

Stereotype Threat and the Intellectual Test Performance of African Americans

Claude M. Steele
Stanford University

Joshua Aronson
University of Texas, Austin

Stereotype threat is being at risk of confirming, as self-characteristic, a negative stereotype about one's group. Studies 1 and 2 varied the stereotype vulnerability of Black participants taking a difficult verbal test by varying whether or not their performance was ostensibly diagnostic of ability, and thus, whether or not they were at risk of fulfilling the racial stereotype about their intellectual ability. Reflecting the pressure of this vulnerability, Blacks underperformed in relation to Whites in the ability-diagnostic condition but not in the nondiagnostic condition (with Scholastic Aptitude Tests controlled). Study 3 validated that ability-diagnosticity cognitively activated the racial stereotype in these participants and motivated them not to conform to it, or to be judged by it. Study 4 showed that mere salience of the stereotype could impair Blacks' performance even when the test was not ability diagnostic. The role of stereotype vulnerability in the standardized test performance of ability-stigmatized groups is discussed.

Not long ago, in explaining his career-long preoccupation with the American Jewish experience, the novelist Philip Roth said that it was not Jewish culture or religion per se that fascinated him, it was what he called the Jewish "predicament." This is an apt term for the perspective taken in the present research. It focuses on a social-psychological predicament that can arise from widely-known negative stereotypes about one's group. It is this: the existence of such a stereotype means that anything one does or any of one's features that conform to it make the stereotype more plausible as a self-characterization in the eyes of others, and perhaps even in one's own eyes. We call this predicament *stereotype threat* and argue that it is experienced, essentially, as a self-evaluative threat. In form, it is a predicament that can beset the members of any group about whom negative stereotypes exist. Consider the stereotypes elicited by the terms *yuppie*, *feminist*, *liberal*, or *White male*. Their prevalence in society raises the possibility for potential targets that the stereotype is true of them and, also, that other people will see them that way. When the allegations of the stereotype are importantly

negative, this predicament may be self-threatening enough to have disruptive effects of its own.

The present research examined the role these processes play in the intellectual test performance of African Americans. Our reasoning is this: whenever African American students perform an explicitly scholastic or intellectual task, they face the threat of confirming or being judged by a negative societal stereotype—a suspicion—about their group's intellectual ability and competence. This threat is not borne by people not stereotyped in this way. And the self-threat it causes—through a variety of mechanisms—may interfere with the intellectual functioning of these students, particularly during standardized tests. This is the principal hypothesis examined in the present research. But as this threat persists over time, it may have the further effect of pressuring these students to protectively disidentify with achievement in school and related intellectual domains. That is, it may pressure the person to define or redefine the self-concept such that school achievement is neither a basis of self-evaluation nor a personal identity. This protects the person against the self-evaluative threat posed by the stereotypes but may have the byproduct of diminishing interest, motivation, and, ultimately, achievement in the domain (Steele, 1992).

The anxiety of knowing that one is a potential target of prejudice and stereotypes has been much discussed: in classic social science (e.g., Allport, 1954; Goffman, 1963), popular books (e.g., Carter, 1991) and essays, as, for example, S. Steele's (1990) treatment of what he called *racial vulnerability*. In this last analysis, S. Steele made a connection between this experience and the school life of African Americans that has similarities to our own. He argued that after a lifetime of exposure to society's negative images of their ability, these students are likely to internalize an "inferiority anxiety"—a state that can be

Claude M. Steele, Department of Psychology, Stanford University; Joshua Aronson, School of Education, University of Texas, Austin. This research was supported by National Institutes of Health Grant MH51977, Russell Sage Foundation Grant 879.304, and by Spencer Foundation and James S. McDonnell Foundation postdoctoral fellowships, and its completion was aided by the Center for Advanced Study in the Behavioral Sciences.

We thank John Butner, Emmeline Chen, and Matthew McGlone for assistance and helpful comments on this research.

Correspondence concerning this article should be addressed to Claude M. Steele, Department of Psychology, Stanford University, Stanford, California 94305, or Joshua Aronson, School of Education, University of Texas, Austin, Texas 78712.

aroused by a variety of race-related cues in the environment. This anxiety, in turn, can lead them to blame others for their troubles (for example, White racism), to underutilize available opportunities, and to generally form a victim's identity. These adaptations, in turn, the argument goes, translate into poor life success.

The present theory and research do not focus on the internalization of inferiority images or their consequences. Instead they focus on the immediate situational threat that derives from the broad dissemination of negative stereotypes about one's group—the threat of possibly being judged and treated stereotypically, or of possibly self-fulfilling such a stereotype. This threat can befall anyone with a group identity about which some negative stereotype exists, and for the person to be threatened in this way, he need not even believe the stereotype. He need only know that it stands as a hypothesis about him in situations where the stereotype is relevant. We focused on the stereotype threat of African Americans in intellectual and scholastic domains to provide a compelling test of the theory and because the theory, should it be supported in this context for this group, would have relevance to an important set of outcomes.

Gaps in school achievement and retention rates between White and Black Americans at all levels of schooling have been strikingly persistent in American society (e.g., Steele, 1992). Well publicized at the kindergarten through 12th grade level, recent statistics show that they persist even at the college level where, for example, the national drop-out rate for Black college students (the percentage who do not complete college within a 6-year window of time) is 70% compared to 42% for White Americans (American Council on Education, 1990). Even among those who graduate, their grades average two thirds of letter grade lower than those of graduating Whites (e.g., Nettles, 1988). It has been most common to understand such problems as stemming largely from the socioeconomic disadvantage, segregation, and discrimination that African Americans have endured and continue to endure in this society, a set of conditions that, among other things, could produce racial gaps in achievement by undermining preparation for school.

Some evidence, however, questions the sufficiency of these explanations. It comes from the sizable literature examining racial bias in standardized testing. This work, involving hundreds of studies over several decades, generally shows that standardized tests predict subsequent school achievement as well for Black students as for White students (e.g., Cleary, Humphreys, Kendrick, & Wesman, 1975; Linn, 1973; Stanley, 1971). The slope of the lines regressing subsequent school achievement on entry-level standardized test scores is essentially the same for both groups. But embedded in this literature is another fact: At every level of preparation as measured by a standardized test—for example, the Scholastic Aptitude Test (SAT)—Black students with that score have poorer subsequent achievement—GPA, retention rates, time to graduation, and so on—than White students with that score (Jensen, 1980). This is variously known as the overprediction or underachievement phenomenon, because it indicates that, relative to Whites with the same score, standardized tests actually overpredict the achievement that Blacks will realize. Most important for our purposes, this evidence suggests that Black-White achievement gaps are not due solely to group differences in preparation. Blacks achieve less

well than Whites even when they have the same preparation, and even when that preparation is at a very high level. Could this underachievement, in some part, reflect the stereotype threat that is a chronic feature of these students' schooling environments?

