, because our hypotheses were directional (e.g., stereotype-threat specific will have a direct negative relationship with test performance), we used one-tailed tests to assess support for our hypotheses.

Measures

Cognitive ability. The Raven APM test was designed to assess general intelligence (Raven, Court, & Raven, 1985). The items consist of a set of matrices, or arrangements of design elements into rows and columns, from each of which a part has been removed. The task is to choose the missing insert from the given alternatives. Language or item content does not culturally bias the test. Furthermore, its ambiguous form can easily be explained to participants as either an intelligence test or a test of some other construct (e.g., perceptual ability). There are a total of 36 items and participants were given 20 min to work on the set.

Stereotype threat-general. Stereotype threat-general was measured with a 3-item scale that assessed the degree to which individuals perceive society views their group in terms of a negative stereotype (Ployhart et al., 2003). The measure is on a 5-point Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). A sample item is, "Some people believe I have less intelligence because of my race." The alpha for the scale was .63.

Stereotype threat-specific. Stereotype threat-specific was measured with a 5-item scale that assessed individuals' perceptions about the test they took (Ployhart et al., 2003). The measure is on a 5-point Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). A sample item is, "During the test, I wanted to show that people of my race could perform well on it." The alpha for the scale was .74.

Anxiety. The measure used to assess anxiety was the Spielberger State Anxiety Scale (Spielberger, Gorsuch, & Lushene, 1970). It is a 19-item measure where respondents are asked to indicate how they felt immediately prior to taking the test. They responded to a number of statements such as, "I feel worried," and "I feel anxious," along a 5-point scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). The alpha for this scale was .92.

Evaluation apprehension. Evaluation apprehension refers to the extent to which one fears how others will view him or her based on test performance. The 4-item measure is on a 5-point Likert scale ranging from 1 (*strongly disagree*) to 5

(*strongly agree*). A sample item is, "People will look down on me if I do not perform well on this test." The alpha for the scale was .86.

Self-efficacy. Self-efficacy was measured with a 6-item scale that was loosely based on Brutus and Ryan (1998). The measure is on a 5-point Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). A sample item is, "I am certain I have the skills to do well on this test." The alpha for the scale was .69.

Cognitive interference. Sarason's (1972) measure of cognitive interference was used to assess task-irrelevant thoughts. The 21 items range on a scale from 1 (*strongly disagree*) to 5 (*strongly agree*). A sample item is, "I thought about the difficulty of the problems." The alpha for the scale was .90.

Belief in tests. Belief in tests measures how much participants value the use of tests in general. It was assessed using 4 items from the Test Attitude Survey (TAS; Arvey, Strickland, Drauden, & Martin, 1990). The measure is on a 5-point Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). A sample item is, "These kinds of tests should be eliminated." The alpha for the scale was .76.

Predictive validity. Predictive validity is the extent to which people believe the test actually predicts what it is supposed to (i.e., intelligence). It was assessed using 5 items from the TAS (Arvey et al., 1990). The measure is on a 5-point Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). A sample item is, "I am confident that this examination can predict how intelligent a person is." The alpha for the scale was .87.

Ethnic identity. Ethnic identity measured how much an individual's self-definition relates to their ethnic group membership. The measure is 8 items and is measured on a 5-point Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*; Helms, 1990). A sample item is, "My race is an important reflection of who I am." The alpha for the scale was .77.

Procedure

One month prior to participating in the study, participants completed the ethnic identity scale as part of a mass testing session in which measures from multiple studies were collected simultaneously. By including the ethnic identity scale in the mass testing session, we were able to gather this information without priming race during the experiment.

On entering the laboratory, participants were told that they would play the role of a job applicant and that the test they would take is commonly used by organizations to hire individuals in architecture, graphic design, and many computer related

jobs. They were also told that the experimenters were interested in job applicants' attitudes and affect about the test. Participants were shown a few sample questions from the Raven APM test and given the test diagnosticity manipulation. Half of the participants were told this test was a measure of intelligence (stereotype-threat condition) and the other half was told that the test was a measure of perceptual ability (non-stereotype threat condition). They were then told that the experimenters were interested in job applicants' reactions to this test and, after reviewing these questions, participants completed the anxiety, cognitive interference, self-efficacy, and evaluation apprehension measures as well as the measures assessing test-taking attitudes (e.g., predictive validity, belief in tests). They then took the Raven APM test for 20 min.

After the time allotted for the test expired, a measure of cognitive interference was given as was the stereotype threat-general and stereotype threat-specific measures. Finally, four items were given to determine whether participants thought they were told the test was a measure of either intelligence or perceptual ability and also whether they believed it or not. After the completion of these final questions and some demographic information participants were fully debriefed.

RESULTS

The means, standard deviations, and intercorrelations including scale reliabilities of the key variables in the study are located in Table 1.

