Journal of Applied Psychology
2006, Vol. 91, No. 4, 979–985

0021-9010/06/$12.00 DOI: 10.1037/0021-9010.91.4.979

The Difference Isn't Black and White: Stereotype Threat and the Race Gap on Raven's Advanced Progressive Matrices

Ryan P. Brown and Eric Anthony Day
The University of Oklahoma

This study addresses recent criticisms aimed at the interpretation of stereotype threat research and methodological weaknesses of previous studies that have examined race differences on Raven's Advanced Progressive Matrices (APM). African American and White undergraduates completed the APM under three conditions. In two threat conditions, participants received either standard APM instructions (standard threat) or were told that the APM was an IQ test (high threat). In a low threat condition, participants were told that the APM was a set of puzzles and that the researchers wanted their opinions of them. Results supported the stereotype threat interpretation of race differences in cognitive ability test scores. Although African American participants underperformed Whites under both standard and high threat instructions, they performed just as well as Whites did under low threat instructions.

Keywords: Stereotype threat, race, Raven's Advanced Progressive Matrices

Proponents of biological explanations for racial disparities on cognitive ability tests have noted that differences between African Americans and Whites are consistent and substantial (typically half to over one standard deviation) even on "culture-free" tests such as Raven's APM (Jensen, 1998). Because its stimuli are nonverbal and do not require a specific knowledge base to be understood, the APM should not be heavily influenced by a respondent's acquired knowledge or reading ability (Saccuzzo & Johnson, 1995). This has led many experts to argue that it is among the purest available measures of general cognitive ability and complex reasoning (e.g., Carpenter, Just, & Snell, 1990; Humphreys, 1984; Jensen, 1980; Snow, Kyllonen, & Marshalek, 1984). Thus, race differences on tests such as Raven's APM are sometimes taken as evidence that the lower test scores of African Americans and other minorities (relative to Whites or Asian Americans) can be attributed to biological differences rather than educational or other environmental factors (e.g., Herrnstein & Murray, 1994; Jensen, 1998).

However, some recent thinking calls this assumption into question. Specifically, research by Steele and colleagues (Steele, 1997; Steele & Aronson, 1995; Steele, Spencer, & Aronson, 2002) on the phenomenon of stereotype threat suggests a plausible environmental explanation that functions independently of item content. In this research, Steele and others have shown that individuals who feel stigmatized with respect to a particular performance domain can experience heightened levels of performance pressure under circumstances in which their stigma seems relevant and salient. This stereotype-driven pressure can then hinder performance, leading to a type of self-fulfilling prophecy. Awareness of a stereotype of inferiority produces an evaluative threat of confirming that stereotype (or being judged by others to have done so), which then leads to the poorer performance predicted by the stereotype.

Evidence pertaining to the stereotype threat hypothesis has grown considerably over the last decade. Researchers have shown, for instance, that African American test takers may perform significantly worse on a difficult verbal test when they are told that it measures their abilities (thus making their intellectual stigma relevant) than when told that it is nondiagnostic of ability (thus negating the relevance of the stigma; Steele & Aronson, 1995). Similar effects have also been obtained with other stereotyped groups, such as Latinos (Gonzales, Blanton & Williams, 2002), the poor (Croizet & Claire, 1998), and women in mathematics (Brown, Charnsangavej, Keough, Newman, & Rentfrow, 2000; Brown & Josephs, 1999; Brown & Pinel, 2003; Spencer, Steele, & Quinn, 1999).

Clearer evidence that reducing stereotype threat can likewise reduce race differences on nonverbal cognitive ability tests like the APM would make a substantial contribution to our understanding of the environmental factors that contribute to race differences on cognitive ability tests. Attempts to obtain such evidence, however, have failed to produce compelling data supporting the stereotype threat interpretation of race differences. For example, McKay, Doverspike, Bowen-Hilton, and Martin (2002) administered the APM to African American and White participants and told them either that the test was a measure of IQ or that it was a measure of pattern completion skills, similar to the test-diagnosticity manipulation of Steele and Aronson (1995). Although McKay et al. found a strong and significant race main effect, they found only a marginal Race × Condition interaction predicting APM scores. Moreover, none of the within-race means was significantly different from one another. Thus, African Americans did not perform significantly worse in the diagnostic than in the nondiagnostic condition. However, prior to the administration of the APM, the researchers gave participants measures of test anxiety and racial identity, as well as demographic questions asking about participants'

Ryan P. Brown and Eric Anthony Day, Department of Psychology, The University of Oklahoma.