Research from the early 1960s—largely that of Irwin Katz and his colleagues (e.g., Katz, 1964) on how desegregation affected the intellectual performance of Black students—shows the sizable influence on Black intellectual performance of factors that can be interpreted as manipulations of stereotype threat. Katz, Roberts, and Robinson (1965), for example, found that Black participants performed better on an IQ subtest when it was presented as a test of eye-hand coordination—a nonevaluative and thus threat-negating test representation—than when it was said to be a test of intelligence. Katz, Epps, and Axelson (1964) found that Black students performed better on an IQ test when they believed their performance would be compared to other Blacks as opposed to Whites. But as evidence that bears on our hypothesis, this literature has several limitations. Much of the research was conducted in an era when American race relations were different in important ways than they are now. Thus, without their being replicated, the extent to which these findings reflect enduring processes of stereotype threat as opposed to the racial dynamics of a specific historical era is not clear. Also, this research seldomly used White control groups. Thus it is difficult to know the extent to which some of the critical effects were mediated by the stereotype threat of Black students as opposed to processes experienced by any students.

Other research supports the present hypothesis by showing that factors akin to stereotype threat—that is, other factors that add self-evaluative threat to test taking or intellectual performance—are capable of disrupting that performance. The presence of observers or coactors, for example, can interfere with performance on mental tasks (e.g., Geen, 1985; Seta, 1982). Being a "token" member of a group—the sole representative of a social category—can inhibit one's memory for what is said during a group discussion (Lord & Saenz, 1985; Lord, Saenz, & Godfrey, 1987). Conditions that increase the importance of performing well—prizes, competition, and audience approval—have all been shown to impair performance of even motor skills (e.g., Baumeister, 1984). The stereotype threat hypothesis shares with these approaches the assumption that performance suffers when the situation redirects attention needed to perform a task onto some other concern—in the case of stereotype threat, a concern with the significance of one's performance in light of a devaluing stereotype.

For African American students, the act of taking a test purported to measure intellectual ability may be enough to induce this threat. But we assume that this is most likely to happen when the test is also frustrating. It is frustration that makes the stereotype—as an allegation of inability—relevant to their performance and thus raises the possibility that they have an inability linked to their race. This is not to argue that the stereotype is necessarily believed; only that, in the face of frustration with the test, it becomes more plausible as a self-characterization and thereby more threatening to the self. Thus for Black students who care about the skills being tested—that is, those who are identified with these skills in the sense of their self-regard being somewhat tied to having them—the stereo-

type loads the testing situation with an extra degree of self-threat, a degree not borne by people not stereotyped in this way. This additional threat, in turn, may interfere with their performance in a variety of ways: by causing an arousal that reduces the range of cues participants are able to use (e.g., Easterbrook, 1959), or by diverting attention onto task-irrelevant worries (e.g., Sarason, 1972; Wine, 1971), by causing an interfering self-consciousness (e.g., Baumeister, 1984), or overcautiousness (Geen, 1985). Or, through the ability-indicting interpretation it poses for test frustration, it could foster low performance expectations that would cause participants to withdraw effort (e.g., Bandura, 1977, 1986). Depending on the situation, several of these processes may be involved simultaneously or in alternation. Through these mechanisms, then, stereotype threat might be expected to undermine the standardized test performance of Black participants relative to White participants who, in this situation, do not suffer this added threat.

Study 1

Accordingly, Black and White college students in this experiment were given a 30-min test composed of items from the verbal Graduate Record Examination (GRE) that were difficult enough to be at the limits of most participants' skills. In the stereotype-threat condition, the test was described as diagnostic of intellectual ability, thus making the racial stereotype about intellectual ability relevant to Black participants' performance and establishing for them the threat of fulfilling it. In the non-stereotype-threat condition, the same test was described simply as a laboratory problem-solving task that was nondiagnostic of ability. Presumably, this would make the racial stereotype about ability irrelevant to Black participants' performance and thus preempt any threat of fulfilling it. Finally, a second nondiagnostic condition was included which exhorted participants to view the difficult test as a challenge. For practical reasons we were interested in whether stressing the challenge inherent in a difficult test might further increase participants' motivation and performance over what would occur in the nondiagnostic condition. The primary dependent measure in this experiment was participants' performance on the test adjusted for the influence of individual differences in skill level (operationalized as participants' verbal SAT scores).

We predicted that Black participants would underperform relative to Whites in the diagnostic condition where there was stereotype threat, but not in the two nondiagnostic conditions—the non-diagnostic-only condition and the non-diagnostic-plus-challenge condition—where this threat was presumably reduced. In the non-diagnostic-challenge condition, we also expected the additional motivation to boost the performance of both Black and White participants above that observed in the non-diagnostic-only condition. Several additional measures were included to assess the effectiveness of the manipulation and possible mediating states.

Method

Design and Participants

This experiment took the form of a 2×3 factorial design. The factors were race of the participant, Black or White, and a test description factor

in which the test was presented as either diagnostic of intellectual ability (the diagnostic condition), as a laboratory tool for studying problem solving (the non-diagnostic-only condition), or as both a problem-solving tool and a challenge (the non-diagnostic-challenge condition). Test performance was the primary dependent measure. We recruited 117 male and female, Black and White Stanford undergraduates through campus advertisements which offered \$10.00 for 1 hr of participation. The data from 3 participants were excluded from the analysis because they failed to provide their verbal SAT scores. This left a total of 114 participants randomly assigned to the three experimental conditions with the exception that we ensured an equal number of participants per condition.

Procedure

Participants who signed up for the experiment were contacted by telephone prior to their experimental participation and asked to provide their verbal and quantitative SAT scores, to rate their enjoyment of verbally oriented classes, and to provide background information (e.g., year in school, major, etc.). When participants arrived at the laboratory, the experimenter (a White man) explained that for the next 30 min they would work on a set of verbal problems in a format identical to the SAT exam, and end by answering some questions about their experience.

