HUMAN PERFORMANCE, *16*(3), 231–259

Understanding Racial Differences on Cognitive Ability Tests in Selection Contexts: An Integration of Stereotype Threat and Applicant Reactions Research

Robert E. Ployhart
Department of Psychology
George Mason University

Jonathan C. Ziegert
Department of Psychology
University of Maryland

Lynn A. McFarland
Department of Psychology
George Mason University

This study integrates research on stereotype threat with research on applicant perceptions to examine how these two paradigms jointly enhance the understanding of racial subgroup cognitive ability test differences in selection contexts. A simulated selection context was used so that both stereotype threat and face validity could be manipulated. Participants were 250 White and 144 Black students. Using a 3 (stereotype threat: diagnostic, non-diagnostic, control) × 2 (face validity: face valid, generic) × 2 (race: Black, White) between-subjects design, our results found that stereotype threat interacted with face validity and race, but only for individuals highly identified with their racial group. Results suggested that Blacks performed best when taking the generic test in the control condition, whereas when taking the face valid test, they performed best in the non-diagnostic condition. Across all threat and face validity conditions, Black performance was worst in the diagnostic condition. In addition, correlational analyses found important individual differences in perceptions of stereotype threat, such that these perceptions contributed to lower face validity,

Requests for reprints should be sent to Robert E. Ployhart, Department of Psychology, George Mason University, Fairfax, VA 152. E-mail: rployhar@gmu.edu

lower test-taking motivation, and higher anxiety. Further, motivation positively and anxiety negatively influenced actual test performance. Thus, this study finds that research on stereotype threat and applicant perceptions are complementary, and together contribute to a better understanding of subgroup differences on cognitive ability tests.

Decades of research suggests that cognitive ability tests are among the most predictive and practically efficient predictors of job performance for most occupations (Hunter & Hunter, 1984; Schmidt & Hunter, 1998). Unfortunately, these tests also produce large racial subgroup differences such that Blacks and Hispanics score lower than Whites. These differences are quite robust; they have been noted as far back as Spearman (1927), are manifest with most measures of general ability, are prevalent in most cultures, and are not a result of test bias (Jensen, 1998; Sackett & Wilk, 1994; Schmidt, 1988). Nearly every study of general cognitive ability finds that Whites score higher than Blacks (Sackett, Schmitt, Ellingson, & Kabin, 2001), with the magnitude of such differences ranging from two–thirds to 1 *SD* unit (e.g., Hough, Oswald, & Ployhart, 2001; Jensen, 1998; Roth, Bevier, Bobko, Switzer, & Tyler, 2001). These differences are so large that, with realistic selection ratios, adverse impact against Blacks is nearly guaranteed (Bobko, Roth, & Potosky, 1999; Sackett & Roth, 1996). Given that organizations must face a tradeoff between diversity and optimal selection when using cognitive ability tests, it is not surprising that there is much debate about how this balance could be best achieved (Sackett et al., 2001).

In recent years, two promising programs of research, stemming from different research paradigms, have been devoted towards trying to understand these racial subgroup differences in cognitive ability test performance. In the social psychological literature, a concept known as "stereotype threat" has emerged as a potential explanation for this issue (Steele & Aronson, 1995). Briefly, stereotype threat occurs when an individual believes that there is a negative stereotype of his or her group's (e.g., race, sex) performance in a particular domain; thus, the threat of confirming this negative stereotype interferes with test performance (provided that certain conditions exist, as discussed shortly). Alternatively, research in industrial and organizational (I–O) psychology has examined how a variety of test-taker perceptions (e.g., motivation, anxiety) can relate to racial subgroup test differences (see Ryan & Ployhart, 2000; Schmitt & Chan, 1998, 1999). Both programs of research offer the potential for a better understanding and thus for a reduction of racial subgroup ability test differences (Ryan, 2001; Sackett et al., 2001; Steele, Spencer, & Aronson, 2002), but to date they have tended to remain relatively isolated from each other.

The purpose of this study is to integrate stereotype threat and applicant reactions research to better understand how and why racial differences exist in cogni-

tive ability test performance. We conduct this study in a simulated selection context so that we can manipulate threat and face validity; something that would be impossible to perform ethically in the real world. By integrating the research on stereotype threat with the research on applicant reactions, and by adopting both experimental and correlational methods, this study tests questions that to date have been neglected in both literatures. Such questions concern how stereotype threat operates in employment contexts (from both an experimental perspective and an individual difference perspective), how face validity manipulations may reduce threat, and how individual differences in perceived threat contribute to individual differences in test perceptions and thus lead to test performance. Several recent calls have been made to examine just these kinds of questions (e.g., Sackett et al., 2001; Steele et al., 2002).

To address these questions, we first discuss previous research on stereotype threat. Next, we discuss research on applicant reactions. Finally, we integrate these two traditions to examine how stereotype threat fits with the research on applicant reactions to influence test performance.

STEREOTYPE THREAT IN EMPLOYMENT TESTING CONTEXTS

Stereotype threat is a phenomenon that occurs when an individual is performing in a domain for which a negative stereotype exists about some group to which the person identifies (e.g., racial, gender); the "threat" of the person's behavior confirming this negative group stereotype has a debilitating effect on the individual's performance. The first study documenting stereotype threat for Black ability test performance was conducted by Steele and Aronson (1995). In this research, four laboratory studies examined threat by either manipulating the diagnosticity of the test or the saliency of the negative stereotype. Their results showed that the introduction of stereotype threat impaired the performance of Black test-takers. An important point to note is that prior ability was controlled for in these analyses; thus the findings of Steele and Aronson actually suggest that stereotype threat exaggerates existing subgroup differences (see Sackett et al., 2001, for a greater discussion).

This initial study on stereotype threat is provocative because it may help explain why Blacks perform more poorly than Whites on measures of cognitive ability in selection contexts. Specifically, the negative stereotype about inferior Black intelligence is one that is widespread (e.g., Devine, 1989). When Blacks enter an employment testing context, they face the threat of performing poorly on this test and thus confirming the negative stereotype. This threat may actually depress their test scores relative to other groups (e.g., Whites, Asians) who do not face such negative stereotypes on cognitive ability.

Since the initial study, stereotype threat has been replicated across several different domains and contexts. For example, one interesting study by Aronson, Lustina, Good, Keough, Steele, and Brown (1999) illustrated that White men can have lowered scores on a math test when their ability was compared to Asians. Specifically, in the stereotype threat condition in which the researchers highlighted Asians' superior mathematical ability, White men scored significantly lower. However, in the condition in which no differences between Asians' and Whites' mathematical ability were stressed, the ethnic differences were minimal. This study replicates and extends Steele and Aronson's (1995) original findings by illustrating that highlighting a negative stereotype of a traditionally nonstigmatized group (e.g., White men) can also lead to decreases in performance. Support for the stereotype threat effect has been found in numerous other domains and with multiple groups (Shih, Pittinsky, & Ambady, 1999; Stone, Lynch, Sjomeling, & Darley, 1999). It is worth noting for issues developed later that in all of these studies, stereotype threat is something that is experimentally manipulated.

This previous research has helped identify some important boundary conditions on observing the stereotype threat effect. Steele et al. (2002) noted several such conditions, but the following three are perhaps the most critical. First, the stereotype threat effect is stronger when individuals are identified with the performance domain (e.g., for cognitive ability tests, how central intelligence is to one's self-concept). Second, although less clear, the effect may be stronger when individuals are more identified with the stereotyped group (e.g., when Blacks see more of their self-identify tied to African–Americans in general). Thus, both domain and racial identity may compose two individual difference influences on the strength of the threat effect. Third, the test must be perceived as diagnostic of the stereotyped construct (e.g., intelligence or cognitive ability). Tests that are not perceived as diagnostic are unlikely to invoke perceptions of threat.

Although research has demonstrated the stereotype threat effect, explaining why it impacts test performance (i.e., the psychological mediators of stereotype threat) has to date remained unknown. Several potential mediators have been hypothesized and examined, including anxiety (Spencer, Steele, & Quinn, 1999; Steele & Aronson, 1995, Study 2; Stone et al., 1999), effort (Gonzales, Blanton, & Williams, 2002, Study 1; Steele & Aronson, 1995, Study 2), performance expectancies (Spencer et al., 1999; Stangor, Carr, & Kiang, 1998; Stone et al., 1999), external attributions (Steele & Aronson, 1995, Study 3), and withdrawal from the performance domain (Major, Spencer, Schmader, Wolfe, & Crocker, 1998). However, despite some interesting results, to date there has been no consistently supported mediator. Steele et al. (2002) noted this lack of consistent mediation most likely occurs because such mediators vary across people, situations, and the nature of the stereotype itself. Fortunately, as discussed shortly, research on applicants' reactions has identified several important test perceptions that could serve as relatively consistent mediators in employment testing contexts.