We thank Winfred Arthur and Joseph Rodgers for their helpful comments on a previous draft of this article.

Correspondence concerning this article can be sent to Ryan Brown at Department of Psychology, The University of Oklahoma, 455 W. Lindsey, DHT #705, Norman, OK 73019. E-mail: rpbrown@ou.edu

race and sex. If participants thought about these measures at all while completing the APM, the ability of the researchers to reduce stereotype threat among African American participants might have been seriously compromised (Steele & Aronson, 1995).[1]

Likewise, Mayer and Hanges (2003) investigated race differences on the APM in a simulated personnel-selection context. These researchers used a manipulation of test diagnosticity similar to that of Steele and Aronson (1995) and McKay et al., (2002), but did not find a Race × Condition interaction with respect to performance. However, the nondiagnostic condition in the Mayer and Hanges study does not appear to have truly reduced the apparent diagnosticity of the test, an inference that is bolstered by the authors' own manipulation check on participants' stereotype threat perceptions (for a similar conclusion, see Steele and Davies [2003]). Furthermore, Mayer and Hanges only gave participants 20 minutes to complete the APM (40 minutes is more typical), which is problematic because the items on the APM increase in difficulty, and stereotype threat effects have been posited to increase as item difficulty increases (e.g., Spencer et al., 1999). Thus, prior studies of stereotype threat effects on Raven's APM scores seem inconclusive.

Together with these questionable tests of stereotype threat theory, Sackett and colleagues (Sackett, Hardison, & Cullen, 2004; Sackett, Schmitt, Ellingson, & Kabin, 2001) have raised a number of criticisms of the literature on this phenomenon, particularly as it relates to differences between African Americans and Whites in cognitive ability test scores. Their primary criticism concerns the statistical analyses used in the original stereotype threat studies by Steele and Aronson (1995), which involved controlling for SAT scores across all participants before examining the effects of the threat manipulations on test performance. This approach, Sackett argues, produces adjusted mean performance levels that can deceptively portray the difference between African Americans and Whites as disappearing under low threat conditions, when all that has really been shown is that the normal gap between African Americans and Whites has been exacerbated under high threat conditions. Although subsequent stereotype threat studies have not all used this analysis of covariance technique, it is also the case that most studies since Steele and Aronson (1995) have examined gender differences rather than race differences. Thus, the concern expressed by Sackett and colleagues about how to interpret the role played by stereotype threat in the gap between African Americans and Whites has merit and should be taken seriously.

Additional concerns about the ability to generalize stereotype threat findings have also been raised by Sackett et al. (2001). The first concern is that Steele and Aronson's (1995) research examined students at an elite university (Stanford) who might not be representative of test takers in the general population. Consistent with this criticism, one might infer that the reason McKay et al. (2002) and Mayer and Hanges (2003) failed to find significant stereotype threat effects on the APM had to do with their use of less selective samples, rather than with a methodological flaw in their studies. After all, Steele (1997) has suggested that stereotype threat should be strongest among individuals who are highly identified with a performance domain. Although subsequent studies have found stereotype threat effects among students at public universities (e.g., Brown & Josephs, 1999; Spencer et al., 1999), these studies have not examined differences between African Americans and Whites.

A second concern raised by Sackett and colleagues (2001) is that the typical instructions given to participants in stereotype threat studies are not ecologically valid. Indeed, they argue that an alternative interpretation of the stereotype threat literature to date is that race differences have simply been increased under artificially created high threat conditions that would never occur in applied settings. For instance, outside of a stereotype threat laboratory study, test administrators would rarely indicate to test takers that they were examining (or expecting to find) race differences on a test of cognitive ability. As with the previous criticisms leveled at stereotype threat studies, this one, too, does not adequately reflect the range of experimental manipulations that have been used in this literature, but it does raise yet another question about the applicability of stereotype threat findings from the laboratory.