The participant was then given a page that stated the purpose of the study, described the procedure for answering questions, stressed the importance of indicating guessed answers (by a check), described the test as very difficult and that they should expect not to get many of the questions correct, and told them that they would be given feedback on their performance at the end of the session. We included the information about test difficulty to, as much as possible, equate participants' performance expectations across the conditions. And, by acknowledging the difficulty of the test, we wanted to reduce the possibility that participants would see the test as a miscalculation of their skills and perhaps reduce their effort. This description was the same for all conditions with the exception of several key phrases that comprised the experimental manipulation.

Participants in the diagnostic condition were told that the study was concerned with "various personal factors involved in performance on problems requiring reading and verbal reasoning abilities." They were further informed that after the test, feedback would be provided which "may be helpful to you by familiarizing you with some of your strengths and weaknesses" in verbal problem solving. As noted, participants in all conditions were told that they should not expect to get many items correct, and in the diagnostic condition, this test difficulty was justified as a means of providing a "genuine test of your verbal abilities and limitations so that we might better understand the factors involved in both." Participants were asked to give a strong effort in order to "help us in our analysis of your verbal ability."

In the non-diagnostic-only and non-diagnostic-challenge conditions, the description of the study made no reference to verbal ability. Instead, participants were told that the purpose of the research was to better understand the "psychological factors involved in solving verbal problems. . . ." These participants too were told that they would receive performance feedback, but it was justified as a means of familiarizing them "with the kinds of problems that appear on tests [they] may encounter in the future." In the non-diagnostic-only condition, the difficulty of the test was justified in terms of a research focus on difficult verbal problems and in the non-diagnostic-challenge condition it was justified as an attempt to provide "even highly verbal people with a mental challenge. . . ." Last, participants in both conditions were asked to give a genuine effort in order to "help us in our analysis of the problem solving process." As the experimenter left them to work on the test, to further differentiate the conditions, participants in the non-diagnostic-only condition were asked to try hard "even though we're not going to evaluate your ability." Participants in the non-diagnostic-challenge

condition were asked to "please take this challenge seriously even though we will not be evaluating your ability."

Dependent Measures

The primary dependent measure was participants' performance on 30 verbal items, 27 of which were difficult items taken from GRE study guides (only 30% of earlier samples had gotten these items correct) and 3 difficult anagram problems. Both the total number correct and an accuracy index of the number correct over the number attempted were analyzed.

Participants next completed an 18-item self-report measure of their current thoughts relating to academic competence and personal worth (e.g., "I feel confident about my abilities," "I feel self-conscious," "I feel as smart as others," etc.). These were measured on 5-point scales anchored by the phrases *not at all* (1) and *extremely* (5). Participants also completed a 12-item measure of cognitive interference frequently used in test anxiety research (Sarason, 1980) on which they indicated the frequency of several distracting thoughts during the exam (e.g., "I wondered what the experimenter would think of me," "I thought about how poorly I was doing," "I thought about the difficulty of the problems," etc.) by putting a number from 1 (*never*) to 5 (*very often*) next to each statement. Participants then rated how difficult and biased they considered the test on 15-point scales anchored by the labels *not at all* (1) and *extremely* (15). Next, participants evaluated their own performance by estimating the number of problems they correctly solved, and by comparing their own performance to that of the average Stanford student on a 15-point scale with the end points *much worse* (1) and *much better* (15). Finally, as a check on the manipulation, participants responded to the question:

The purpose of this experiment was to: (a) provide a genuine test of my abilities in order to examine personal factors involved in verbal ability; (b) provide a challenging test in order to examine factors involved in solving verbal problems; (c) present you with unfamiliar verbal problems to measure verbal learning.

Participants were asked to circle the appropriate response.

Results

Because there were no main or interactive effects of gender on verbal test performance or the self-report measures, we collapsed over this factor in all analyses.

Manipulation Check

Chi-square analyses performed on participants' responses to the postexperimental question about the purpose of the study revealed only an effect of condition, $\chi^2(2) = 43.18, p < .001$. Participants were more likely to believe the purpose of the experiment was to evaluate their abilities in the diagnostic condition (65%) than in the nondiagnostic condition (3%), or the challenge condition (11%).

Test Performance

The ANCOVA on the number of items participants got correct, using their self-reported SAT scores as the covariate (Black mean = 592, White mean = 632) revealed a significant condition main effect, $F(2, 107) = 4.74, p < .02$, with participants in the non-diagnostic-challenge condition performing higher than participants in the non-diagnostic-only and diagnostic conditions, respectively, and a significant race main effect, $F(1, 107)$

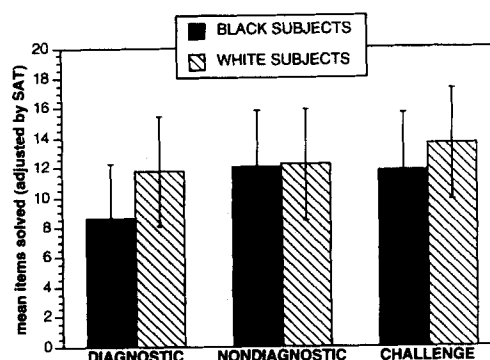


Figure 1. Mean test performance Study 1.

= 5.22, $p < .03$, with White participants performing higher than Black participants.¹ The race-by-condition interaction did not reach conventional significance ($p < .19$). The adjusted condition means are presented in Figure 1.

If making the test diagnostic of ability depresses the performance of Black students through stereotype threat, then their performance should be lower in the diagnostic condition than in either the non-diagnostic-only or non-diagnostic-challenge conditions which presumably lessened stereotype threat, and it should be lower than that of Whites in the diagnostic condition. Bonferroni contrasts² with SATs as a covariate supported this reasoning by showing that Black participants in the diagnostic condition performed significantly worse than Black participants in either the nondiagnostic condition, $t(107) = 2.88, p < .01$, or the challenge condition, $t(107) = 2.63, p < .01$, as well as significantly worse than White participants in the diagnostic condition $t(107) = 2.64, p < .01$.

But, as noted, the interaction testing the differential effect of test diagnosticity on Black and White participants did not reach significance. This may have happened, however, because an incidental pattern of means—Whites slightly outperforming Blacks in the nondiagnostic-challenge condition—undermined the overall interaction effect. To pursue a more sensitive test, we constructed a weighted contrast that compared the size of the race effect in the diagnostic condition with that in the nondiagnostic condition and assigned weights of zero to the White and Black non-diagnostic-challenge conditions. This analysis (including the use of SATs as a covariate) reached marginal significance, $F(1, 107) = 3.27, p < .08$. In sum, then, the hypothesis was supported by the pattern of contrasts, but when tested over the whole design, reached only marginal significance.