Manipulation Check

Two items were used to assess whether the stereotype threat manipulation was salient to participants. Participants responded to the following two questions: "I was told that this test measures intelligence" and "I was told that this test measures perceptual ability." An independent sample's *t* test was conducted to ensure that participants in each condition were aware of how the experimenter labeled the diagnosticity of the test. Participants in the intelligence condition scored significantly higher on the item asking them if they were told the test measured intelligence than participants in the perceptual ability group, $t(139) = 11.38, p < .05$. Similarly, participants in the perceptual ability condition had a significantly higher score on the question that asked them whether they were told the test measures perceptual ability than those in the intelligence condition, $t(134) = -9.45, p < .05$. Thus, it appears that participants picked up on how the experimenter labeled the diagnosticity of the test.

Test of Hypotheses

Hypothesis 1 posited that African Americans would experience more stereotype threat-general than White participants. Consistent with this hypothesis, African

TABLE 1
Means, Standard Deviations, and Intercorrelations for All Measures

<i>Measures</i>	<i>M</i>	<i>SD</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>	<i>8</i>	<i>9</i>	<i>10</i>
Predictive validity	2.46	.84	(.87)									
Belief in tests	3.08	.78	.67	(.76)								
Ethnic identity	3.78	.73	-.02	-.09	(.77)							
Anxiety	3.90	.61	.08	-.07	-.08	(.92)						
Cognitive interference	2.51	.71	-.01	.03	.15	.47	(.90)					
Self-efficacy	4.04	.62	-.08	.10	-.05	-.35	-.25	(.69)				
Evaluation apprehension	1.71	.79	.27	.12	-.14	.38	-.27	-.40	(.86)			
Stereotype threat-specific	1.93	.69	-.01	-.14	.27	.08	.08	-.09	.23	(.74)		
Stereotype threat-general	2.68	1.09	-.04	-.05	.40	-.14	-.03	.02	-.07	.60	(.63)	
Raven test score	19.63	4.83	.16	.21	-.21	.02	-.09	.15	-.01	-.15	-.30	

Note. All correlations greater than .16 are significant, $p < .05$. Scale reliabilities are shown along the diagonal.

Americans ($M = 3.57$, $SD = .90$) reported significantly more stereotype threat—general than did Whites ($M = 2.09$, $SD = .75$), $t(150) = -10.98$, $p < .001$. Thus, Hypothesis 1 was supported.

Hypothesis 2 predicted that once the participants were equalized in the perceived predictive validity of the test, belief in tests, and in ethnic identity, the differences between the two racial groups on stereotype threat—specific would be larger in the intelligence condition than in the perceptual ability condition. We control for these variables to be consistent with Steele and Aronson (1995) boundary conditions for the stereotype threat effect. As indicated earlier, stereotype threat is expected to occur only when individuals identify with the targeted group and believe that the test measures something relevant to the feared stereotype.²

We tested this hypothesis by conducting a hierarchical regression in which perceived predictive validity, belief in tests, and ethnic identity were entered first, then dummy coded variables representing the two main effects (i.e., race and test diagnosticity condition) were entered, and finally, a single variable representing the interaction of the two main effects was entered into the regression equation. Consistent with Steele and Aronson's (1995) discussion, the three covariates contributed significantly as a block, $F(3, 115) = 3.97$, $p \leq .01$. More specifically, ethnic identity was the only significant predictor ($\beta = 0.25$), $t(115) = 2.78$, $p \leq .01$, of stereotype threat—specific among these block of variables.

We found a significant Race \times Test Diagnosticity interaction, $F(1, 112) = 5.02$, $p \leq .05$, $\Delta R^2 = 0.03$, and the plot of this interaction using means corrected for the three covariates is shown in Figure 1. Consistent with our hypothesis, the largest White/African American difference on stereotype threat experienced due to the test was found in the intelligence condition. The discrepancy between African Americans ($M = 2.52$) and Whites ($M = 1.56$) on stereotype threat—specific in the intelligence condition was larger than the difference between African Americans ($M = 2.23$) and Whites ($M = 1.76$) in the perceptual ability condition.

Hypothesis 3 posited that African Americans who identify with the stigmatized group and who do not discount the test will show greater decrements in test performance than will White participants when they are told that the test measures intelligence. To test this hypothesis, we ran a hierarchical regression similar to the one conducted for Hypothesis 2 except that the Raven APM test score was used as the dependent variable in this analysis. Unfortunately, in contrast to the findings of

² We statistically removed these test-taking attitudes (e.g., perceived predictive validity, belief in tests) and ethnic identity out of our data because Steele and Aronson (1995) indicated that variation on these constructs create noise in the stereotype threat effect. Although we could have assessed the moderating effect of these variables on the stereotype threat effect, this would have required assessing a five-way interaction after controlling for all lower level interactions. Given that the self-report measures were correlated with each other and the likelihood of interpreting a five-way interaction, we decided not to pursue this approach.