The present study was an attempt to improve upon the experimental designs used in previous studies of stereotype threat to understand race differences on the APM. First, we avoided inadvertently priming participants with any suggestion that we were interested in race differences prior to the experiment. Second, we were very careful to create a low threat condition in which any suggestion that the APM would be used to assess or evaluate participants' intellectual abilities was eliminated. In this condition, the APM items were referred to as "puzzles," and participants were told that we were interested in their evaluations of the task (rather than the task's evaluation of them). Third, we gave participants 40 minutes to complete the test, increasing the probability that all participants would encounter the more difficult test items on the APM. Fourth, our sample was composed of students from only a modestly selective public university in the southwestern United States, allowing for greater generalization than is possible with students from highly selective schools like Stanford. Fifth, the statistical analysis we used avoided the interpretive problems inherent in the ANCOVA approach used by Steele and Aronson (1995). Finally, we also included two conditions designed to elicit stereotype threat: one condition that explicitly presented the test as an IQ measure (high threat) and one condition that used the standard wording that is supposed to be used with the APM (standard threat). The inclusion of both of these conditions allowed us to determine not only whether stereotype threat can occur under novel circumstances with the APM (i.e., when participants are told that the APM is a measure of IQ) but also whether stereotype threat can occur under standard testing conditions. This standard condition thus helps to address the criticism by Sackett and others (2004) that race differences in previous studies have simply been increased under high threat conditions rather than decreased in low threat conditions. Consistent with the extensive literature on IQ and cognitive ability, we predicted that Whites would perform better than African Americans on the APM in both the standard threat and high threat conditions. However, consistent with the stereotype threat hypothesis, we predicted that this performance gap between African Americans and Whites would be significantly reduced in the low threat condition.

[1] It is also possible that describing the APM as a measure of "pattern completion skills" failed to remove all of the evaluative threat of the test in that a measure of any kind of cognitive "skill" might seem stereotype-relevant to intellectually stigmatized individuals.

Method

Participants

Fifty-nine African American students and 83 White students enrolled in introductory psychology courses at The University of Oklahoma participated in partial fulfillment of a course research requirement. Data from 4 participants (3 African American men and 1 African American woman) were excluded from analyses because they failed to report an ACT score, and data from 2 additional participants (1 African American man and 1 African American woman) were excluded because they did not believe the cover story about the purpose of the study. The mean age of the final sample (53 African American participants and 83 White participants) was 19.23 years ($SD = 3.0$). Seventy percent of the final sample were women.

Test Performance

We used scores on Raven's APM (Raven, Raven, & Court, 1998) to operationalize performance. The APM consists of 36 design problems arranged in an ascending order of difficulty. Each problem involves a logical pattern with a piece missing. The respondent's task is to select the piece that best completes the pattern from eight alternatives. The APM test manual reports a test–retest reliability of 0.91 and strong evidence of convergent validity (Raven et al., 1998). We obtained a Spearman-Brown odd-even split-half reliability estimate of 0.82 for the APM scores. We also examined the extent to which the reliability estimate varied as a function of condition. As shown in Table 1, no statistically reliable differences were observed.

Design and Procedure

Data were collected within a 2 (race: African American or White) × 3 (instructions: standard threat, high threat, or low threat) between-groups design. Each participant was tested individually in a small room. After having participants read and sign consent forms, the experimenter explained that the purpose of the study was to examine how different people work on puzzle-solving tasks. The experimenter then provided participants with written instructions for the APM, which were also verbally summarized, according to one of the three experimental conditions, to which participants were randomly assigned within race.

In the standard threat condition, participants were given instructions consistent with the published manual for the APM (Raven et al., 1998). Specifically, participants were told, "The Advanced Progressive Matrices is a measure of observation and clear thinking." The APM was referred to as a "test" several times during these instructions. Participants in the high threat condition were told, "The task you will be working on is an IQ test. Like the SAT and ACT, this test is frequently used to measure individuals' intelligence and ability." The items were referred to as an "IQ test" several times during these instructions. Finally, in the low threat condition, participants were told, "The task you will be working on is a series of puzzles. Please take these puzzles seriously. When you are finished working on the puzzles, we would like to ask you some questions about the puzzles and get your thoughts and reactions about them." At no time during the low threat instructions were the items referred to as a test; rather, items were consistently referred to as "puzzles." By combining this consistent reference to the APM items as puzzles with the indication that we desired participants' feedback about the items, we intended to remove any suggestion that the task was evaluative in nature—thus reducing the relevance of racial stereotypes about intellectual abilities. After providing these condition-specific instructions, the experimenter (a White man) provided all participants with specific instructions on how the items work.

All participants had 40 minutes to work on the APM. Afterward, participants answered some demographic questions and reported their best score on the ACT, as well as what they believed was the purpose of the task they had just completed. Finally, participants were probed for suspicion and were fully debriefed about the true purposes of the study.