¹ Because we did not warn participants to avoid guessing in these experiments, we do not report the performance results in terms of the index used by Educational Testing Service, which includes a correction for guessing. This correction involves subtracting from the number correct, the number wrong adjusted for the number of response options for each wrong item and dividing this by the number of items on the test. Because 27 of our 30 items had the same number of response options (5), this correction amounts to adjusting the number correct almost invariably by the same number. All analyses are the same regardless of the index used.

² All comparisons of adjusted means reported hereafter used the Bonferroni procedure.

Accuracy

An ANCOVA on accuracy, the proportion correct of the number attempted, with SATs as the covariate, found that neither condition main effect nor the interaction reached significance, although there was a marginally significant tendency for Black participants to evidence less accuracy, $p < .10$. This tendency was primarily due to Black participants in the diagnostic condition who had the lowest adjusted mean accuracy of any group in the experiment, .420. The adjusted means for the White diagnostic, White non-diagnostic-only, White non-diagnostic-challenge, Black non-diagnostic-only, and Black diagnostic-challenge conditions were, .519, .518, .561, .546, and .490, respectively. Bonferroni tests revealed that Black participants in the diagnostic condition were reliably less accurate than Black participants in the non-diagnostic-only condition and White participants in the diagnostic condition, $t(107) = 2.64$, $p < .01$, and $t(107) = 2.13$, $p < .05$, respectively.

No condition or interaction effects reached significance for the number of items completed or the number of guesses participants recorded on the test (all $F_s < 1$). The overall means for these two measures were 22.9 and 4.1, respectively.

Self-Report Measures

There were no significant condition effects on the self-report measure of academic competence and personal worth or on the self-report measure of disruptive thoughts and feelings during the test. Analysis of participants' responses to the question about test bias yielded a main effect of race, $F(1, 107) = 10.47$, $p < .001$. Black participants in all conditions thought the test was more biased than White participants.

Perceived Performance

Participants' estimates of how many problems they solved correctly and of how they compared to other participants both showed significant condition main effects, $F(2, 106) = 7.91$, $p < .001$, and $F(2, 107) = 3.17$, $p < .05$, respectively. Performance estimates were higher in the non-diagnostic-only condition ($M = 11.81$) than in either the diagnostic ($M = 9.20$) or non-diagnostic-challenge conditions ($M = 8.15$). Bonferroni tests showed that Black participants in the diagnostic condition ($M = 4.89$) saw their relative performance as poorer than Black participants in the non-diagnostic-only condition ($M = 6.54$), $t(107) = 2.81$, $p < .01$, and than Black participants in the non-diagnostic-challenge condition ($M = 6.30$), $t(107) = 2.40$, $p < .02$, while test description had no effect on the ratings of White participants. The overall mean was 5.86.

Discussion

With SAT differences statistically controlled, Black participants performed worse than White participants when the test was presented as a measure of their ability, but improved dramatically, matching the performance of Whites, when the test was presented as less reflective of ability. Nonetheless, the race-by-diagnosticity interaction testing this relationship reached only marginal significance, and then, only when participants from the non-diagnostic-challenge condition were excluded

from the analysis. Thus there remained some question as to the reliability of this interaction.

We had also reasoned that stereotype threat might undermine performance by increasing interfering thoughts during the test. But the conditions affected neither self-evaluative thoughts nor thoughts about the self in the immediate situation (Sarason, 1980). Thus to further test the reliability of the predicted interaction and explore the mediation of the stereotype threat effect, we conducted a second experiment.

Study 2

We argued that the effect of stereotype threat on performance is mediated by an apprehension over possibly conforming to the negative group stereotype. Could this apprehension be detected as a higher level of general anxiety among stereotype-threatened participants? To test this possibility, participants in all conditions completed a version of the Spielberger State Anxiety Inventory (STAI) immediately after the test. This scale has been successfully used in other research to detect anxiety induced by evaluation apprehension (e.g., Geen, 1985). We also measured the amount of time they spent on each test item to learn whether greater anxiety was associated with more time spent answering items.

Method

Participants

Twenty Black and 20 White Stanford female undergraduates were randomly assigned (with the exception of attaining equal cell sizes) to either the diagnostic or the nondiagnostic conditions as described in Study 1, yielding 10 participants per condition. Female participants were used in this experiment because, due to other research going on, we had considerably easier access to Black female undergraduates than to Black male undergraduates. This decision was justified by the finding of no gender differences in the first study, or, as it turned out, in any of the subsequent studies reported in this article—all of which used both men and women.

Procedure

This experiment used the same test used in Study 1, with several exceptions; the final three anagram problems were deleted and the test period was reduced from 30 to 25 min. Also, the test was presented on a Macintosh computer (LCII). Participants controlled with the mouse how long each item or item component was on the screen and could, at their own pace, access whatever item material they wanted to see. The computer recorded the amount of time the items, or item components were on the screen as well as the number of referrals between item components (as in the reading comprehension items)—in addition to recording participants' answers.

Following the exam, participants completed the STAI and the cognitive interference measure described for Study 1. Also, on 11-point scales (with end-points *not at all* and *extremely*) participants indicated the extent to which they guessed when having difficulty, expended effort on the test, persisted on problems, limited their time on problems, read problems more than once, became frustrated and gave up, and felt that the test was biased.

Results and Discussion

The ANCOVA performed on the number of items correctly solved yielded a significant main effect of race, $F(1, 35) =$

10.04, $p < .01$, qualified by a significant Race $\times$ Test Description interaction, $F(1, 35) = 8.07$, $p < .01$. The mean SAT score for Black participants was 603 and for White participants 655. The adjusted means are presented in Figure 2. Planned contrasts on the adjusted scores revealed that, as predicted, Blacks in the diagnostic condition performed significantly worse than Blacks in the nondiagnostic condition $t(35) = 2.38$, $p < .02$, than Whites in the diagnostic condition $t(35) = 3.75$, $p < .001$, and than Whites in the nondiagnostic condition $t(35) = 2.34$, $p < .025$.

For accuracy—the number correct over the number attempted—a similar pattern emerged: Blacks in the diagnostic condition had lower accuracy ($M = .392$) than Blacks in the nondiagnostic condition ($M = .490$) or than Whites in either the diagnostic condition ($M = .485$) or the nondiagnostic condition ($M = .435$). The diagnosticity-by-race interaction testing this pattern reached significance, $F(1, 35) = 4.18$, $p < .05$. But the planned contrasts of the Black diagnostic condition against the other conditions did not reach conventional significance, although its contrasts with the Black nondiagnostic and White diagnostic conditions were marginally significant, with p s of .06 and .09 respectively.