Although it is possible that stereotype threat may help us understand why Blacks score lower on cognitive ability tests than Whites, all of the aforementioned research has been conducted in non-employment testing contexts. Thus, the extent to which existing stereotype threat research may generalize to employment testing contexts is unclear for several reasons. First, many previous studies have used very strong manipulations of diagnosticity, such as explicitly telling participants that the test is diagnostic of a particular construct and that the stereotyped groups perform more poorly on it (see Aronson et al., 1999; Leyens, Desert, Croizet, & Darcis, 2000). Clearly no such statements would be performed as part of employment testing. Second and more important, some have argued that the very nature of using cognitive ability tests in selection contexts may itself be enough to elicit the threat (Steele et al., 2002). Due to the facts that (a) the knowledge of the negative stereotype of Black intellectual functioning is so widespread (Devine, 1989) and (b) many popular measures of cognitive ability can be easily recognized as diagnostic of such, stereotype threat may be almost assured.

Even given these questions of applying threat to an employment context, we propose that stereotype threat could still help to illuminate why racial differences in selection procedures exist. In particular, by understanding applicant test-taker perceptions and reactions, we predict that stereotype threat may yet have important consequences, although in a manner different from its usual treatment in social psychological research.

Test-Taker Perceptions and Applicant Reactions

Within the last decade, a number of studies have found that test-perceptions are related to test performance, that there are sizeable racial subgroup differences on test perceptions favoring Whites (e.g., test-taking motivation, face validity, belief in tests), and that these perceptions may thus contribute to racial subgroup test differences (e.g., Hough et al., 2001; Ryan, 2001; Sackett et al., 2001). For example, Arvey, Strickland, Drauden, and Martin (1990) found that test-taking motivation was lower for Blacks than for Whites, and that controlling for test-taking motivation reduced Black–White differences on a work sample test. In later work, Schmit and Ryan (1992) found that test-taking motivation moderated the criterion-related validity of cognitive ability and personality tests. For cognitive ability, the validity was greater for those with more test-taking motivation.

A program of research conducted by Chan and colleagues (Chan, 1997; Chan & Schmitt, 1997; Chan, Schmitt, Deshon, Clause, & Delbridge, 1997) has provided perhaps the most convincing evidence that applicant perceptions relate to test performance. Chan (1997) found that Whites perceive cognitive ability tests as more valid than Blacks. Chan et al. (1997) further showed that face validity influences test-taking motivation, which in turn influences test performance. In this study, Whites and Blacks completed two parallel versions of a cognitive ability test, but

between administrations, participants also completed measures of face validity and test-taking motivation. Although initial test performance was a strong determinant of performance on the second test, initial test performance also impacted face validity. Face validity then impacted test-taking motivation, which influenced test performance on the second test. This study is particularly important because it suggests that if the face validity perceptions, and thus test-taking motivation, of Blacks can be improved, subgroup differences in test performance may be reduced. It is important to note that nearly all of the aforementioned studies were correlational in nature, and one should be careful about making strong causal inferences.

Thus, this applicant reactions research suggests that there are racial differences in test perceptions, and that these test perceptions may influence test performance (e.g., Chan, Schmitt, Sacco, & DeShon, 1998). One implication of this research is that making tests more face valid may contribute to greater motivation which may lead to higher test performance. Given that Blacks hold more negative face validity perceptions, face validity manipulations that improve Black perceptions may enhance their test performance (Sackett et al., 2001). But to improve face validity perceptions, one must first understand what factors contribute to face validity. This question of understanding what individual differences contribute to test perceptions has been proposed by Ryan (2001; Ryan & Ployhart, 2000), but to date has no answer. Clearly the appearance of the test is one important factor (Chan, Schmitt, Sacco, et al., 1998); we propose a second being individual differences in perceptions of stereotype threat.

THIS STUDY: INTEGRATING STEREOTYPE THREAT WITH APPLICANT REACTIONS

This study integrates the research on stereotype threat with the research on applicant reactions to propose and test hypotheses relevant to both literatures, with the ultimate goal of understanding what influences test performance and the resulting Black–White differences. We use a simulated selection context to manipulate both threat and face validity. Further, our hypotheses are framed around both experimental and correlational approaches, following their respective research traditions.

Experimental Hypotheses

Two manipulations are performed, one for stereotype threat and one for face validity. Stereotype threat is invoked via a diagnosticity manipulation. In the *diagnostic* condition, we use a manipulation similar to Steele and Aronson (1995) and describe the test as being diagnostic of cognitive ability. In addition, we include a

control condition where no such instructions are provided. However, we also include a third condition where the test is described as non-diagnostic of cognitive ability. In the *non-diagnostic* condition, the test is described as being diagnostic of retail manager skills (the fictitious job used in this study) rather than of cognitive ability. Steele et al. (2002) noted that this may be a reasonable way to reduce threat in real-world contexts. Research on applicant reactions suggests that describing why a test is face valid or job related enhances test-taker perceptions (e.g., Gilliland, 1993; Horvath, Ryan, & Stierwalt, 2000; Truxillo, Bauer, Campion, & Paronto, 2002). Other research suggests that merely describing the test as measuring more malleable skills, rather than relatively fixed intelligence, results in greater focus on effort and increased action in the face of setbacks (Hong, Chiu, Dweck, Lin, & Wan, 1999). Thus, one would expect test performance to be highest in the non-diagnostic condition, followed by the control condition and then the diagnostic condition.

We also manipulate the face validity of the test by altering the test's appearance and by changing the test name and item format. Specifically, in the *face valid* condition, the cognitive ability test and items are portrayed in a retail context and the test is named the Retail Management Skills Test, whereas in the *generic* (control) condition, the unaltered cognitive test is administered and is named the General Intelligence and Aptitude Test. This manipulation is based on recommendations from several researchers to couch ability items in job-related contexts (e.g., Anastasi & Urbina, 1997; Jensen, 2000; Rynes, 1993). Theoretically, the face valid version should result in higher performance than the generic version (e.g., Chan & Schmitt, 1997).

We expect these effects to be conditional on each other and on race, thus our first hypothesis proposes a Stereotype Threat × Face Validity × Race interaction on test performance. That is, it is not only how the test is described (diagnosticity), but also how it looks (face validity), that affects test performance. For Whites, we propose that the highest performance will occur when the face valid test is presented in the non-diagnostic condition (e.g., Chan & Schmitt, 1997), after which performance should be higher in the face valid versions than the generic versions across the control and diagnostic conditions (these latter two conditions should not differ from each other). For Blacks, we similarly propose that the highest performance will occur when the test is described as a retail manager test (non-diagnostic condition), followed by the control and then the diagnostic conditions. Across all diagnosticity conditions, Black test performance will be higher when the face valid test is used than when the generic test is used. Thus, the main difference between Blacks and Whites is that Whites should experience no difference between control and diagnostic conditions, whereas Blacks will experience a difference and score lower in the diagnostic condition. One implication of this hypothesis is that the smallest subgroup difference should occur when the face valid test is provided in the non-diagnostic condition.

Note that for testing this hypothesis, and consistent with stereotype threat theory, it is possible that these effects may be found only when participants are highly identified with the performance domain, their racial group, or both. Thus, we analyze Hypothesis 1 with and without controlling for domain and racial identity.

Correlational Hypotheses

Selection decisions are made based on individual test scores, making an understanding of how individual differences in test perceptions relate to individual differences in test performance important. For example, research on applicant reactions suggests that how individuals perceive selection procedures is often more important than how they are actually administered (Chan, Schmitt, Jennings, Clause, & Delbridge, 1998; Ployhart & Ryan, 1997; Ryan & Ployhart, 2000; Truxillo & Bauer, 1999). This research further shows how individual differences in perceptions contribute to test performance (e.g., Arvey et al., 1990; Chan et al., 1997). This finding may be particularly important with understanding stereotype threat in selection contexts because such contexts may be inherently threat provoking (Steele et al., 2002). Specifically, even if threat was ever-present in selection contexts, it does not mean that all participants will perceive threat in the same way, or that this threat would have the same consequences for all people.

Research on stigma consciousness supports this claim. In particular, Pinel's (1999, 2002) research found that individual differences in *stigma consciousness* (the extent to which people expect to be stereotyped by others) influence a variety of perceptions and behaviors, including the kinds of situations and activities people enter, perceptions of discrimination, and interpersonal behavior. Stigma consciousness is not the same as stereotype threat; stigma consciousness only represents the extent to which people believe they are being stereotyped, whereas stereotype threat carries an additional concern about one's behavior in the stereotyped context (Pinel, 1999). However, this distinction provides a rationale to expect individual differences in perceptions of stereotype threat (see also Steele et al., 2002). Consistent with the research on stigma consciousness and applicant reactions, one might expect that individual differences in perceived threat would relate to important applicant perceptions, reactions, and behavior.