Table 1
Spearman-Brown Odd-Even Split-Half Reliabilities of APM Scores by Condition

	White			African American		
	n	r_{xx}	95% C.I.	*n*	r_{xx}	95% C.I.
Low threat	27	.68	.28; .85	17	.88	.65; .96
Standard threat	29	.77	.49; .89	19	.80	.47; .93
High threat	27	.84	.65; .93	17	.91	.75; .97

Note. C.I. = confidence interval.

Results

Previous stereotype threat studies that have examined race differences in performance on cognitive ability tests (e.g., Steel & Aronson, 1995) have typically included standardized test scores (e.g., SAT) as a covariate to control for prior ability differences. However, this approach can be criticized for its circularity because the meaning of group differences in ability is exactly what is being explained (Sackett et al., 2001, 2004). To avoid this circularity but still equate experimental conditions for prior ability within racial groups, we followed the approach advocated by Brown and Josephs (1999), which was a modified covariance analysis on within-race residualized performance means. Specifically, we first conducted simple regression analyses within each race, regressing APM scores on ACT scores across experimental conditions and saving the residuals for further analysis. Second, after adding the race-specific mean performance levels to the unstandardized residuals of these separate regression analyses, we submitted the residualized test scores to a 2 (race: African American or White) × 3 (instructions: standard threat, high threat, or low threat) between-groups general linear model analysis of variance (GLM-ANOVA). This approach effectively controls for within-race differences in ACT scores across experimental conditions without making adjustments to performance means across race. As shown in Table 2, such an adjustment was important in the present study in that the mean ACT scores of African American participants differed across experimental conditions, particularly between the standard threat and high threat conditions. The statistical power for finding medium ($\eta^2 = .06$) and large ($\eta^2 = .14$) interaction effects with our design and sample size was 73% and 99%, respectively. The statistical power for finding significant differences between African Americans in the low threat condition and the standard and high threat conditions was 51% (for a medium effect; $d = 0.50$) and 85% (for a large effect; $d = 0.80$). In their seminal research, Steele and Aronson (1995; Study 2) found a large Race × Condition interaction ($\eta^2 = .18$) and a large difference ($d = 0.80$) between African Americans in a diagnostic versus a nondiagnostic condition.[2]

The mean residualized APM scores in each condition are shown in Table 2, along with raw APM scores unadjusted for ACT. Our

[2] These effect sizes were calculated based on the *F*, *t*, and degrees of freedom reported by Steele and Aronson (1995). However, it should be noted that effect sizes derived from Steele and Aronson (1995) are not in the same metric as the effect sizes observed in the present study because Steele and Aronson partialled SAT scores out of between-group differences.

Table 2
Means and Standard Deviations by Condition

	White		African American		
	M	*SD*	*M*	*SD*	*d*[a]
ACT scores					
Low threat	25.11	3.82	22.88	3.28	0.62
Standard threat	24.83	4.15	24.11	4.81	0.16
High threat	24.81	3.44	21.06	2.79	1.18
Raw APM scores					
Low threat	22.44	4.56	24.29	5.05	−0.39
Standard threat	24.28	4.08	22.42	5.15	0.41
High threat	24.67	3.89	19.41	5.60	1.16
Adjusted APM scores[b]					
Low threat	22.35	3.94	24.21	4.06	−0.47
Standard threat	24.32	3.55	21.35	4.56	0.75
High threat	24.72	3.55	20.79	4.50	1.00
Accuracy[b]					
Low threat	0.67	0.13	0.70	0.13	−0.23
Standard threat	0.71	0.12	0.62	0.16	0.66
High threat	0.74	0.10	0.59	0.13	1.34
Items attempted[b]					
Low threat	33.84	3.69	34.90	1.76	−0.36
Standard threat	34.70	2.51	34.82	2.43	−0.05
High threat	33.60	3.37	35.30	1.80	−0.62

[a] Effects calculated as mean for Whites minus mean for African Americans. [b] Scores are adjusted within race for prior ACT scores.