Blacks completed fewer items than Whites, $F(1, 35) = 9.35$, $p < .01$, and participants in the diagnostic conditions tended to complete fewer items than those in the nondiagnostic conditions, $F(1, 35) = 3.69$, $p < .07$. The overall interaction did not reach significance. But planned contrasts revealed that Black participants in the diagnostic condition finished fewer items ($M = 12.38$) than Blacks in the nondiagnostic condition ($M = 18.53$), $t(35) = 2.50$, $p < .02$; than Whites in the diagnostic condition ($M = 20.93$), $t(35) = 3.39$, $p < .01$; and than Whites in the nondiagnostic condition ($M = 21.45$), $t(35) = 3.60$, $p < .01$.

These results establish the reliability of the diagnosticity-by-race interaction for test performance that was marginally significant in Study 1. They also reveal another dimension of the effect of stereotype threat. Black participants in the diagnostic condition completed fewer test items than participants in the other conditions. Test diagnosticity impaired the rate, as well as the accuracy of their work. This is precisely the impairment caused by evaluative pressures such as evaluation apprehension, test anxiety, and competitive pressure (e.g., Baumeister, 1984). But one might ask why this did not happen in the near-identical Study 1. Several factors may be relevant. First, the most involved test items—reading comprehension items that

took several steps to answer—came first in the test. And second, the test lasted 25 min in the present experiment whereas it lasted 30 min in the first experiment. Assuming, then, that stereotype threat slowed the pace of Black participants in the diagnostic conditions of both experiments, this 5-min difference in test period may have made it harder for these participants in the present experiment to get past the early, involved items and onto the more quickly answered items at the end of the test, a possibility that may also explain the generally lower scores in this experiment.

This view is reinforced by the ANCOVA (with SATs as a covariate) on the average time spent on each of the first five test items—the minimum number of items that all participants in all conditions answered. A marginal effect of test presentation emerged, $F(1, 35) = 3.52$, $p < .07$, but planned comparisons showed that Black participants in the diagnostic condition tended to be slower than participants in the other conditions. On average they spent 94 s answering each of these items in contrast to 71 for Black participants in the nondiagnostic condition, $t(35) = 2.39$, $p < .05$; 73 s for Whites in the diagnostic condition, $t(35) = 2.12$, $p < .05$, and 71 s for Whites in the nondiagnostic condition, $t(35) = 2.37$, $p < .05$. Like other forms of evaluative pressure, stereotype threat causes an impairment of both accuracy and speed of performance.

No differences were found on any of the remaining measures, including self-reported effort, cognitive interference, or anxiety. These measures may have been insensitive, or too delayed. Nonetheless, we lack an important kind of evidence. We have not shown that test diagnosticity causes in Black participants a specific apprehension about fulfilling the negative group stereotype about their ability—the apprehension that we argue disrupts their test performance. To examine this issue we conducted a third experiment.

Study 3

Taking an intellectually diagnostic test and experiencing some frustration with it, we have assumed, is enough to cause stereotype threat for Black participants. In testing this reasoning, the present experiment examines several specific propositions.

First, if taking or expecting to take a difficult, intellectually diagnostic test makes Black participants feel threatened by a specifically racial stereotype, then it might be expected to activate that stereotype in their thinking and information processing. That is, the racial stereotype, and perhaps also the self-doubts associated with it, should be more cognitively activated for these participants than for Black participants in the nondiagnostic condition or for White participants in either condition (e.g., Dovidio, Evans, & Tyler, 1986; Devine, 1989; Higgins, 1989). Accordingly, in testing whether test diagnosticity arouses this state, the present experiment measured the effect of conditions on the activation of this stereotype and of related self-doubts about ability.

Second, if test diagnosticity makes Black participants apprehensive about fulfilling and being judged by the racial stereotype, then these participants, more than participants in the other conditions, might be motivated to disassociate themselves from the stereotype. Brent Staples, an African American editorialist for the *New York Times*, offers an example of this in his recent autobiography, *Parallel Time*. He describes beginning graduate school at the University of Chicago and finding that as

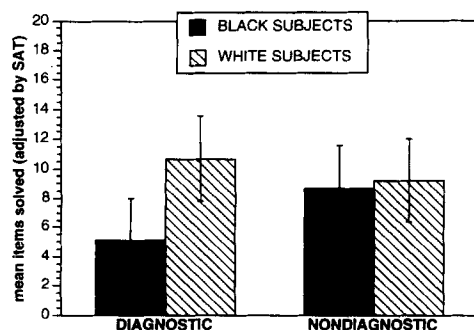


Figure 2. Mean test performance Study 2.

he walked the streets of Hyde Park he made people uncomfortable. They grouped more closely when he walked by, and some even crossed the street to avoid him. He eventually realized that in that urban context, dressed as a student, he was being perceived through the lens of a race-class stereotype as a potentially menacing Black man. To deflect this perception he learned a trick; he would whistle Vivaldi. It worked. Upon hearing him do this, people around him visibly relaxed and he felt out of suspicion. If it is apprehension about being judged in light of the racial stereotype that interferes with the performance of Black participants in the diagnostic condition, then these participants, like Staples, might be motivated to deflect such a perception by showing that the broader racial stereotype is not applicable to them. To test this possibility, the present experiment measured the effect of conditions on participants' stated preferences for such things as activities and styles of music, some of which were stereotypic of African Americans.

Third, by adding to the normal evaluative risks of test performance the further risk of self-validating the racial stereotype, the diagnostic condition should also make Black participants more apprehensive about their test performance. The present experiment measured this apprehension as the degree to which participants self-handicapped their expected performance, that is, endorsed excuses for poor performance before the test.

The experiment took the form of a 2×3 design in which the race of participants (African American or White) was crossed with diagnostic, nondiagnostic, and control conditions. The diagnostic and nondiagnostic conditions were the same as those described for Study 2, while in the control condition participants completed the critical dependent measures without expecting to take a test of any sort. In the experimental conditions, the dependent measures were administered immediately after the diagnosticity instructions and just before the test was ostensibly to be taken. These included measures of stereotype activation, stereotype avoidance, and, as a measure of general performance apprehension, participants' willingness to self-handicap. Participants in this experiment never took the test. The measures of stereotype activation and stereotype avoidance, we felt, could activate the racial stereotype and stereotype threat among Black participants in both the diagnostic and nondiagnostic conditions, making performance results difficult to interpret.