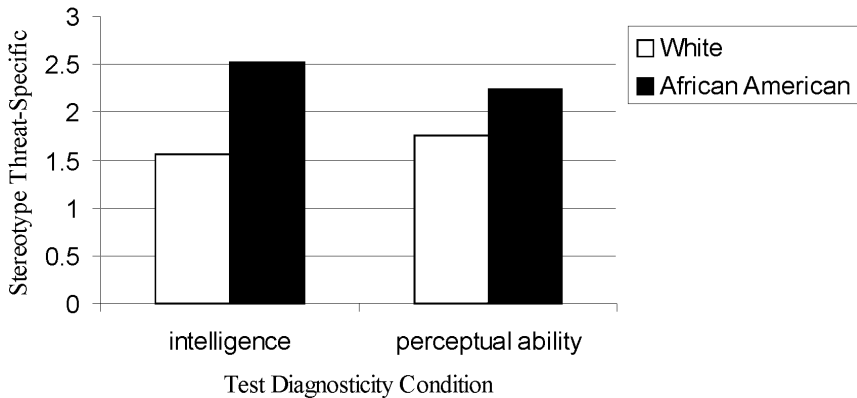


FIGURE 1 Race \times Test Diagnosticity Condition interaction on stereotype threat-specific.

McKay et al. (2002), we did not find a significant interaction, $F(1, 112) = .004$, $p > .05$; $\Delta R^2 = 0.00$. The adjusted-mean difference between Whites ($M = 21.83$) and African Americans ($M = 19.11$) in the intelligence condition did not significantly differ from the difference between Whites ($M = 18.96$) and African Americans ($M = 16.14$) in the perceptual ability condition.

The first three hypotheses attempted to replicate and extend the findings of McKay et al. (2002). The next two hypotheses attempt to further extend their findings by examining whether self-report measures of stereotype threat predict test score and to test for mediators of the effect. Hypothesis 4 predicted a negative relationship between stereotype threat-specific, stereotype threat-general, and test performance. We tested this hypothesis by examining the correlations among the constructs using one-tailed tests of significance.

The first analysis was to determine whether stereotype threat-specific significantly predicted test score. There was a significant negative relationship found between stereotype threat-specific and test score, $r(151) = -.15$, $p < .05$. Thus, there was support for the hypothesis that stereotype threat resulting from the specific testing situation has a negative effect on test scores. The same analysis was performed with stereotype-threat general as a predictor of Raven's Test performance. There was a significant negative relationship between stereotype threat-general and test score, $r(151) = -.30$, $p < .001$. The results of Hypothesis 4 suggest that stereotype threat that comes from the test has an effect on test performance but a more general feeling of stereotype threat (i.e., how you feel society perceives your group in terms of a particular stereotype) appears to have an even greater impact.

Hypothesis 5 explored the role of anxiety, self-efficacy, cognitive interference, and evaluation apprehension as mediator variables for the effect of stereotype threat on test scores. To test this hypothesis, we used MPlus (Muthen & Muthen, 1998) to perform two structural equation models (SEM). In the first model, we

tested whether the four identified variables mediated the relationship between stereotype threat–specific and test scores. Figure 2 shows the standardized path coefficients for the SEM testing the stereotype threat–specific variable. All the paths not shown in this figure were forced to zero. The fit of this model was very good, $\chi^2(8) = 7.72, p > .05$; CFI = 1.00; RMSEA = 0.00; SRMR = .051.

As can be seen in Figure 2, stereotype threat–specific only significantly affected evaluation apprehension. Consistent with our hypothesis, as stereotype threat–specific increased, so did evaluation apprehension. However, contrary to expectations, stereotype threat–specific did not affect any of the other mediators. Further, contrary to the mediation hypothesis, evaluation apprehension was not significantly related to test scores and there was still a significant direct relationship between stereotype threat–specific and test score. The only hypothesized mediator that was related to the test score was efficacy, but stereotype threat–specific was not related to this variable. Overall, these results failed to support this hypothesis for stereotype threat–specific.

The second SEM model that we tested explored whether the four aforementioned variables mediated the relationship between stereotype threat–general and test scores. Figure 3 shows the standardized path coefficients for the stereotype threat–general SEM. All the paths not shown in this figure were forced to zero. Once again, the fit of the model was very good, $\chi^2(9) = 8.01, p > .05$; CFI = 1.00; RMSEA = 0.00; SRMR = .050.

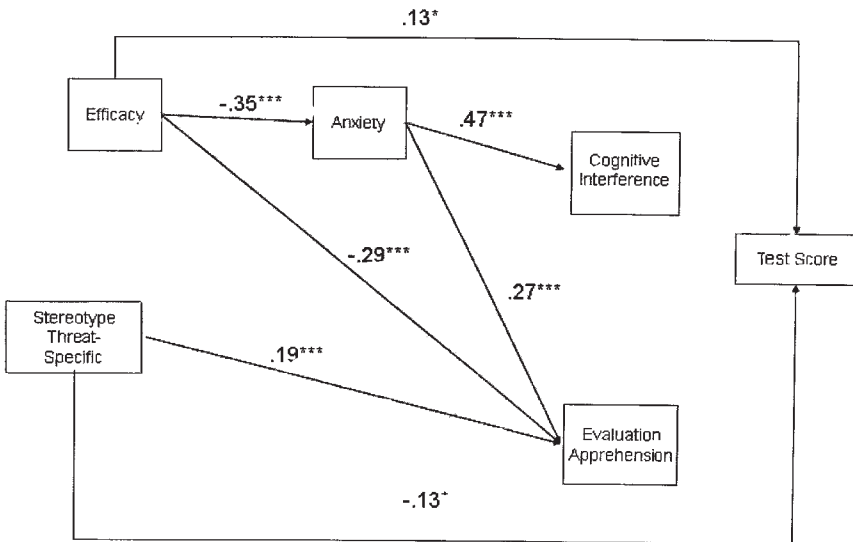


FIGURE 2 Mediation path model for stereotype threat–specific.