Following this reasoning, we examine how individual differences in perceptions of stereotype threat relate to test performance through various mediating mechanisms. The specific mediators examined in this study were chosen because of their theoretical relationships to stereotype threat (see Steele et al., 2002), and because of their importance in the applicant reactions research discussed earlier. Figure 1 provides an overview of the theoretical model.

As can be seen in the figure, we first expect that race, racial identity, and domain identity will influence perceptions of stereotype threat. Blacks, those higher in ra-

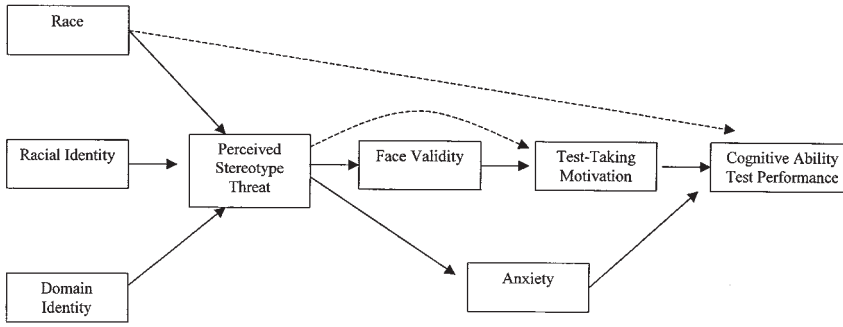

FIGURE 1 Hypothesized Model 1. Inclusion of dashed paths tests partial versus fully mediated models. Note that perceived stereotype threat is used in these analyses.

cial identity, and those higher in domain identity should exhibit higher levels of threat (e.g., Aronson et al., 1999; Steele & Aronson, 1995; Steele et al., 2002).

Second, perceived stereotype threat should influence the mediators of anxiety and face validity perceptions. Anxiety has long been offered as a theoretically important mediator of the threat effect (Steele & Aronson, 1995). Some research has documented this positive relationship (e.g., Osborne, 2001; Spencer et al., 1999), such that threat is associated with greater anxiety. Although this relationship has by no means been a consistent finding in the literature, prior research has tended to focus on manipulations of threat rather than on individual differences in perceived threat. We also expect perceived stereotype threat to negatively influence test-taking motivation, but drawing from the applicant reactions literature (see Chan et al., 1997), we expect this relationship to be indirect through face validity perceptions. There is no research we are aware of to support this threat-face validity relationship, but previous research has shown that Blacks hold more negative perceptions of face validity than Whites (Chan, 1997; Ryan, 2001). As we also expect racial differences on perceptions of threat, it seems reasonable that those perceiving higher threat will see the test as less face valid. Indeed, perceiving the test to be less face valid may be a way for those threatened to devalue the relevance of the testing domain (e.g., Steele & Aronson, 1995). Further, face validity perceptions should be positively related to test-taking motivation. This relationship has been found in numerous studies, including Arvey et al. (1990) and Chan et al. (1997).

Finally, we expect that test-taking motivation and anxiety will both jointly influence cognitive ability test performance. Evidence of positive test-taking motivation and negative anxiety relationships with test performance has been found in several studies, including Arvey et al. (1990), Chan et al. (1997), and Sanchez, Truxillo, and Bauer (2000; see Ryan, 2001 and Ryan & Ployhart, 2000, for a review).

Beyond this hypothesized model in Figure 1, we also examine alternative models that test different theoretical predictions. For example, we test whether the

model in Figure 1 accounts for the direct effects of race on test performance, or whether the effect of race on performance is only partially mediated. Similarly, we examine whether stereotype threat has direct effects on motivation, or whether such effects are fully mediated through face validity perceptions. Such model comparisons provide important tests of different theoretical predictions. However, given a lack of theory linking these two literatures, we are forced to cautiously explore these alternative models rather than hypothesize them a priori.

METHOD

Participants and Design

Participants were 394 students at a large Eastern university. Two hundred fifty were White and 144 were Black. There were approximately the same number of male ($n = 193$) and female ($n = 201$) participants who were young ($M = 18.90$, $SD = 1.91$). Their average combined SAT verbal and quantitative score was 1190.42 ($SD = 130.97$). These participants had applied for an average of 4.32 jobs ($SD = 3.67$) over the previous year.

A 3 (stereotype threat: diagnostic, control, non-diagnostic) × 2 (face validity: generic, control) × 2 (race: White, Black) between-subjects design was used. Participants were run in small groups consisting of both Whites and Blacks; no session was run with only Black participants because this would have been unusual and possibly alerted participants to the purpose of the study. All participants were randomly assigned to conditions.

Procedure

Participants were first told that the purpose of the study was to understand how to select people for retail managerial positions. We described what this job entailed, how common it was, and how there were few good tests that could be used to hire retail managers. Therefore, we explained that we wanted participants to take a test that could possibly be used for such a purpose. We used this job as there is no strong stereotype as to either Whites or Blacks performing better as retail managers, as both Blacks and Whites commonly hold these types of positions. Although it is possible that any type of management position may be a stereotype relevant context, it is unlikely that a strong stereotype exists for a retail management position due to the fact that there are many Black retail managers. We instructed participants to take the test like they would if they were an applicant applying for a retail manager position. To further enhance the simulated selection context, we used an approach followed in several studies (e.g., Chan et al., 1997; Horvath, Ryan, &

Stierwalt, 2000; McFarland & Ryan, 2000) and informed participants that those who scored in the top 15% would receive $20.

The diagnosticity manipulation involved verbal instructions explaining what the test measured and was presented immediately before they received the test. In the diagnostic condition (n = 130), we used a manipulation nearly identical to Steele and Aronson (1995). Specifically, participants were told:

> This test is designed to measure your intelligence, which means your general quantitative, verbal, and reasoning skills. Immediately after you complete the test, your test will be scored and you will be given feedback on your test performance (just like if you took the test for a job). This information may be helpful to you by familiarizing you with some of your strengths and weaknesses in quantitative and verbal problem solving. This test is a difficult but genuine test of your cognitive abilities and limitations, so that we might better understand the factors involved in both.

Notice that the focus of the diagnostic information is on general intelligence, which is the same as general cognitive ability. Conversely, in the non-diagnostic condition, there was no mention made of either cognitive ability or general intelligence. In particular, in the non-diagnostic condition (n = 136), participants were told:

> This test is designed to measure your skills as a retail manager, which means your ability to check inventory, develop store display layouts, create work schedules, and similar things that retail managers do. Immediately after you complete the test, your test will be scored and you will be given feedback on your test performance (just like if you took the test for a job). This information is given to you simply to get you familiarized with the content of these types of tests; because many of you will work in retail settings in college this information might be useful. This test is a difficult but genuine measure of your retail management skills, so that we might better understand the factors involved in them.

Notice that the focus of the non-diagnostic information is on retail manager skills, not on cognitive ability. Finally, in the control condition (n = 128), participants were merely told, "Please take the following test. This test is a difficult test."

After providing these instructions, participants were then administered the test. The face validity manipulation was provided by simply giving participants either the generic (n = 194) or retail-specific (n = 200) version of the test. Further, the generic or retail instructions from the cover page were read to participants, and they were then instructed to begin. All participants received 25 min to take the test, as pilot testing indicated that most people could complete the questions in about 20 min. Once the time limit was reached, the tests were collected and participants were then administered the posttest questionnaire containing the remaining measures described below (e.g., perceived stereotype threat, racial identity, motivation,

etc.). The items from each of these measures were randomized in the posttest questionnaire. After this questionnaire was finished, participants were debriefed as to the purpose of the study and thanked for their time.

Measures

Unless otherwise noted, all measures were assessed using a 5-point, *strongly disagree* to *strongly agree* format (higher numbers indicate more of the latent construct). Internal consistency reliability estimates are shown along the diagonal in Table 1.

Manipulation checks. Two manipulation checks were used to assess if participants were aware of the simulated selection context and that their scores on the test would influence their chance to earn $20. A sample manipulation check item is, "These test scores will be used to determine who gets the $20." The reliability of this scale was .88.