If test diagnosticity threatens Black participants with a specifically racial stereotype, then Black participants in the diagnostic condition, more than participants in the other conditions, should show greater cognitive activation of the stereotype and ability-related self-doubts, greater motivation to disassociate themselves from the stereotype, and greater performance apprehension as indicated by the endorsement of self-handicapping excuses.

Method

Participants

Thirty-five Black (9 male, 26 female) and 33 White (20 male, 13 female) Stanford undergraduates were randomly assigned to either a diagnostic, nondiagnostic, or control condition, yielding from 10 to 12 participants per experimental group.

Procedure

A White male experimenter gave a booklet to participants as they arrived that explained that the study was examining the relationship between two types of cognitive processes: lexical access processing (LAP) and higher verbal reasoning (HVR). They were told that they would be asked to complete two tasks, one of which measured LAP—"the visual and recognition processing of words"—and the other of which measured HVR—"abstract reasoning about the meaning of words." Test diagnosticity was manipulated as in Study 1 with the following written instructions to further differentiate the conditions:

Diagnostic: Because we want an accurate measure of your ability in these domains, we want to ask you to try as hard as you can to perform well on these tasks. At the end of the study, we can give you feedback which may be helpful by pointing out your strengths and weaknesses.

Nondiagnostic: Even though we are not evaluating your ability on these tasks, we want to ask you to try as hard as you can to perform well on these tasks. If you want to know more about your LAP and HVR performance, we can give you feedback at the end of the study.

Finally, participants were shown one sample item from the LAP (an item of the same sort as used in the fragment completion task) and three sample items from the HVR—difficult verbal GRE problems. The purpose of the HVR sample items was to alert participants to the difficulty of the test and the possibility of poor performance, thus occasioning the relevance of the racial stereotype in the diagnostic condition.

Participants in the control condition arrived at the laboratory to find a note on the door from the experimenter apologizing for not being present. The note instructed them to complete a set of measures lying on the desk in an envelope with the participant's name on it. The envelope contained the LAP word fragment measure and the stereotype avoidance measure (described below) with detailed instructions. No mention of verbal ability evaluation was made.

Measures

Stereotype activation. Participants first performed a word-fragment completion task, introduced as the "LAP task," versions of which have been shown to measure the cognitive activation of constructs that are either recently primed or self-generated (Gilbert & Hixon, 1991; Tulving, Schacter, & Stark, 1982). The task was made up of 80 word fragments with missing letters specified as blank spaces (e.g., ___ C E). Twelve of these fragments had as one possible solution a word reflecting either a race-related construct or an image associated with African Americans. The list was generated by having a group of 40 undergraduates (White students from the introductory psychology pool) generate a set of words that reflected the image of African Americans. From these lists, the research team identified the 12 most common constructs (e.g., lower class, minority) and selected single words to represent those constructs on the task. For example, the word "race" was used to represent the construct "concerned with race" on the task. Then, for each of the words placed on the task, at least two letter spaces were omitted and the word was checked again to determine whether other, non-stereotype-related associations to the word stem were possible. Leaving at least two letter spaces blank in each word fragment greatly unconstrains the number of word completions possible for each fragment when compared to leaving only one letter space blank. This reduces the chance of ceiling effects in which virtually all participants would think of the race-related fragment completion. The complete list was as follows: ___ C E (RACE); L A ___ (LAZY); ___ A C K (BLACK); ___ O R (POOR); C L ___ S ___ (CLASS); B R ___ (BROTHER); ___ T E (WHITE); M I ___

(MINORITY); W E L ____ (WELFARE); C O ____ (COLOR); T O ____ (TOKEN).

We included a fairly high number (12) of target fragments so that if ceiling or floor effects occurred on some fragments it would be less likely to damage the sensitivity of the overall measure. To reduce the chance that participants would become aware of the racial nature of the target fragments, they were spaced with at least three filler items between them, and there were only two target fragments per page in the task booklet. Participants were instructed to work quickly, spending no more than 15 s on each item.

Self-doubt activation. Seven word fragments reflecting self-doubts about competence and ability were included in the 80-item LAP task: L O ____ (LOSER); D U ____ (DUMB); S H A ____ (SHAME); ____ E R I O R (INFERIOR); F L ____ (FLUNK); ____ A R D (HARD); W ____ K (WEAK). These were generated by the research team, and again included at least two blank letter spaces in each fragment. As with the racial fragments, these were separated from one another (and from the racial fragments) by at least three filler items.

Stereotype avoidance. This measure asked participants to rate their preferences for a variety of activities and to rate the self-descriptiveness of various personality traits, some of which were associated with images of African Americans and African American life. Participants in the diagnostic and nondiagnostic conditions were told that these ratings were taken to give us a better understanding of the underpinnings of LAP and HVR processes. Control participants were told that these measures were being taken to assess the typical interests and personality traits of Stanford undergraduates. The measure contained 57 items asking participants to rate the extent to which they enjoyed a number of activities (e.g., pleasure reading, socializing, shopping, traveling, etc.), types of music (e.g., jazz, rap music, classical music), sports (e.g., baseball, basketball, boxing), and finally, how they saw themselves standing on various personality dimensions (e.g., extroverted, organized, humorous, etc.). All ratings were made on 7-point Likert scales with 1 indicating the lowest preference or degree of trait descriptiveness. Some of these activities and traits were stereotypic of African Americans. For an item to be selected as stereotypic, 65% of our pretest sample of 40 White participants had to have generated the item when asked to list activities and traits they believed to be stereotypic of African Americans. In the activities category, the stereotype-relevant items were: "How much do you enjoy sports?" and "How much do you enjoy being a lazy 'couch potato'?" The stereotype-relevant music preference item was *rap music*; the stereotype-relevant sports preference item was *basketball*; and the stereotype-relevant trait ratings were *lazy* and *aggressive/belligerent*.

Participants also completed a brief demographic questionnaire (asking their age, gender, major, etc.) just before they expected to begin the test. As another measure of participants' motivation to distance themselves from the stereotype, the second item of this questionnaire gave them the option of recording their race. We reasoned that participants who wanted to avoid having their performance viewed through the lens of a racial stereotype would be less willing to indicate their race.

Self-handicapping measure. This measure just preceded the demographic questionnaire. The directions stated "as you know, student life is sometimes stressful, and we may not always get enough sleep, etc. Such things can affect cognitive functioning, so it will be necessary to ask how prepared you feel." Participants then indicated the number of hours they slept the night before in addition to responding, on 7-point scales (with 7 being the higher rating on these dimensions) to the following questions: "How able to focus do you feel?;" "How much stress have you been under lately?;" "How tricky/unfair do you typically find standardized tests?"