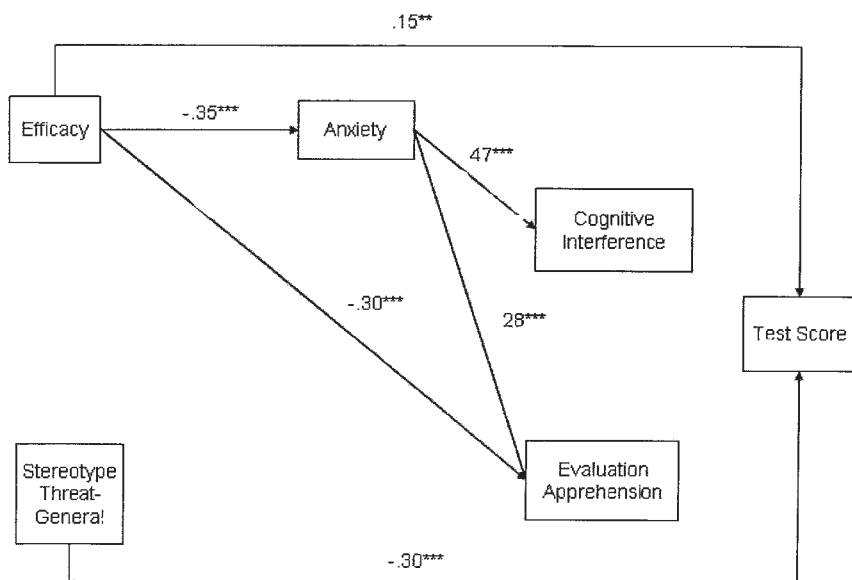


FIGURE 3 Mediation path model for stereotype threat-general.

As can be seen in this figure, contrary to our expectations, stereotype threat-general was not significantly related to any of the hypothesized mediators.³ Also, there was a significant direct relationship between stereotype threat-general and test scores. Once again, efficacy significantly affected test scores, but stereotype threat-general was not significantly related to efficacy.

Finally, since both stereotype threat-specific and stereotype threat-general were believed to have unique effects, we ran a third SEM which combined the models shown in Figures 1 and 2. The only change to these two models is that we predicted that participants' level of stereotype threat-specific would be influenced by their predisposition towards stereotype threat (i.e., stereotype threat-general).

Figure 4 shows the standardized path coefficients for this combined SEM. All the paths not shown in this figure were forced to zero. Although the fit of this combined model was not as good as the prior models, it fit extremely well, $\chi^2(12) = 20.84, p > .05$; CFI = .96; RMSEA = 0.07; SRMR = .062. As can be seen in Figure 4, the general level of stereotype threat experienced by participants (i.e., stereotype

³It should be noted that the path from stereotype threat-general to anxiety was technically significant but in the wrong direction (i.e., more stereotype threat-general tended to decrease anxiety). Given our use of a one-tailed test and the unusual direction of this relationship, we concluded that this relationship was not theoretically meaningful. Indeed, when we forced the stereotype threat-general to anxiety relationship to be zero, there was no loss in model fit.

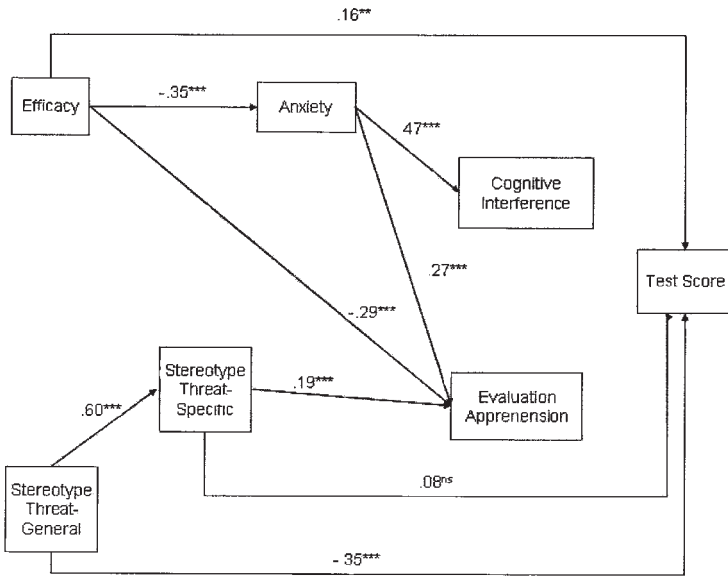


FIGURE 4 Mediation path model for both stereotype threat-specific and stereotype threat-general.

threat-general) substantially influenced the extent to which they experienced threat in our laboratory study (i.e., stereotype threat-specific). This figure also shows that one previously significant path in a prior SEM is no longer significant. In particular, the significant path between stereotype threat-specific and Raven’s APM score (see Figure 2) was no longer significant when stereotype threat-general is also in the model. Thus, it appears that only stereotype threat-general and efficacy directly affect the Raven’s APM score in our study. In summary, all three SEM models failed to support Hypothesis 5.