Cognitive ability tests. The cognitive ability test items were derived from sample items used in the GRE and GMAT. Eleven questions were related to analytical reasoning and 10 questions were related to quantitative reasoning. No verbal items were included because of the difficulty in adapting these to work versions. Thus, the test consisted of 21 questions. Items of moderate to high difficulty were chosen because the difficulty of the test is related to the magnitude of threat experienced (Spencer et al., 1999). These items were dichotomously scored (correct/in-

TABLE 1
Descriptive Statistics and Intercorrelations for All Measures[a]

	M	*SD*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Race	.37	.48	—							
Test score	7.78	3.33	–.50	.67						
Face validity perceptions	3.20	.93	.00	.06	.87					
Stereotype threat test specific	2.11	.81	.39	–.14	–.16	.81				
Motivation	3.30	.87	.03	.20	.26	–.23	.93			
Anxiety	2.81	.59	.18	–.37	–.16	.19	–.24	.75		
Racial identity	2.41	.73	.36	–.19	–.10	.70	–.12	.22	.80	
Identification with intelligence	2.94	.96	–.10	.19	.15	–.06	.53	–.08	.02	.76

Note. All values greater than .10 are significant at the .05 level and values greater than .13 are significant at the .01 level for a two-tailed test. Race is coded as 0 = White and 1 = Black. Reliabilities for the scales are reported along the diagonal.

[a] $n = 394$.

correct). The original 21 GRE and GMAT items were then altered to be specific to a management position in a retail setting. Following item cloning procedures outlined in Clause, Mullins, Nee, Pulakos, and Schmitt (1998), we kept the basic structure and purpose of each item intact but made the content appear to reflect retail settings. For example, a generic item:

> In a group of people solicited by a charity, 30% contributed $40 each, 45% contributed $20 each, and the rest contributed $12 each. What percentage of the total contributed came from people who gave $40?
>
> (A) 25%
> (B) 30%
> (C) 40%
> (D) 45%
> (E) 50%

was rewritten to be face valid for a retail setting as follows:

> Out of the total number of shirt sales in a retail store on a given day, 30% of the shirts sold for $40 each, 45% of the shirts sold for $20 each, and the rest of the shirts sold for $12 each. What percentage of the total sales came from shirts that sold for $40?
>
> (A) 25%
> (B) 30%
> (C) 40%
> (D) 45%
> (E) 50%

To ensure the adequacy of these items, they were pilot tested on 601 participants in a mass-testing session. In this pilot test, participants were either given the face valid or generic items. Participants were instructed to rate the items on 7-point scales ranging from *strongly disagree* to *strongly agree* as to the degree to which they measured intelligence and retail manager potential. Participants viewed the retail items as assessing more retail manager potential than the generic items (face valid $M = 4.11$, generic $M = 3.78$), $t(599) = 3.63$, $p < .001$. In the actual experimental sessions, the test booklet cover and instructions also differed between the face validity and control conditions. The generic test had a title page calling it the General Intelligence and Aptitude Test whereas the face valid version's cover page was labeled the Retail Management Skills Test. Further, the generic test had instructions calling it a test of general intellectual ability and the test publisher was named Psychological Testing Systems. The face valid test had instructions calling it a test of general managerial skills required for retail positions and the test publisher was named Managerial Assessment Systems.

Perceptions of stereotype threat. A self-report measure for detecting stereotype threat was constructed based on prior scales from Steele and Aronson (1995), Chatman (1999), and McKay (1999). After examining these existing items, it was apparent that different researchers were defining and measuring threat differently; some items referred to threat as specific to a particular test and others referred to threat as general to a construct. Thus, we aimed to identify a scale capable of measuring stereotype threat in employment testing contexts. Originally, 15 items were included; however, 7 items that were long and poorly worded exhibited poor factor loadings in an initial confirmatory factor analysis (CFA) and were eliminated. Thus, we focused on 8 items that were measured on a 5-point Likert scale.

Using the more refined items, we ran a series of nested CFA models for both Blacks and Whites and found that a two factor model for Blacks, $\chi^2(19) = 25.77$; standardized root square residual (SRMR) = .06; comparitive fit index (CFI) = .95; root mean square error of approximation (RMSEA) = .05; adjusted goodness of fit (AGFI) = .91, fit significantly better than a one factor model, $\chi^2(20) = 42.66$; SRMR = .08; CFI = .82; RMSEA = .09; AGFI = .86, as indicated by a change in chi-square test, $\Delta\chi^2(1) = 16.89$; $p < .05$. Further, the disattentuated correlation among these latent factors was only .51. Conversely, in the White sample we did not find a two factor solution, $\chi^2(19) = 21.55$; SRMR = .02; CFI = 1.00; RMSEA = .02; AGFI = .96, fit better than a one factor solution, $\chi^2(20) = 21.85$; SRMR = .02; CFI = 1.00; RMSEA = .02; AGFI = .96, as the change in chi-square from the one factor model to a two factor model was only 0.30 and the correlation of these latent factors for the two factor solution was 1.00.

Thus, in the Black sample there appeared to be two constructs; after considering the items, we labeled these constructs as *test-specific threat* (5 items) and *generalized threat* (3 items). An example of a test specific threat item is, "A negative opinion exists about how people from my race perform on this type of test." An example of a general threat item is, "Some people feel that I have less intelligence because of my race." Because the two-factor solution was necessary to adequately represent threat in the Black sample, and there was no difference between the one- and two-factor solutions in the White sample, we chose to focus on the test-specific threat measure for the remainder of this study. Of the two constructs, this one is more relevant to personnel selection contexts and is closer with the conceptualization of individual differences in perceived threat (as it is situationally bound).

Face validity. Face validity was measured with the 5-item Face Validity scale from Smither, Reilly, Millsap, Pearlman, and Stoffey (1993). This measure has been used considerably in previous research (see Ryan & Ployhart, 2000). A sample item is, "I could not see any relationship between the test and what is required on the job" (reverse scored).

Test-taking motivation. Test-taking motivation was assessed from Arvey et al.'s (1990) Test Attitude Survey (TAS). Subscales from the TAS, including those used in this study, have been used frequently (e.g., Chan et al., 1997; Schmit & Ryan, 1992). A sample item is, "I was extremely motivated to do well on this test."

Anxiety. Anxiety was measured with the 10-item comparative anxiety scale from Arvey et al. (1990). A sample item is, "During the test, I got so nervous that I couldn't do as well as I should have."

Racial identity. Racial identity was measured with a 7-item scale constructed with items from Helms (1990). A sample item is, "It is important that I am strongly affiliated with others from my race."

Domain identification. Individuals' degree of identification with intelligence testing was assessed with a 3-item scale. A sample item for identification is, "Scoring well on intelligence tests means a lot to me." These items were modeled after academic identification items presented in Steele and Aronson (1995).

RESULTS

Preliminary Analyses

Table 1 shows the descriptive statistics for all measures. Overall, the mean on the cognitive ability test was 7.78 (SD = 3.33), indicating that the test was difficult (participants on average got 37% of the questions correct). Again, it is important that the test is difficult (so long as it is not impossible) because more challenging tests lead to greater amounts of stereotype threat (Spencer et al., 1999). No participant got all of the questions right or wrong. The White mean on the test was 9.03 (SD = 3.14) and the Black mean was 5.61 (SD = 2.43). Overall, the mean difference (in SD units) between these groups was d = 1.18, a finding slightly higher, but largely consistent with the previous research on subgroup differences in cognitive ability tests (Jensen, 1998; Roth et al., 2001).

As mentioned earlier, two manipulation check items were administered after the test to ensure that participants were aware of the simulated selection context and that their test scores would influence their chance to earn $20. The mean of these two items was 3.78 and the median was 4, indicating that individuals were cognizant and aware of this selection testing context. Indeed, the mean of these items significantly differed from the midpoint of the scale, $t(393) = 13.27$, $p < .001$, suggesting that individuals were aware of the selection context.

To assess the discriminant validity of the stereotype threat and test perception scales, a confirmatory factor analysis was conducted. Two models were tested, a

one factor model and the hypothesized six factor model (i.e., perceptions of stereotype threat, racial identity, domain identity, face validity, anxiety, and test-taking motivation). Results found that the six factor model fit better, $\chi^2(725) = 2{,}220.36$, $p < .05$, SRMR = .093, CFI = .79, RMSEA = .072, AGFI = .71, than the one factor model, $\chi^2(740) = 5{,}006.56$, $p < .05$, SRMR = .15, CFI = .41, RMSEA = .12, AGFI = .40, as indexed by a change in chi square test, $\Delta\chi^2(15) = 2{,}786.20$, $p < .05$. Thus, the measures show reasonable discriminant validity. Note that although the fit is mixed for the six factor model, it is also based on a rather large model with 740 degrees of freedom. Fit indexes are affected by model complexity (Hu & Bentler, 1999); given the large degrees of freedom as well as the fact that we used previously established scales, we felt the model fit the data reasonably well.