analysis revealed a main effect of race, $F(1, 130) = 5.78$, $p < .05$, $\eta^2 = .04$, such that African Americans ($M = 22.12$, $SD = 4.55$) scored lower on average than did Whites ($M = 23.80$, $SD = 3.78$). In addition, the analysis revealed a Race × Instructions interaction, $F(2, 130) = 6.44$, $p < .01$, $\eta^2 = .09$, $MSE = 15.73$. Consistent with our predictions, planned comparisons revealed that although the difference between African Americans and Whites under standard testing instructions replicated the typical race difference favoring Whites, $F(1, 130) = 6.44$, $p < .05$, $d = 0.75$, a difference that was even larger in the high threat condition, $F(1, 130) = 10.22$, $p < .01$, $d = 1.00$, this difference was actually reversed in the low threat condition, although the difference here was not statistically significant, $F(1, 130) = 2.30$, $p > .10$, $d = -0.47$. Likewise, and in contrast to the results reported by McKay et al. (2002), African Americans in the low threat condition scored significantly better than did African Americans in either the standard or the high threat conditions (ps $< .05$). It is especially noteworthy that African Americans in the low threat condition performed as well as Whites did in either the standard or high threat conditions. Thus, removing the relevance of the racial intelligence stereotype increased the APM scores of African American participants to essentially the same level as that of White participants in the conditions in which Whites performed their best. Adding gender of participant as a covariate in these analyses did not change any of our conclusions, and gender itself was not a significant predictor of ACT-adjusted APM scores ($r = .16$, $p > .05$; male = 1, female = 0).

Further examination revealed that these performance effects were largely due to decrements in performance accuracy (number of problems answered correctly divided by number attempted) rather than mere persistence. As with basic performance, there was a significant race main effect on accuracy (residualized within race for prior ACT scores), $F(1, 130) = 8.57$, $p < .01$, $\eta^2 = .06$, and a significant Race × Instructions interaction, $F(1, 130) = 5.20$, $p < .01$, $\eta^2 = .07$, $MSE = 0.016$. As shown in Table 2, African American participants were significantly less accurate than Whites in the standard and high threat conditions (ps $< .05$). In contrast, African American participants in the low threat condition were slightly, but not significantly, more accurate than Whites in the low threat condition, $F < 1$. The Race × Instructions interaction was not significant for number of problems attempted (also adjusted within race for prior ACT scores), $F < 1$, although African Americans overall ($M = 35.0$, $SD = 2.01$) did answer slightly more items than did Whites ($M = 34.06$, $SD = 3.21$), $F(1, 130) = 3.77$, $p = .05$, $d = 0.33$.

In a final exploratory analysis, we also examined item difficulty curves as a function of race and experimental condition. Accordingly, we regressed item difficulties (i.e., % correct) on race and item number of the APM, including the quadratic term for item number as well as all interaction terms. We observed no significant interactions involving race. Likewise, we observed similar curves across all three experimental conditions. Specifically, there was a systematic increase in difficulty across items in all three conditions, with a more pronounced increase in difficulty for the later items of the test (i.e., a significant quadratic effect). These item difficulty curves are similar to the results reported in the APM manual (Raven et al., 1998) and by Arthur and Day (1994).

Discussion

The results of the present study offer strong support for the hypothesis that race differences in cognitive ability test scores could be accounted for with a simple, contextual variable that is independent of biological factors and even test content. Specifically, when African American participants were told that the Raven's APM was merely a set of puzzles for which we wanted their feedback, they performed significantly better than they did when they were told that the APM was either a test of "observation and clear thinking" (the standard instructions used with the APM) or an IQ test. Indeed, the differences within African American participants were not only statistically significant, they were also substantial—approximately three fourths of a standard deviation—in stark contrast to recent claims that stereotype threat effects are "typically very small" (Reeve & Hakel, 2002). The present results also stand in contrast to several previous studies that failed to find convincing evidence of stereotype threat on the APM (e.g., Mayer & Hanges, 2003; McKay et al., 2002), but which, as noted previously, suffered from methodological weaknesses that precluded an adequate test of the stereotype threat hypothesis.

That the performance effects we observed in this study were a function of accuracy rather than problems attempted is also worth noting. This result is consistent with several previous findings (e.g., Shih, Pittinsky, & Ambady, 1999; Steele & Aronson, 1995), including evidence that stereotype threat influences problem-solving strategies (Quinn & Spencer, 2001) and reduces working memory capacity (Schmader & Johns, 2003). Thus, this pattern of results suggests that stereotype threat may operate by impeding the production of correct answers rather than by diminishing effort or motivation. However, several studies have also demonstrated that

Table 3
APM–ACT Correlations by Condition

	White		African-American		Collapsed across race	
	r	95% C.I.	*r*	95% C.I.	*r*	95% C.I.
Low threat	.51	.16; .75	.60	.17; .84	.45	.18; .66
Standard threat	.50	.16; .73	.52	.09; .79	.51	.26; .69
High threat	.41	.04; .68	.64	.23; .86	.62	.40; .77
Collapsed across threat	.46	.40; .70	.56	.22; .65	.53	.40; .64

Note. C.I. = confidence interval.

stereotype threat can induce behavioral self-handicapping—a strategic reduction in preparation effort prior to performance designed to reduce the impact of failure on the self-concept (e.g., Brown, 1999; Stone, 2002). In sum, stereotype threat could impede performance through multiple mechanisms, any of which could be sufficient to produce group differences in performance.