Additional Analyses

After failing to find any support for the hypothesized mediators, we wondered whether participants believed the information provided in the diagnostic manipulation. Did participants who were told that the Raven’s APM test measured intelligence actually believe that the test measured intelligence? Also, did participants who were told that the Raven’s APM test measured perceptual ability actually believe that the test measured perceptual ability? Our manipulation checks verified that participants correctly indicated what they were told about the APM. However, it is possible that they were more willing to believe one of these explanations over the other. Indeed, prior research has shown that people commonly believe that the

Raven's APM test measures perceptual ability (Jensen, 1980). However, empirical work has repeatedly shown it to be one of the purest measures of general intelligence that exists (Jensen, 1980, 1998). To conduct these exploratory analyses, we included two 5-point questions in our analyses. The first question ("perception belief") asked participants to rate the extent to which they believed that the test measured perceptual ability 1 (*strongly disagree*) to 5 (*strongly agree*). The second question ("intelligence belief") asked participants to rate the extent to which they believed that the Raven's APM test measured intelligence 1 (*strongly disagree*) to 5 (*strongly agree*).

We correlated these single items with the dummy coded diagnostic manipulation variable. There was a significant correlation between the perception belief question and the diagnostic manipulation variable, $r(137) = -.21, p < .05$, but not between the intelligence belief question and the diagnostic manipulation variable, $r(141) = .03, p > .05$. Consistent with expectations, participants in the non-diagnostic condition (i.e., experimenter indicated that the test measured perceptual ability) were significantly more likely to believe that the test measured perceptual ability than were participants in the diagnostic condition (i.e., experimenter indicated that the test measured cognitive ability). However, contrary to expectations, participants were no more likely to believe that the Raven's APM test measured intelligence if they were in the diagnostic as opposed to the non-diagnostic condition. These results imply that participants believed the test measured perceptual ability as opposed to intelligence even when explicitly told that the Raven's APM test measures intelligence.

Even more interestingly, we found a significant Race \times Test Diagnosticity interaction with regard to the belief that the test measured perceptual ability, $t(97) = -2.10, p < .05$, and this interaction is shown in Figure 5. African American participants in the diagnostic condition (i.e., told the test measures intelligence) were more likely to believe the test actually measured perceptual ability than were Afri-

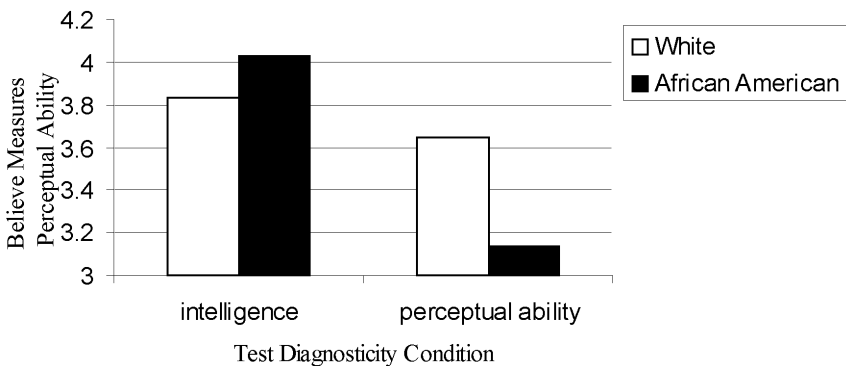


FIGURE 5 Race \times Test Diagnosticity Condition on perception belief.

can Americans in the non-diagnostic condition (i.e., told that the test measures perceptual ability). This interaction suggests that when the African American participants were put into a threatening situation (i.e., told the test measures intelligence), the ambiguous nature of the Raven's APM test allowed them to discount the experimenter and re-label the construct measured by the test as something less threatening (i.e., a perceptual ability measure). These results combined with the empirical evidence that strongly indicates that the Raven's APM test is a factor pure measure of intelligence seems to suggest that the face validity of the test might be an important moderator of the stereotype threat effect.

DISCUSSION

The purpose of this study was to explore the mechanism underlying Steele and Aronson's (1995) stereotype threat effect. In particular, we simultaneously explored the role of four hypothesized mediators (i.e., anxiety, self-efficacy, cognitive interference, evaluation apprehension) for the stereotype threat effect. In addition, we replicated the work of McKay et al. (2002) by exploring the effect of stereotype threat in a more realistic selection setting. We extended the McKay et al. work by using a self-report measure of stereotype threat that differentiated between test-specific and more general (trait-like) feelings of stereotype threat. We explored the relationship between these two stereotype threat measures and the various mediators as well as the Raven APM test performance.