Experimental Analyses

The first hypothesis predicted a Stereotype Threat × Validity interaction on test performance. Thus, a 3 (stereotype threat condition) × 2 (face validity) × 2 (race) analysis of variance (ANOVA) was conducted. Results indicate statistically significant main effects only for diagnosticity, $F(2, 382) = 3.45$, $p < .05$, and race, $F(1, 382) = 127.81$, $p < .001$. Table 2 shows the means for each condition, along with the standardized mean differences (*d*). Again, the significant main effect for race indicates that Whites ($M = 9.03$) scored higher than Blacks ($M = 5.61$) across all of the conditions, with standardized differences ranging from .91 to 1.64. Notice that the control-generic condition *d* of .94 maps similarly back on to previous meta-analytic estimates (e.g., Roth et al., 2001). Across the diagnosticity conditions, test performance was highest in the control condition ($M = 8.36$), followed by the non-diagnostic ($M = 7.81$) and then the diagnostic ($M = 7.18$) condition. Follow-up

TABLE 2
Race × Diagnosticity × Face Validity Cognitive Ability Test Score Means and Standard Deviations[a]

	Marginal Means		*Generic*					*Face Valid*				
			White		*Black*			*White*		*Black*		
Threat Condition	*M*	*SD*	*M*	*SD*	*M*	*SD*	*d*	*M*	*SD*	*M*	*SD*	*d*
Control	8.36	3.47	9.50	3.29	6.59	2.67	.94	9.88	3.17	5.26	2.03	1.64
Diagnostic	7.18*	3.21	8.65	3.11	5.77	2.07	1.05	8.31	3.14	4.60	2.24	1.31
Non-diagnostic	7.81	3.25	8.84	2.97	5.17	2.50	1.30	9.00	3.09	6.32	2.69	.91
Marginal *M*					7.85	3.27				7.72	3.41	

Note. Marginal means represent the weighted means for the three stereotype threat conditions and the two face validity conditions across race.

[a] $n = 394$.

*Mean difference significant ($p < .05$) from control condition.

planned comparisons found that only the control and diagnosticity conditions differed significantly, $p < .05$. Thus, although diagnosticity appeared to negatively impact test performance as previous research has found, it did not affect one racial group more than another.

Overall, Hypothesis 1 was not supported. Although it is possible that an ineffective threat manipulation could explain the lack of support for this hypothesis, the manipulation used was modeled after Steele and Aronson (1995), so this possibility is minimized to some degree. It is also possible that random assignment did not work in this instance, or that participants did not equally care about the retail management position. We conducted analyses to examine these alternative explanations by covarying total SAT scores and a measure of identification with the retail management position (measured only as a potential covariate); these did not change any of the conclusions noted earlier and are not discussed further. Although Hypothesis 1 was not supported, notice that the subgroup difference was exaggerated in the threat conditions, particularly when a face valid test was used or the test was described as non-diagnostic. Yet in contrast, the smallest subgroup difference occurred when the test was face valid and described as diagnostic of managerial skills. This may indicate that for subgroup differences to be reduced, the test must not only look face valid, but also be explained as such.

Prior research has illustrated that identification with the domain being tested is an important factor in the strength of the stereotype threat effect (Aronson et al., 1999), such that threat effects exist most strongly for individuals who are most highly identified with the performance domain. We therefore reran the analysis mentioned earlier after controlling for individuals' identification with their race, their identification with intelligence testing, and with both covariates simultaneously. The results in each of these analyses were identical to those described in the previous paragraphs in that none of these new analyses reached significance.

This general lack of support is in contrast to the published work on stereotype threat. One potential reason for this discrepancy is due to the selection of the participants in our sample versus other studies. Indeed, many prior studies on stereotype only select participants who are highly identified with the domain (Steele et al., 2002). Therefore, we ran post hoc analyses in which we repeated the three-way interaction mentioned earlier after selecting only individuals who were highly identified with the intelligence testing domain. Specifically, we only selected individuals who scored in top 50% on the identification with intelligence testing scale (White: $n = 115$; Black: $n = 74$). This analysis did not find a significant three-way interaction. Although this lack of statistical significance may be partly due to a reduced sample size resulting in a reduction of power, the effect sizes were simply not large enough to indicate a sizeable effect was present.

Next, we re-ran the same $3 \times 2 \times 2$ ANOVA, but only analyzed participants who were highly identified with their racial group; we operationalized this as participants scoring in the top 50% on the racial identity scale (White: $n = 116$; Black: $n =$

72). For this restricted sample, the three-way Stereotype Threat × Face Validity × Race interaction was significant, $F(2, 176) = 3.57$, $p < .05$; the cell means are shown in Table 3 and graphed in Figure 2. Interestingly, the face validity manipulation seems to change the effects of the stereotype threat manipulation. When the generic test is administered, Blacks score best in the control condition, with the difference between the control and diagnostic conditions being significant; $t(21) = 2.16$, $p < .05$. The opposite is found when the face valid test is administered, such that Blacks score best in the non-diagnostic condition, with the difference between the non-diagnostic and diagnostic conditions being significant; $t(22) = 1.96$, $p < .05$. There are no significant White mean differences across any of these conditions. Particularly interesting is the finding that Black performance on the face valid test in the control condition is nearly the same as Black performance on the face valid test in the diagnostic condition. Apparently, lacking a description of what the test measures, Blacks assume the face valid measure taps cognitive ability.

Correlational Analyses

The model shown in Figure 1 was tested using path analysis in LISREL 8.30 with maximum likelihood estimation. Please recall that the stereotype threat variable used in these analyses is the continuous perceived stereotype threat measure (i.e., the self-report measure, not the manipulated variable). To examine model fit, we focus on the chi-square and change in chi-square when comparing models, SRMR, CFI, RMSEA, and AGFI. Following common suggestions (Browne & Cudeck,

TABLE 3
Pre-Selected Analyses: Race × Diagnosticity × Face Validity Cognitive Ability Test Score Means and Standard Deviations[a]

	Marginal Means		*Generic*					*Face Valid*				
			White		*Black*			*White*		*Black*		
Threat Condition	*M*	*SD*	*M*	*SD*	*M*	*SD*	*d*	*M*	*SD*	*M*	*SD*	*d*
Control	7.69	3.27	7.69	3.61	7.25[b,e]	3.11	.13	9.43[d]	2.69	5.00[d,e]	1.91	1.82
Diagnostic	6.64	3.43	8.14[c]	3.23	4.91[b,c]	1.87	1.24	8.31[d]	3.57	4.25[b,d]	2.56	1.28
Non-diagnostic	8.00	3.27	9.30[c]	2.87	5.67[c]	2.57	1.31	8.50	3.32	6.58[b]	3.23	.58
Marginal *M*					7.58	3.26				7.45	3.44	

Note. Participants in this table were pre-selected for being highly identified with their racial group, such that only those in the top 50% on the racial identity scale are included. Marginal means represent the weighted means for the three stereotype threat conditions and the two face validity conditions across race. Within each column, means with the same subscript are significantly different. Within each row, means with the same subscript are significantly different. All significance tests use $p < .05$.

[a]$n = 188$.