Perhaps the most important finding was that although African Americans in the standard and high threat conditions underperformed their White counterparts to the same degree that they do in normative samples (0.75 to 1.0 standard deviation), this difference disappeared in the low threat (puzzles) condition. Unlike some previous stereotype threat studies, these results cannot be criticized on the grounds that we statistically adjusted the APM scores for prior ability levels (thus removing any prior race differences in ability from the performance means; Sackett et al., 2001, 2004) because we made this adjustment only within race to control for ability differences across experimental conditions. Thus, the African American participants in the low threat condition scored about as well as the White participants in the standard and high threat conditions, despite having scored lower than Whites on the ACT. Because prior differences in ACT scores were not covaried out of APM scores *across race* in our analyses, our experimental manipulation appears to account for the entire race difference in APM performance in this study, contrary to recent criticisms by Sackett et al. (2004). Likewise, the less select nature of the sample in the present study does not support the contention that stereotype threat only explains differences between African Americans and Whites among students at elite universities (Sackett et al., 2001).

Despite the strength of the present results, three important limitations to this study are apparent. First, as with most prior stereotype threat studies, the present results are not accompanied by any direct mediating evidence (cf. Schmader & Johns, 2003) or even a manipulation check to verify that stereotype-related concerns were primed in African American participants by our instructions. Thus, we cannot be sure that stereotype threat was, in fact, the mechanism that produced the performance effects in our study. Because of this, it is plausible that mechanisms unrelated to social stereotypes accounted for our observed performance effects. For example, one could hypothesize that unfamiliarity with tests such as the APM might be more likely with African Americans than Whites. This lack of familiarity could be the basis for an evaluative threat that might explain at least part of the race gap in performance.[3] Although previous research on stereotype threat supports the view that manipulations such as ours can lead to stereotype-based performance concerns (e.g., Brown & Josephs, 1999; Brown & Pinel, 2003; Steele & Aronson, 1995), caution in generalizing from those studies to the present investigation is warranted. The difficulty of measuring stereotype activation without inadvertently priming stereotypes in low threat conditions remains a challenge to researchers searching for mediational evidence, and this challenge makes additional research in this area an important pursuit.

A second limitation of our research is that we do not present any evidence that the condition designed to reduce stereotype threat among African American participants did not also affect the *predictive validity* of their APM scores. The extent to which stereotype threat influences predictive validity will depend on the degree to which stereotype threat differentially influences predictor and criterion scores (see Cullen, Hardison, & Sackett, 2004). Indeed, the possibility that predictor variables such as the SAT and ACT are influenced by stereotype threat to similar degrees as criterion variables such as college GPA (Aronson, Fried, & Good, 2002; Brown et al., 2000; Brown & Lee, 2005) might explain why differential predictive validity for African Americans and Whites is not typically observed with standardized cognitive ability tests. Although our study did not include a criterion measure for assessing predictive validity, the same reasoning might apply to the associations between APM scores and participants' prior ACT scores. Thus, we might expect to find the largest APM–ACT correlations within the standard and high threat conditions and the lowest correlations within the low threat conditions. Table 3 shows these correlations across the six experimental conditions. As shown, there were no reliable differences among the six conditions. Collapsing these correlations across race reveals a pattern that does conform more to the above expectations, although again the differences were not statistically reliable. Future studies that manipulate stereotype threat during the administration of a cognitive ability test and that assess its predictive validity with respect to subsequent performance indices (e.g., GPA) would greatly enhance our understanding of how stereotype threat relates to the criterion validity of test scores. Of course, different criterion measures might vary in the extent to which they are susceptible to stereotype threat effects, and this susceptibility would also influence the predictive validity of a cognitive ability test. Investigating the extent to which (and reasons why) different kinds of criterion measures (e.g., tests of job knowledge, supervisory ratings) are

[3] We acknowledge an anonymous reviewer for suggesting this alternative explanatory framework for our findings.

susceptible to stereotype threat effects could be a worthwhile avenue for future research.