Overall, many of our hypotheses were supported. In particular, we found that it was meaningful to separate stereotype threat into two components. As predicted, we found that the diagnosticity manipulation significantly affected stereotype threat-specific but not stereotype threat-general. In addition, the SEM analyses demonstrated that these two kinds of stereotype threat operated through different mechanisms. In particular, stereotype threat-specific was found to be positively related to evaluation apprehension whereas stereotype threat-general was found to be positively related to anxiety. These findings for the two facets of stereotype threat make conceptual sense and justify their separate measurement.

Although a few studies have directly measured participant's reactions to the diagnosticity manipulations (e.g., Steele & Aronson, 1995; McKay et al., 2002), these studies used more global measures that did not distinguish between the situation-specific and more general forms of stereotype threat. Future research is needed to understand the nomological network underlying the stereotype threat-general construct. For example, the work of Pinel (1999) on stigma consciousness provides a related perspective on the types of lenses people use to perceive the social world. Stigma consciousness refers to the perceived probability of being stereotyped (Pinel, 1999). Although Pinel distinguished between stereotype threat and stigma consciousness, they are certainly in the same family of con-

structs. The key point highlighted in this study is not so much the differentiation between stereotype threat–general and other like constructs such as stigma consciousness. Our intent is to elucidate the notion that the measurement of stereotype threat (which is a very important measurement issue to verify that performance decrements are in fact due to threat) may be more complex than originally conceived of and we may be tapping multiple constructs with the measures that have been used in the literature thus far. Separate measurement of these two forms is important because it provides direct confirmation that the experimental manipulation affects the test score through the hypothesized mechanism.

Further, we predicted that both of these forms of stereotype threat would have negative direct effects on test performance. Consistent with our expectations, both stereotype threat–specific and stereotype threat–general were found to have negative relationships with performance on the Raven’s APM test. However, we found that the stereotype threat–general facet had a stronger effect on the Raven’s APM test performance in this study than either the experimental manipulation or the stereotype threat–specific measure. Perhaps it was the selection context of this study or perhaps it was the Raven’s APM test and participants’ misinterpretation of this test as a measure of perceptual ability that diminished the effect of the situational stereotype threat (i.e., diagnosticity manipulation) on the test scores. Future research is needed to disentangle the effect of these two factors.

Unfortunately, we were not successful in identifying the mediators of the stereotype threat effect. Although both facets of stereotype threat were significantly related to certain mediator variables, these mediators were not subsequently related to test scores. Of the four mediators identified in the literature, only one (i.e., self-efficacy) exhibited a significant relationship to the Raven’s APM test scores. Consistent with the self-efficacy literature, participants with higher levels of self-efficacy performed better on the Raven’s APM test than did participants with lower levels of self-efficacy. Unfortunately, neither stereotype threat facet was significantly related to self-efficacy. Thus, the question remains: What is the mechanism by which the stereotype threat effect affects cognitive ability test scores?

Perhaps our lack of support for mediation was due to us leaving out a critical mediator from our models. Indeed, Steele (1999) recently suggested another mediator through which the stereotype threat effect may operate. Specifically, he stated, “underperformance appears to be rooted less in self-doubt than in social mistrust” (p. 52). In other words, rather than the mediators tested in this study (e.g., anxiety, evaluation apprehension), Steele suggested that a general distrust of others might be causing test performance decrements. Consistent with Steele’s arguments, recent studies have demonstrated that societal mistrust can have a negative impact on cognitive ability test performance of African Americans (Hayes & Ployhart, 2002). Although interesting, these findings suggest that societal mistrust is only a partial explanation for the stereotype threat effect. Specifically, societal mistrust appears to be a stable attitude that people have across a variety of situations. Thus,

it is most likely a useful mediator for the effect of the stereotype threat–general facet as opposed to the stereotype threat–specific facet. Also, this study did find some evidence that anxiety and evaluation apprehension were affected by stereotype threat. Given these caveats, it does appear that societal mistrust might be a useful variable to include in future studies.

Stereotype Threat in a Realistic Environment

In addition to assessing the mediators of the stereotype threat effect, another important goal of this study was to try to create stereotype threat perceptions using more realistic manipulations that resembled the kinds of experiences of job applicants in the real world. Although clearly this study took place in an artificial laboratory context, we did not manipulate stereotype threat by reminding participants about the feared stereotype before they took the exam (as is done in prior laboratory studies). Nor did we attempt to make race salient to participants in our study. In fact, we took steps to minimize the salience of race in our study. We did this to create conditions that would most likely occur in actual employment testing settings. In this study, we only attempted to manipulate stereotype threat by informing participants about the construct being measured by the test.

Despite the subtle manipulation, we still obtained a significant race by diagnostic manipulation on the stereotype threat–specific measure. Thus, it appears that we were successful in creating conditions of stereotype threat even with a more realistic but subtle manipulation. However, the most critical consequence of the stereotype threat effect (its negative influence on test scores) was minimal in this study. Specifically, we failed to find a significant relationship between the diagnosticity manipulation and the Raven's APM test score. Overall, our results suggest that, in more realistic settings, feelings of stereotype threat may occur but their actual influence on test scores may be minimal.