Results

Stereotype Activation

A 2 (race) $\times$ 3 (condition: diagnostic, nondiagnostic, or control) ANCOVA (with verbal SAT as the covariate: Black

mean = 581, White mean = 650) was performed on the number of target word fragments filled in with stereotypic completions. This analysis yielded significant main effects for both race, $F(1, 61) = 13.77, p < .001$, and for experimental condition, $F(2, 61) = 5.90, p < .005$. These main effects, however, were qualified by a significant Race $\times$ Condition interaction, $F(2, 61) = 3.30, p < .05$. Figure 3 shows that as expected, the diagnostic condition significantly increased the number of race-related completions of Black participants but not of White participants. Black participants in the diagnostic condition produced more race-related completions ($M = 3.70$) than Black participants in the nondiagnostic condition ($M = 2.10$), $t(61) = 3.53, p < .001$, or for that matter, more than participants in any of other conditions, all $ps < .05$.

Self-Doubt Activation

It did the same for their self doubts. The number of self-doubt-related completions of self-doubt target fragments were submitted to an ANCOVA (as described above) yielding a main effect of experimental condition, $F(2, 61) = 4.33, p < .02$, and a Race $\times$ Condition interaction, $F(2, 61) = 3.34, p < .05$. As Figure 3 shows, Black participants in the diagnostic condition, as predicted, generated the most self-doubt-related completions, significantly more than Black participants in the nondiagnostic condition, $t(61) = 3.52, p < .001$, and more than participants in any of the other conditions as well, all $ps < .05$.

Stereotype Avoidance

The six preference and stereotype items described above were summed to form an index of stereotype avoidance that ranged from 6 to 42 with 6 indicating high avoidance and 42 indicating low avoidance (Cronbach's alpha = .65). When these scores were submitted to the ANCOVA they yielded a significant effect of condition, $F(2, 61) = 4.73, p < .02$, and a significant Race $\times$ Condition interaction, $F(2, 61) = 4.14, p < .03$. As can be seen in Figure 3, Black participants in the diagnostic condition were the most avoidant of conforming to stereotypic images of African Americans ($M = 20.80$), more so than Black participants in the nondiagnostic condition ($M = 29.80$), $t(61) = 3.61, p < .001$, and/or White participants in either condition, all $ps < .05$.

Indicating Race

Did the ability diagnosticity of the test affect participants' tendency to indicate their race on the demographic questionnaire? Among Black participants in the diagnostic condition, only 25% would indicate their race on the questionnaire, whereas 100% of the participants in each of the other conditions would do so. Using a 0/1 conversion of the response frequencies (with 0 = refusal to indicate race and 1 = indication of race) the standard ANCOVA performed on this measure revealed a marginally significant effect of race, $F(1, 61) = 3.86, p < .06$, a significant effect of condition, $F(2, 61) = 3.40, p < .04$, and a significant Race $\times$ Condition interaction, $F(1, 61) = 6.60, p < .01$, all due, of course, to the unique unwillingness of Black participants in the diagnostic condition to indicate their race.

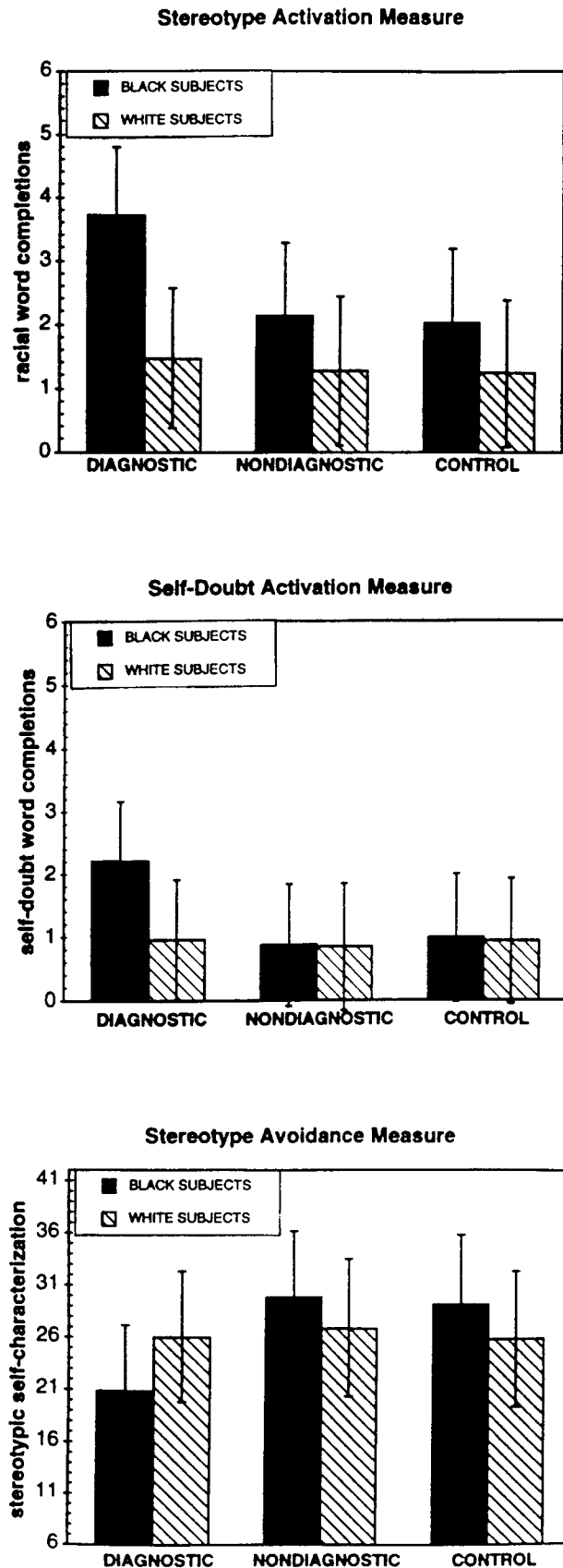


Figure 3. Indicators of stereotype threat.