A third limitation to our study is that we administered the APM in an individual rather than a group setting, and this is not often the case when cognitive ability tests are given outside the laboratory for predictive purposes. Although previous research has shown that the social make-up of the testing environment itself can be enough to induce stereotype threat (e.g., Inzlicht & Ben-Zeev, 2000), additional research on this issue is clearly warranted. It may be, for instance, that reducing stereotype threat effects will prove more challenging in group contexts than in individual ones. Alternatively, the feeling of anonymity and related dynamics deriving from large-group settings could conceivably be enough to reduce stereotype threat effects. All such possibilities, of course, have important implications for the ways in which cognitive ability tests are used and administered (see also Inzlicht & Ben-Zeev, 2003).

Overall, the data in the present study were consistent with predictions derived from stereotype threat theory. Somewhat surprisingly, though, Whites in the low threat condition actually performed worse than Whites in the other conditions. Although not predicted, this trend is consistent with results reported by McKay et al. (2002), who found a somewhat smaller difference in the same direction among Whites in the nondiagnostic compared to the diagnostic condition. Whether this difference represents a decrement among White participants in the low threat (or nondiagnostic) condition (as McKay et al. argued) or an increment in the standard and high threat (or diagnostic) conditions is a matter for speculation. However, the latter interpretation is consistent with a meta-analysis by Walton and Cohen (2003) on "stereotype lift" effects among positively stereotyped groups taking ability-diagnostic tests.

Conclusions and Implications

This study has important implications for how we understand race differences on cognitive ability tests. What must always be kept in mind when interpreting such differences is that a score on a cognitive ability test reflects an individual's *performance* on that test, which is a function of developed ability, motivation, and a host of other influences. One of these "other influences" may be stereotype threat, and the million-dollar question is whether this nongenetic factor can account for a substantial portion of the performance gap between African Americans and Whites. The present study suggests that it can. Thus, whatever contribution genes might make to individual differences in cognitive ability within race, it appears that stereotype threat might be capable of accounting for a substantial portion of the mean difference between races. To the extent that environmental factors such as stereotype threat can influence performance, the conclusion that race differences in cognitive ability scores are due to intractable, biological differences seems unwarranted. It is important to note, however, that we do not believe that our findings necessarily diminish the importance of other contributing factors in the performance gap between African Americans and Whites, including differences in educational and economic opportunities (Sackett et al., 2004). Rather, we believe that stereotype threat and other context-driven effects, among various social, educational, and economic factors, can play an important role in group differences in achievement.

What are the applied implications of our study? Although this research was carried out in a laboratory environment, we believe that it has several important implications for the world outside the laboratory. First, our use of standard instructions in one condition of our study allows us to avoid the criticism that test administrators must explicitly invoke racial stereotypes or use particularly race-biased language to produce stereotype threat effects (see also Spencer et al., 1999). Our data suggest that just the implication that a test is intellectually evaluative is enough to diminish performance among African American respondents. Second, the size of the effects we obtained might also have important implications for our understanding of when stereotype threat might produce group differences in performance. As Jensen (1998) has noted, the size of race differences on cognitive ability tests is associated with the extent to which tests are *g*-loaded (i.e., the extent to which they capture general intelligence rather than acquired knowledge or skills). But Schmader and Johns (2003) have demonstrated that stereotype threat may influence test performance by reducing working memory capacity, which is itself a strong predictor of performance on highly *g*-loaded tests (Kyllonen & Christal, 1990). Thus, stereotype threat effects might be particularly large on the very tests that past research has shown produce the largest race differences. Consistent with this contention, the size of the race difference that we obtained with the APM under the threat conditions (and that we eliminated in the low threat condition) was almost exactly the size of the typical race difference found on the APM, which is one of the largest race differences in the cognitive ability literature (Jensen, 1998). One implication of all of this might be that attempts to reduce adverse impact on tests that are highly *g*-loaded might benefit more from considerations of stereotype threat effects than would such attempts on less *g*-loaded tests. This, of course, is mere speculation at this point, but it is an interesting possibility that merits investigation.