Alternative Forms of Testing

Finally, it is possible that the stereotype threat effect might have been minimized in this study by our use of a non-verbal measure of intelligence. We discovered that, despite being told that the Raven's Test measured intelligence, participants tended to believe that it was a measure of perceptual ability. Even after accounting for participants' belief in tests and their perceived predictive validity of the Raven's APM test, participants still discounted the possibility that the test measured cognitive ability. This result suggests that racial differences could be minimized in employment settings if job applicants believe that a test is measuring some relevant but non-threatening construct. Perhaps increasing the physical fidelity of an exam, without sacrificing its psychological fidelity, will provide applicants with a non-threatening context that will minimize perceptions of stereotype threat.

Practical Implications

There are a number of practical implications that can be drawn from this study. One key for organizations is to try not to prime a stereotype or ethnic identity prior to administering a personnel selection exam. Words such as “intelligence” or “cognitive ability” should be avoided when providing instructions in an attempt to not prime negative stereotypes. Similarly, in an attempt to not make race salient, it is important to pay attention when many applicants are taking a test at the same time. Research has demonstrated that group composition can be sufficient to elicit stereotype threat, such that being in the minority can hurt performance (Inzlicht & Ben-Zeev, 2000).

In addition to these issues related to priming race and stereotypes, the results of this study suggest that using alternate forms of testing may be helpful in reducing stereotype threat. Although African Americans did experience more stereotype threat related to the test when in the intelligence condition, there was not a decrement in test performance when using the Raven APM test—a non-verbal measure of intelligence. Thus, although stereotype threat does appear to persist even with alternate forms of cognitive ability testing, it is possible that non-verbal measures can help attenuate its deleterious effects.

Finally, the reputation of an organization can influence the extent to which applicants trust its hiring procedures. For example, when participants were led to believe that a test was “race-fair”, they had more trust and subsequently did not perform as poorly (Steele, 1999). Thus, perhaps procedurally just organizations invoke trust from applicants and this can help placate the negative influence of stereotype threat on test performance.

Limitations

One limitation of this study is that we purport to create a realistic environment, but the experiment is conducted in the laboratory with college students. Although the manipulation and personnel selection context were explicitly set up to create a realistic environment, the generalizability of the results are uncertain.

A second limitation of this experiment is that participants were not pre-selected based on their identification with intelligence as prescribed by some researchers (Steele & Aronson, 1995). However there is some inconsistency about what constitutes domain identification as some researchers have used SAT score (Aronson et al., 1999; Schmader, 2002; Shih, Pittinsky, & Ambady, 1999, Spencer et al., 1999), others have used the reputation of the academic institution (Steele & Aronson, 1995), some have asked a single question about the domain importance (Cheryan & Bodenhausen, 2000), and a large number of published studies make no mention of pre-selecting participants based on domain identification (Blascovich et al., 2001; Brown & Josephs, 1999; Quinn & Spencer, 2001; Stangor et al.,

1998). Further, a recent meta-analysis assumes that all college students are at least moderately identified with intelligence (Jones & Stangor, 2002). Despite the inconsistency in the operationalization of domain identification, this study fails to pre-select participants based on any of these alternative tactics.

A third limitation of this study is that we only gave participants 20 min to work on the Raven's APM. Typically, the Raven's test is administered for 45 min. This might be problematic because the Raven's APM was constructed so that questions increase in difficulty. Thus, it is possible that the dependent variable used in the study was not difficult enough to induce a substantial amount of threat. Although this might be a limitation, it should be noted that anecdotal evidence from watching participants take the exam as well as comments made during participants' debriefing suggests that most participants found the exam to be challenging.

Finally, one limitation that extends to the entire stereotype threat empirical literature is that it has basically relied on one type of research methodology (i.e., experimentation) to assess whether stereotype threat accounts for racial group mean differences on cognitive ability tests. As Sackett et al. (2001) pointed out, this experimental paradigm only provides evidence that the mean differences on cognitive ability tests can be enhanced. It does not provide evidence that stereotype threat is responsible for the mean differences obtained in applied settings. Rather, a key to determining how useful an explanation stereotype threat is for explaining existing mean differences across racial subgroups is to use measures of stereotype threat in applied settings. If the measures were administered in a real testing context and the influence of these measures were statistically controlled for, the ability of the stereotype threat to eliminate mean differences could be directly tested. Unfortunately, the sample and experimental design of this study did not allow for such analyses.

CONCLUSION

The results of this study demonstrate that African Americans who are told a test is diagnostic of intelligence experience more stereotype threat related to that test. In addition, both stereotype threat related to the test as well as a more general sense of threat have deleterious effects on tests performance. Although no mediators were found, the three models presented provide an examination of the relationships between two types of self-report measures of stereotype threat (i.e., specific and general), potential mediators, and test performance. These results lend support for the stereotype threat effect with African Americans on a non-verbal measure of cognitive ability even when a subtle manipulation is used; however, this effect may not be particularly strong and future research must be conducted to uncover its underlying mechanisms and its generalizability.