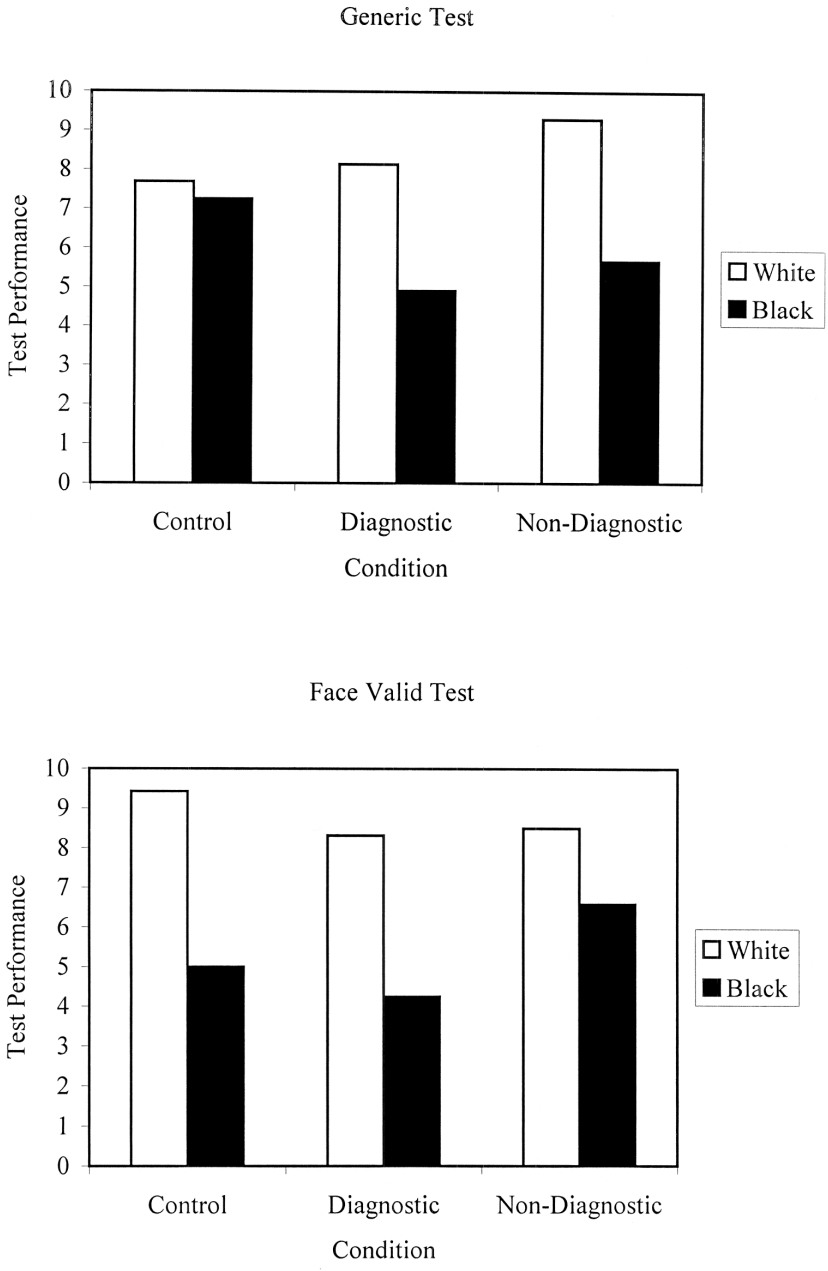

FIGURE 2 Three-way interaction between stereotype threat, face validity, and race (using only those scoring in the top 50% on racial identity).

1993; Hu & Bentler, 1999), good fit is indicated by (approximately) SRMR ≤ .08, CFI ≥ .90, RMSEA ≤ .10, and AGFI ≥ .90. No corrections for measurement error were made.

Table 4 shows the fit indexes for these models. In examining the hypothesized Model 1 (Figure 1), the fit indexes clearly show a poor fit to the data. The most prominent cause of misfit was solved by adding a path from domain identity to test-taking motivation (while dropping the path to stereotype threat). Apparently, domain identification has no direct impact on individual differences in perceived stereotype threat, but rather has a direct effect on test-taking motivation. In hindsight, this makes sense as we are all familiar with individuals who try harder on tasks they find more important.

Therefore, we revised Figure 1 and ran Model 2 with the direct path from domain identity to motivation. As shown in Table 4, although this change results in a substantially better fit, it still did not approach acceptable standards. This is not surprising, as we had specified the model such that all of the effects of race on ability test performance were mediated through stereotype threat and test perceptions. By allowing a direct path from race to test performance (Model 3), the data now approached a moderate fit (the difference between Models 2 and 3 was also statistically significant). Thus, individual differences in stereotype threat and test perceptions partially mediate the effects of race on test performance.

Because no prior research has examined individual differences in perceptions of stereotype threat, much less integrated stereotype threat and applicant reactions research, we sought to further test the nomological network surrounding perceptions of threat and test perceptions. We therefore tested two additional theoretical

TABLE 4
Fit Indexes for Path Models[a]

Model	χ^2	*df*	$\Delta\chi^2$	Δdf	*SRMR*	*CFI*	*RMSEA*	*AGFI*
Hypothesized model (Figure 1)	323.62*	17	—	—	.14	.59	.19	.69
Revised model 2 (domain identity → motivation)	205.54*	17	—	—	.11	.75	.16	.78
Revised model 3 (race → test performance)	98.21*	16	107.33*	1	.078	.89	.11	.88
Revised model 4 (stereotype threat → motivation)	80.74*	15	17.74*	1	.070	.91	.10	.89
Revised model 5 (motivation → anxiety)	63.79*	14	16.95*	1	.057	.93	.092	.90

Note. SRMR = standardized root square residual; CFI = comparitive fit index; RMSEA = root mean square error of approximation; AGFI = adjusted goodness-of-fit index.

[a]$n = 394$.

*$p < .05$.

models. First, we allowed a direct path from stereotype threat to motivation (Model 4), as many have suggested that effort (motivation) may be a mediator of stereotype threat (Aronson et al., 1999; Steele & Aronson, 1995; Steele et al., 2002). Second, we allowed a direct path from motivation to anxiety (Model 5), as research suggests that goals (a motivational end-state) may contribute to test anxiety (e.g., Elliot & McGregor, 1999). Likewise, research on stress and coping suggests that those who exhibit more active forms of coping experience less stress (Carver, Scheier, & Weintraub, 1989; Folkman, Lazarus, Gruen, & DeLongis, 1986). Table 4 shows the fit of these revised models. As shown in the table, both revisions improved model fit, with Model 4 fitting significantly better than Model 3, and Model 5 fitting significantly better than Model 4. We thus consider Model 5 to be an adequate representation of the data.

Figure 3 shows the revised Model 5 graphically with standardized path coefficients (all are significant at $p < .05$). This model indicates several new and theoretically interesting relationships among race, perceived threat, test perceptions, and test performance. Starting from the left of the figure, individual differences in stereotype threat are largely driven by racial identity. Consistent with previous research, stereotype threat effects are stronger for those more identified with their race. Next, although Ryan (2001; Ryan & Ployhart, 2000) questioned what individual differences drive test perceptions, Steele et al. (2002) questioned what constructs mediate the effects of stereotype threat. The middle portion of Figure 2 helps answer both of these questions, as all test perceptions are affected by individual differences in stereotype threat. The greater one's perceptions of threat, the less face valid the test appears, the less motivated one is, and the more anxious one

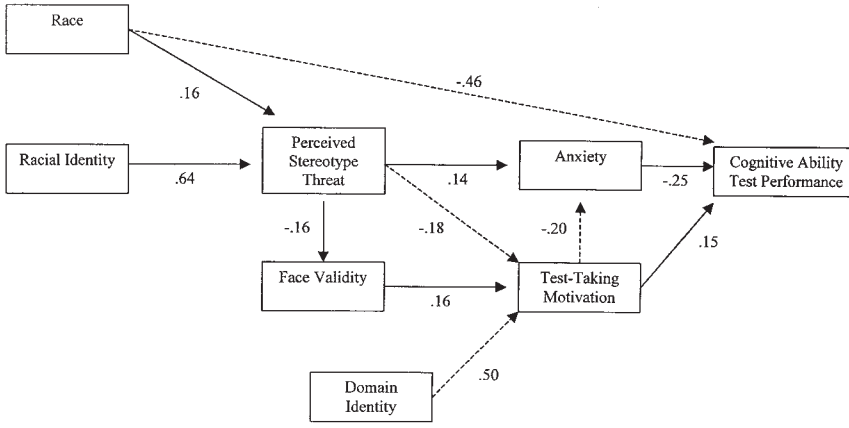

FIGURE 3 Final Revised Model 5. Path coefficients are completely standardized estimates. All paths in dashes are those not originally hypothesized. Note that perceived stereotype threat is used in these analyses.

feels. These relationships suggest that individual differences in threat have a fundamental impact on test perceptions. Interestingly, and contrary to stereotype threat research and theory, domain identity operates primarily to influence more specific test-taking motivation rather than indirectly through stereotype threat. Finally, beyond race, anxiety negatively impacts test performance whereas motivation improves it. Together, these construct relations help illuminate the processes through which race relates to test performance.

DISCUSSION

The purpose of this study was to integrate research on stereotype threat with research on applicant reactions to better understand racial subgroup differences on cognitive ability tests in employment contexts. The experimental analyses suggested that, when preselecting those more strongly identified with their race, stereotype threat lowered Black test performance. However, the nature of this effect for the other two threat conditions was conditional on the face validity of the test, such that the generic test produced higher performance in the control condition, whereas the face valid test produced higher performance in the non-diagnostic condition. The correlational analyses suggested that individual differences in perceptions of threat contribute to face validity perceptions, motivation, and anxiety, and it is through these mechanisms that stereotype threat influences test performance. Together, these analyses provide several theoretical implications.

First, these results contribute to an understanding of racial subgroup differences in cognitive ability test performance. In particular, the results suggest that cognitive ability test performance is affected by both distal (perceptions of stereotype threat) and proximal (face validity, test-taking motivation, anxiety) influences. It should come as no surprise that explanations for racial subgroup differences are complex and multi-determined, as many have argued that research must identify the psychological reasons why race is related to test performance (Helms, 1992; Jensen, 1998). Although this study does not find a fully mediated relationship between race and test performance, the results do support a partially mediated model. As such, Figure 3 provides a starting point for future research to identify other potential psychological mediators.