Third, our results suggest that people interested in reducing adverse impact need not look simply to test content to find a solution to the problem of minority underperformance (e.g., Pine, Church, Gialluca, & Weiss, 1980). Instead, an alternative (and perhaps more productive) approach might be to focus on the testing context. Successfully reducing the stereotype-related evaluative threat of the testing context might go a long way toward reducing race differences on many cognitive ability tests, as it did in the present study with the APM. Of course, we will be the first to admit that simply telling test takers in applied settings that a test like the APM is "just a set of puzzles" or that their opinion of the test is of primary interest is not feasible. When selection tests are given outside of the laboratory, test administrators would be hard pressed to convince anyone that they are not diagnostic of ability. However, recent demonstrations that group differences can be reduced via contextual modifications aimed at reducing race or gender salience (Inzlicht & Ben-Zeev, 2000; Steele & Aronson, 1995; Stricker, 1998), reframing the implications of the test without reducing its apparent diagnosticity (Brown & Josephs, 1999), or altering test-takers' levels of self-construal (e.g., asking individuating questions regarding personal interests and strengths; Ambady, Paik, Steele, Owen-Smith, & Mitchell, 2004) are a source of optimism alongside the challenge of minimizing the impact of stereotype threat in applied settings. We hope that the present study encourages researchers to see the potential gain in such an endeavor.

References

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

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

Arthur, W., Jr., & Day, D. V. (1994). Development of a short form for the Raven Advanced Progressive Matrices Test. *Educational and Psychological Measurement, 54,* 394–403.

Brown, R. P. (1999). *Sex differences in self-handicaps: The relevance of performance stereotypes.* Unpublished doctoral dissertation: The University of Texas at Austin.

Brown, R. P., Charnsangavej, T., Keough, K., Newman, M., & Rentfrow, P. (2000). Putting the "affirm" into affirmative action: Preferential selection and academic performance. *Journal of Personality and Social Psychology, 79,* 736–747.

Brown, R. P., & Josephs, R. A. (1999). A burden of proof: Stereotype relevance and gender differences in math performance. *Journal of Personality and Social Psychology, 76,* 246–257.

Brown, R. P., & Lee, M. N. (2005). Stigma consciousness and the race gap in college academic achievement. *Self & Identity, 4,* 149–157.

Brown, R. P., & Pinel, E. C. (2003). Stigma on my mind: Individual differences in the experience of stereotype threat. *Journal of Experimental Social Psychology, 39,* 626–633.

Carpenter, P. A., Just, M. A., & Snell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices test. *Psychological Review, 97,* 404–431.

Croizet, J. C., & Claire, T. V. (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.

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

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

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

Humphreys, L. G. (1984). General intelligence. In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives on bias in mental testing* (pp. 221–247). New York: Plenum Press.

Inzlicht, M., & Ben-Zeev, T. (2000). A threatening intellectual environment: Why women are susceptible to experiencing problem-solving deficits in the presence of men. *Psychological Science, 11,* 365–371.

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

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.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working memory capacity?! *Intelligence, 14,* 389–433.

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

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.

Pine, S. M., Church, A. T., Gialluca, K. A., & Weiss, D. J. (1980). Effects of computerized adaptive testing on Black and White students. *JSAS Catalog of Selected Documents in Psychology, 10,* 8.

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., Raven, J., & Court, J. H. (1998). *A manual for Raven's Progressive Matrices and Vocabulary Scales.* London: H. K. Lewis.

Reeve, C. L., & Hakel, M. D. (2002). Asking the right questions about *g*. *Human Performance, 15,* 47–74.

Saccuzzo, D. P., & Johnson, N. E. (1995). Traditional psychometric tests and proportionate representation: An intervention and program evaluation study. *Psychological Assessment, 7,* 183–194.

Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2004). On interpreting stereotype threat as accounting for African American – White differences on cognitive tests. *American Psychologist, 59,* 7–13.

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

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

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

Snow, R. E., Kyllonen, P. C., & Marshalek, B. (1984). The topography of ability and learning correlations. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 2, pp. 47–103). Hillsdale, NJ: Erlbaum.

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

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

Steele, C. M., & Davies, P. G. (2003). Stereotype threat and employment testing: A commentary. *Human Performance, 16,* 311–326.

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 34, pp. 379–440). New York: Academic.

Stone, J. (2002). Battling doubt by avoiding practice: The effects of stereotype threat on self-handicapping in white athletes. *Personality and Social Psychology Bulletin, 28,* 1667–1678.

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

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

Received March 1, 2004
Revision received March 21, 2005
Accepted April 29, 2005 ■