REFERENCES

- 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.
- Arvey, R. D., Strickland, W., Drauden, G., & Martin C. (1990). Motivational components of test-taking. *Personnel Psychology, 43*, 695–716.
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.
- Blascovich, J., Spencer, S. J., Quinn, D., & Steele, C. (2001). African Americans and blood pressure: The role of stereotype threat. *Psychological Science, 12*, 225–229.
- Brown, R. P., & Josephs, A. (1999). A burden of proof: Stereotype relevance and gender differences in math performance. *Journal of Personality and Social Psychology, 76*, 246–257.
- Brutus, S., & Ryan, A. M. (1998). A new perspective on preferential treatment: The role of ambiguity and self-efficacy. *Journal of Business and Psychology, 13*, 157–178.
- Cheryan, S., & Bodenhausen, G. V. (2000). When positive stereotypes threaten intellectual performance: The psychological hazards of “model minority” status. *Psychological Science, 11*, 399–402.
- Croizet, J., & Claire, T. (1998). Extending the concept of stereotype threat to social class: The intellectual underperformance of students from low socioeconomic backgrounds. *Personality and Social Psychology Bulletin, 24*, 588–594.
- Hayes, S., & Ployhart, R. E. (2002). Using cultural mistrust to understand Black–White test taking and test performance differences. Manuscript under review.
- Helms, J. E. (1990). *Black and White racial identity: Theory, research, and practice*. New York: Greenwood.
- Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72–98.
- 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.
- Jensen, A. R. (1980). *Bias in Mental Testing*. New York: Free Press.
- Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.
- Jones, P. R., & Stangor, C. (2002). The mediators and moderators of stereotype threat: A meta-analysis. Manuscript under review.
- Lee, Y., & Ottati, V. (1995). Perceived in-group homogeneity as a function of group membership salience and stereotype threat. *Personality and Social Psychology Bulletin, 21*, 610–619.
- Levy, B. (1996). Improving memory in old age through implicit self-stereotyping. *Journal of Personality and Social Psychology, 71*, 1092–1107.
- McKay, P. F., Doverspike, D., Bowen-Hilton, D., & Martin, Q. D. (2002). Stereotype threat effects on the Raven Advanced Progressive Matrices scores of African Americans. *Journal of Applied Social Psychology, 32*, 767–787.
- Muthen, L. K., & Muthen, B. O. (1998). *M-Plus user's guide*. Los Angeles, CA: Author.
- Oswald, D. L., & Harvey, R. D. (2000). Hostile environments, stereotype threat, and math performance among undergraduate women. *Current Psychology, 19*, 338–355.
- Pinel, E. C. (1999). Stigma consciousness: The psychological legacy of social stereotypes. *Journal of Personality and Social Psychology, 76*, 114–128.
- Ployhart, R. E., Ziegert, J. C., & McFarland, L. A. (2003). Understanding racial differences on cognitive ability tests in selection contexts: An integration of stereotype threat and applicant reactions research. *Human Performance, 16*, 231–259.
- Quinn, D. M., & Spencer, S. J. (2001). The interference of stereotype threat with women's generation of mathematical problem-solving strategies. *Journal of Social Issues, 57*, 55–71.

- Raven, J. C., Court, J. H., & Raven, J. (1985). *A manual for Raven's Progressive Matrices and vocabulary scales*. London: H. K. Lewis.
- Sackett, P. R., Schmitt, N., Ellingson, E. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing and higher education. *American Psychologist, 56*, 302–318.
- Sarason, I. G. (1972). Experimental approaches to test anxiety: Attention and the uses of information. In C. D. Spielberger (Ed.), *Anxiety: Current trends in theory and research* (pp. 383–401). New York: Academic.
- Schmader, T. (2002). Gender identification moderates the stereotype threat effects on women's math performance. *Journal of Experimental Social Psychology, 38*, 194–201.
- Shih, M., Pittinsky, T. L., & Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological Science, 10*, 80–83.
- 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.
- Spielberger, C. D., Gorsuch, R. R., & Lushene, R. (1970). *The State-Trait Anxiety Inventory (STAI) test manual*. Palo Alto, CA: Consulting Psychologists Press.
- Stangor, C., Carr, C., & Kiang, L. (1998). Activating stereotypes undermines task performance expectations. *Journal of Personality and Social Psychology, 75*, 1191–1197.
- Steele, C. M., & Aronson, J. (1995). Contending with a stereotype: African-American intellectual test performance and stereotype threat. *Journal of Personality and Social Psychology, 69*, 797–811.
- Steele, C. M. (1999, August). Thin ice: "Stereotype threat" and Black college students. *The Atlantic Monthly, 44*–54.
- Stone, J., Lynch, C. I., Sjomeling, M., & Darley, J. M. (1999). Stereotype threat effects on Black and White athletic performance. *Journal of Personality and Social Psychology, 77*, 1213–1227.

Copyright of Human Performance is the property of Lawrence Erlbaum Associates and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.