Second, this study contributes to the literature on stereotype threat in at least two ways: (a) by examining threat in selection contexts and (b) by studying individual differences in perceptions of threat. Interestingly, we found no interaction between race and stereotype threat. This lack of support suggests that either threat was present for Blacks in all of the conditions examined in this study, or that Whites were also affected by the threat manipulation. Although we cannot rule out the first explanation, our data are more consistent with the second rationale as White test performance tended to be lowest in the diagnostic condition. In contrast

to much previous research (e.g., Aronson et al., 1999; Leyens et al., 2000), this study did not explicitly tell participants that Whites perform better on the test than Blacks. Regardless of whether threat affects both Whites and Blacks, or whether threat is a constant for Blacks in most employment testing contexts, one implication of this study is that by itself, reducing stereotype threat would not lead to practically important reductions in subgroup test differences.

However, more consistent with previous stereotype threat research, when we selected only those individuals who were highly identified with their race, the diagnostic condition always produced the lowest Black test performance. Interestingly, the nature of the threat effect differed by face validity condition for the other two threat conditions: when the generic test was administered, Black performance was highest in the control condition; when the face valid test was administered, Black performance was highest in the non-diagnostic condition. From these results, it appears that Blacks taking the generic test still felt threat even when told it measured retail manager skills, whereas Blacks taking the face valid test perceived it as a measure of intelligence unless they were told otherwise.

Also of importance in this study was the preliminary support found for the key role played by perceptions of threat. Previous research has questioned whether stereotype threat could account for racial differences in real selection contexts because such contexts would be inherently threat provoking (Steele et al., 2002). Our study suggests that although this possibility may be true, it is more important how individuals perceive threat. Similar to the work of Pinel (1999) on stigma consciousness, our results suggest that there are important individual differences in how individuals exposed to the same stimuli perceive and react to those stimuli. The importance of such individual differences has been a common finding in the applicant reactions literature (e.g., Arvey et al., 1990; Chan et al., 1997; Ployhart & Ryan, 1997), but to date has been relatively unexamined in the stereotype threat literature. Further, we suspect that one of the reasons research has not found consistent mediators of stereotype threat is because it has attempted to find individual difference mediators of between-condition threat manipulations. As such, within-condition variance may largely reflect individual differences in perceptions of threat, and to the extent that it is this within-condition variance that relates to other individual difference constructs, examining perceptions of threat may help unlock the psychological mediators of stereotype threat.

Another interesting result of this study is the apparently pivotal role played by racial identity. We found few effects of domain identity (intelligence) in this study; however, we found the Stereotype Threat × Face Validity × Race interaction with individuals who were highly identified with their race. The path analyses also suggested that racial identity is the primary driver of perceptions of stereotype threat, as opposed to domain identity. Thus, our results suggest both racial and domain identity are important, but they differ in how they are important and what constructs they impact. This impact of racial identity is in line with the theoretical as-

pects of stereotype threat. In particular, as threat is thought to develop from an individual being concerned about confirming a group's stereotype, it stems that one who highly identifies with the group would experience the most threat. This impact of group identity on stereotype threat has also been illustrated in recent research by Schmader (2002). In particular, when experiencing high stereotype threat, women who were highly identified with their gender performed worse than women who were not highly identified with their gender. Thus, being highly identified with the stereotyped group appears to highlight the negative aspects of stereotype threat.

Third, these findings contribute to the literature on applicant reactions. Most of this research has focused on proximal, test-specific perceptions (test-taking motivation, test anxiety). Although much research suggests a small, but important, link between these test-specific perceptions and test performance (Chan et al., 1997; Ryan, 2001; Sackett et al., 2001), the factors that contribute to test perceptions are less well understood. In particular, the reasons why subgroup differences exist in test perceptions remains largely unknown (Ryan, 2001). Our study suggests that these more proximal test perceptions may be driven by larger issues surrounding racial stereotypes, racial identity, and perceptions of threat. Such fundamental and enduring views may influence a host of processes and perceptions for those afflicted by them. For example, it may be that every testing context would evoke stereotype threat, even when the test is noncognitive in nature. Likewise, it is possible that stereotype threat could also influence on the job behavior and work performance (cf. Pinel, 1999, 2002).

An additional contribution to research on applicant reactions is that we manipulated face validity by alternating the appearance of the test. Interesting, this suggestion has been offered on numerous occasions (e.g., Anastasi & Urbina, 1997; Rynes, 1993), but few studies have actually examined its effects on cognitive ability test performance and applicant perceptions. We found that the face validity manipulation interacted with the stereotype threat manipulation in different ways for Whites and Blacks. Yet by itself, the face validity manipulation did not have an effect on racial subgroup differences or test performance. However, this lack of a direct effect does not mean it is unimportant. Jensen (2000) noted that test developers must begin to adopt such practices if cognitive ability tests are going to gain widespread acceptance from the general public. There are many positive benefits that may accrue from using tests that are face valid. For example, such tests may reduce the likelihood of litigation against the organization, enhance perceptions of test fairness and appropriate treatment, and not adversely affect one's intention to purchase products from the organization (e.g., Smither et al., 1993). Given the relatively low cost for adapting a traditional cognitive ability test into one that appears more job related, there are few good reasons not to make this a standard practice for test development.

Fourth, these results highlight the utility of taking a broader perspective when trying to understand racial subgroup test differences. We see relatively few references in the I–O literature to the work in social psychology, and even fewer references in the social psychological literature to the work in I–O psychology. This is unfortunate because both domains ultimately seek to understand the same phenomena, and isolation in such instances leads to duplication of effort and resources. In this study, we found that each approach had implications for the other, and the best understanding of racial subgroup differences came from the joint consideration of both perspectives. As shown in Table 2, the smallest subgroup difference occurred in the non-diagnostic, face valid condition. Although the positive benefit of this may be small (the subgroup difference was only reduced .03 *SD* units from the generic control condition), it is the negative consequences of not paying attention to threat or face validity that are most telling. For example, using a face valid test without a description of what the test measures had a dramatic negative effect on Black test performance; $d = 1.31$.

Finally, these results may have some important implications for practice. Most important is that test administers should be sure not only to use face valid tests, but also to explain why the test is face valid. As shown in Table 2, Blacks appeared to respond to the face valid test in the control condition much the same as they did to the face valid test in the diagnostic condition, suggesting that unless test administrators explain what the test measures, Blacks may interpret it as a measure of cognitive ability. And yet simply providing explanations for how the test relates to the knowledge, skills, and ability (KSA) required on the job also slightly lowered test performance for Blacks. These analyses would suggest that practitioners may do more harm than good by neglecting either the appearance or explanations surrounding the use of the test. Further, it will also be important for practitioners not to treat all Blacks as holding the same perceptions. For example, only those most highly identified with their racial group may hold negative perceptions. Likewise, we found important individual differences in perceptions of stereotype threat.

These results and implications leave us with the question that initially motivated this research, "Does stereotype threat account for racial subgroup differences in cognitive ability tests in selection contexts?" Consistent with the reasoning of Sackett et al. (2001), our results suggest the answer is, "By itself, no." In hindsight this question may have been the wrong one to ask; a better question would explore whether stereotype threat contributes to an understanding of these racial differences, and to this question, our study would clearly say "Yes." But so too does an understanding of test-taker perceptions and reactions. Thus, stereotype threat by itself will not fully account for racial subgroup differences (as few would ever claim), but an understanding of how threat operates through various mediating processes may contribute to a greater understanding of the psychology of the test-taker.

Limitations and Directions for Future Research

To our knowledge, this is one of the first studies to formally integrate research on applicant reactions with research on stereotype threat. Although we tried to be true to the methods used in both literatures, there are still limitations of this study that future research should seek to address. First, we did not have a non-selection context condition, making it difficult to directly compare our findings back to those in the social psychological literature. Although our primary interest was only in testing contexts, it would be helpful for future research to tease apart the effects of stereotype threat in these two contexts (selection vs. non-selection). Such a study would help determine how the effect size of threat changes by context.

Second, we measured all of our test perceptions after test performance. This was necessary because asking such questions before the test would be likely to prime the negative stereotype before test administration; something we obviously did not wish to do. Likewise, some test perceptions such as anxiety are best assessed after one has experienced the test. On the other hand, it is possible that participants responded to the posttest survey in a manner that reflected their perceived test performance (e.g., "I think I performed poorly, so I don't think this test is face valid"; Chan, Schmitt, Sacco, et al., 1998). Readers should also not overly interpret our correlational analyses as proving causality, and future experimental research will ultimately be necessary to make stronger causal inferences.

Third, the generalizability of these findings may be questioned on the grounds that (a) participants were all students who (b) were engaged in a simulated selection procedure. Our results may be slightly restricted in that participants had moderate SAT total scores. Future studies may use other variations of our design to examine the effects on real applicants, although manipulating threat in the real world would clearly be inappropriate. However, using the self-report measure of perceived threat would allow an examination of stereotype threat in real selection contexts. Of course, future research should continue to assess the validity of this measure and our other measures, as some of these (e.g., racial identity) were based on composites from previous research. Future research should also cross-validate the path model shown in Figure 3, as the final model was arrived at through several iterations.

Finally, the manipulations of stereotype threat and face validity may have affected each other, although we found no evidence that this affected our results in any substantive manner. A related concern is that we used a very simple manipulation of face validity. One might find stronger effects if the test format was altered from a paper-and-pencil test to one using a higher fidelity format (e.g., video), as was found in Chan and Schmitt (1997).

CONCLUSION

This study finds that the stereotype threat and applicant reactions literatures are quite complementary. Together, they increase our understanding of social and per-

sonal factors that contribute to cognitive ability test performance and Black–subgroup differences. Future research should further explore how these diverse research traditions may fit together to reduce subgroup test differences. Such benefits may extend well beyond any particular study, applicant, or organization; they extend to the hope of a society that promotes equality and equal access for all.

ACKNOWLEDGMENT

We thank Elyse Wallach and Elyssa Ivers for their help with data collection.

REFERENCES

Anastasi, A., & Urbina, S. (1997). *Psychological testing*. Upper Saddle River, NJ: Prentice-Hall.

Aronson, J., Lustina, M., Good, C., Keough, K., Steele, C., & 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.

Bobko, P., Roth, P. L., & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology, 52,* 561–589.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Carver, C. S., Scheier, M. F., & Weintraub, J. W. (1989). Assessing coping strategies: A theoretically based approach. *Journal of Personality and Social Psychology, 56,* 267–283.

Chan, D. (1997). Racial subgroup differences in predictive validity perceptions on personality and cognitive ability tests. *Journal of Applied Psychology, 82,* 311–320.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82,* 143–159.

Chan, D., Schmitt, N., DeShon, R. P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology, 82,* 300–310.

Chan, D., Schmitt, N., Jennings, D., Clause, C. S., & Delbridge, K. (1998). Applicant perceptions of test fairness integrating justice and self-serving bias perspectives. *International Journal of Selection & Assessment, 6*, 232–239.

Chan, D., Schmitt, N., Sacco, J. M., & DeShon, R. P. (1998). Understanding pretest and posttest reactions to cognitive ability and personality tests: Performance-reactions relationships and their structural invariance across racial groups. *Journal of Applied Psychology, 83,* 471–485.

Chatman, C. M. (1999). *The identity paradox model: Explaining school performance among African–American college students.* Unpublished doctoral dissertation, Rutgers University, New Brunswick, NJ.

Clause, C. S., Mullins, M. D., Nee, M. T., Pulakos, E., & Schmitt, N. (1998). Parallel test form development: A procedure for alternative predictors and an example. *Personnel Psychology, 51,* 193–208.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56,* 5–18.

Elliot, A. J., & McGregor, H. A. (1999). Test anxiety and the hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology, 76*, 628–644.

Folkman, S., Lazarus, R. S., Gruen, R. J., & DeLongis, A. (1986). Appraisal, coping, health status, and psychological symptoms. *Journal of Personality and Social Psychology, 50*, 571–579.

Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *Academy of Management Review, 18,* 694–734.

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.

Helms, J. E. (1990). *Black and White racial identity: Theory, research, and practice.* New York: Greenwood.

Helms, J. E. (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing? *American Psychologist, 47,* 1083–1101.

Hong, Y. Y., Chiu, C. Y., Dweck, C. S., Lin, D. M. S., & Wan, W. (1999). Implicit theories, attributions, and coping: A meaning system approach. *Journal of Personality and Social Psychology, 77,* 588–599.

Horvath, M., Ryan, A. M., & Stierwalt, S. L. (2000). The influence of explanations for selection test use, outcome favorability, and self-efficacy on test-taker perceptions. *Organizational Behavior and Human Decision Processes, 83,* 310–330.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001).Determinants, detection, and amelioration of adverse impact in personnel selection procedures: Issues, evidence, and lessons learned. *Internal Journal of Selection and Assessment, 9,* 152–194.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72–98.

Jensen, A. R. (1998). *The g factor.* Westport, CT: Praeger.

Jensen, A. R. (2000). Testing: The dilemma of group differences. *Psychology, Public Policy, and Law, 6,* 121–127.

Leyens, J. P., Desert, M., Croizet, J. C., & Darcis, C. (2000). Stereotype threat: Are lower status and history of stigmatization preconditions of stereotype threat? *Personality and Social Psychology Bulletin, 26,* 1189–1199.

Major, H. R., Spencer, S. J., Schmader, T., Wolfe, C. T., & Crocker, J. (1998). Coping with negative stereotypes about intellectual performance: The role of psychological disengagement. *Personality and Social Psychology Bulletin, 24,* 34–50.

McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across non-cognitive measures.*Journal of Applied Psychology, 85*, 812–821.

McKay, P. F. (1999). *Stereotype threat and its effect on the cognitive ability test performance of African–Americans: The development of a theoretical model.* Unpublished doctoral dissertation, University of Akron, Akron, Ohio.

Osborne, J. W. (2001). Testing stereotype threat: Does anxiety explain race and sex differences in achievement? *Contemporary Educational Psychology, 26,* 291–310.

Pinel, E. C. (1999). Stigma consciousness: The psychological legacy of social stereotypes. *Journal of Personality and Social Psychology, 76,* 114–128.

Pinel, E. C. (2002). Stigma consciousness in intergroup contexts: The power of conviction. *Journal of Experimental Social Psychology, 38,* 178–185.

Ployhart, R. E., & Ryan, A. M. (1997). Toward an explanation of applicant reactions: An examination of organizational justice and attribution frameworks. *Organizational Behavior and Human Decision Processes, 72*, 308–335.

Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., III., & Tyler, P. (2001). Ethnic subgroup differences in cognitive ability in employment and educational settings. A meta-analysis. *Personnel Psychology, 54*, 297–330.

Ryan, A. M. (2001). Explaining the Black/White test score gap: The role of test perceptions. *Human Performance, 14,* 45–75.

Ryan, A. M., & Ployhart, R. E. (2000). Applicants' perceptions of selection procedures and decisions: A critical review and agenda for the future. *Journal of Management, 26,* 565–606.

Rynes, S. L. (1993). Who's selecting whom? Effects of selection practices on applicant attitudes and behavior. In W. C. Borman & N. Schmitt (Eds.), *Personnel selection in organizations* (pp. 240–274). San Francisco: Jossey-Bass.

Sackett, P. R., & Roth, L. (1996). Multi-stage selection strategies: A monte carlo investigation of effects on performance and minority hiring. *Personnel Psychology, 49,* 1–18.

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

Sackett, P. R., & Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist, 49,* 929–954.

Sanchez, R. J., Truxillo, D. M., & Bauer, T. N. (2000). Development and examination of an expectancy-based measure of test-taking motivation. *Journal of Applied Psychology, 85,* 739–750.

Schmader, T. (2002). Gender identification moderates stereotype threat effects on women's math performance. *Journal of Experimental Social Psychology, 38,* 194–201.

Schmidt, F. L. (1988). The problem of group differences in ability test scores in employment selection. *Journal of Vocational Behavior, 33,* 272–292.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262–274.

Schmit, M. J., & Ryan, A. M. (1992). Test-taking dispositions: A missing link? *Journal of Applied Psychology, 77,* 629–637.

Schmitt, N., & Chan, D. (1998). *Personnel selection: A theoretical approach.* Thousand Oaks, CA: Sage.

Schmitt, N., & Chan, D. (1999). The status of research on applicant reactions to selection tests and its implications for managers. *International Journal of Management Reviews, 1,* 45–62.

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

Smither, J. W., Reilly, R. R., Millsap, R. E., Pearlman, K., & Stoffey, R. W. (1993). Applicant reactions to selection procedures. *Personnel Psychology, 46,* 49–76.

Spearman, C. (1927). *The abilities of man: Their nature and measurement.* New York: Macmillan.

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.

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). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69,* 797–811.

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

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.

Truxillo, D. M., & Bauer, T. N. (1999). Applicant reactions to test score banding in entry-level and promotional contexts. *Journal of Applied Psychology, 84,* 322–339.

Truxillo, D. M., Bauer, T. N., Campion, M. A., & Paronto, M. E. (2002). Selection fairness information and applicant reactions: A longitudinal field study. *Journal of Applied Psychology*, *87*, 1020–1031.