Self-Handicapping

Four measures assessed participants' desire to claim impediments to performance. Because participants in the control conditions did not complete this measure, these responses were submitted to separate 2(race) $\times$ 2(diagnosticity) ANCOVAs. Cell means are presented in Table 1. Framing the verbal tasks as diagnostic of ability had significant effects on three of the four measures. For the number of hours of sleep, the ANCOVA yielded a significant effect of race, $F(1, 39) = 8.22, p < .01$, and a significant effect of condition, $F(1, 39) = 6.53, p < .02$. These effects were qualified by a significant Race $\times$ Condition interaction, $F(1, 39) = 4.1, p < .01$. For participants' ratings of their ability to focus, a similar result emerged: main effects of race, $F(1, 39) = 7.26, p < .02$, and condition, $F(1, 39) = 10.67, p < .01$, and a significant qualifying interaction, $F(1, 39) = 5.73, p < .03$. And finally, the same pattern of effects emerged for participants' ratings of how tricky or unfair they generally find standardized tests to be: a race main effect, $F(1, 39) = 13.24, p < .001$, a condition main effect, $F(1, 39) = 13.42, p < .001$, and a marginally significant, qualifying interaction, $F(1, 39) = 3.58, p < .07$. No significant effects emerged on participants' ratings of their current stress.

Discussion

We had assumed that presenting an intellectual test as diagnostic of ability would arouse a sense of stereotype threat in Black participants. The present results dramatically support this assumption. Compared to participants in the other conditions—that is, Blacks in the nondiagnostic condition and Whites in either condition—Black participants expecting to take a difficult, ability-diagnostic test showed significantly greater cognitive activation of stereotypes about Blacks, greater cognitive activation of concerns about their ability, a greater tendency to avoid racially stereotypic preferences, a greater tendency to make advance excuses for their performance, and finally, a greater reluctance to have their racial identity linked to their performance even in the pedestrian way of recording it on their questionnaires. Clearly the diagnostic instructions caused these participants to experience a strong apprehension, a distinct sense of stereotype threat.

Table 1
Self-Handicapping Responses in Study 3

Measure	Experimental condition			
	Diagnostic		Nondiagnostic	
	Blacks (n = 12)	Whites (n = 11)	Blacks (n = 11)	Whites (n = 10)
Hours of sleep	5.10 _a	7.48 _b	7.05 _b	7.70 _b
Ability to focus	4.03 _a	5.88 _b	5.85 _b	6.16 _b
Current stress	5.51 _a	5.24 _a	5.00 _a	5.02 _a
Tests unfair	5.46 _a	2.78 _b	3.14 _b	2.04 _b

Note. Means not sharing a common subscript differ at the .01 level according to Bonferroni procedure. Means sharing a common subscript do not differ.

So far, then, we have shown that representing a difficult test as diagnostic of ability can undermine the performance of Black participants, and that it can cause in them a distinct sense of being under threat of judgment by a racial stereotype. This manipulation of stereotype threat—in terms of test diagnosticity—is important because it establishes the generality of the effect to a broad range of real-life situations.

But two questions remain. The first is whether stereotype threat itself—in the absence of the test being explicitly diagnostic of ability—is sufficient to disrupt the performance of these participants on a difficult test. That is, we do not know whether mere activation of the stereotype in the test situation—without the test being explicitly diagnostic of ability—would be enough to cause such effects. A second question is whether the disruptive effect of the diagnosticity manipulation was in fact mediated by the stereotype threat it caused. Showing first that test diagnosticity disrupts Black participants' performance and then, separately, that it causes in these participants to be threatened by the stereotype, does not prove that the effect of test diagnosticity on performance was mediated by the stereotype threat it caused. The performance effect could have been mediated by some other effect of the diagnosticity manipulation. We conducted a fourth experiment to address these questions, and thereby, to test the replicability of the stereotype threat effect under different conditions.

Study 4

This experiment again crossed a manipulation of stereotype threat with the race of participants in a 2×2 design with test performance as the chief dependent measure. We addressed the first question above by representing the test in this experiment as nondiagnostic of ability. If stereotype threat then depressed Black participants' performance, we would know that stereotype threat is sufficient to cause this effect even when the test is not represented as diagnostic of ability. We addressed the second question by taking from Study 3 a dependent measure of stereotype threat that had been significantly affected by the diagnosticity manipulation, and manipulating that variable as an independent variable in the present experiment. If this manipulation then affects Black participants' performance, we would know that at least one aspect of the stereotype threat caused by the diagnosticity manipulation was able to impair performance. This would mean that the effect of that manipulation on performance was, or could have been, mediated by the stereotype threat it caused.

The variable that we manipulated in the present study was whether or not participants were required to list their race before taking the test. Recall that in Study 3, 75% of the Black participants in the diagnostic condition refused to record their race on the questionnaire when given the option, whereas all of the participants in the other conditions did. On the assumption that this was a sign of their stereotype avoidance, we reasoned that having participants record their race just prior to the test should prime the racial stereotype about ability for Black participants, and thus make them stereotype threatened. If this threat alone is sufficient to impair their performance, then, with SATs covaried, these participants should perform worse than White participants in this condition.

In the non-stereotype-threat conditions, the demographic questionnaire simply omitted the item requesting participants' race and, otherwise, followed the nondiagnostic procedures of Studies 1 and 2. Without raising the specters of ability or race-relevant evaluation, we expected Black participants in this condition to experience no stereotype threat and to perform (adjusted for SATs) on par with White participants.

Method

Design and Participants

This experiment took the form of a 2×2 design in which participants' race was crossed with whether or not they recorded their ethnicity on a preliminary questionnaire. Twenty-four Black (6 male, 18 female) and 23 White (11 male, 12 female) Stanford undergraduates were randomly assigned to either the race-prime condition or the no-race-prime condition. Data from two Black participants were discarded because they arrived with suspicions about the racial nature of the study. One White student failed to provide her SAT score and was discarded from data analyses. These participants were replaced to bring the number of participants in each of the four conditions to 11.

Procedure

The procedure closely paralleled that of the nondiagnostic conditions in Studies 1 and 2. After explaining the purpose and format of the test, the experimenter (White man) randomly assigned the participant to the race-prime or no-race-prime condition by drawing a brief questionnaire (labeled "personal information") from a shuffled stack. This questionnaire comprised the experimental manipulation. It was identical for all participants—asking them to provide their age, year in school, major, number of siblings, and parents' education—except that in the race-prime condition the final item asked participants to indicate their race. Because this questionnaire was given to the participant immediately prior to the test, the experimenter remained blind to the participant's condition throughout the pretest interaction. After ensuring that the participant had completed the questionnaire, the experimenter started the test and left the room. Twenty-five minutes later he returned, collected the test, and gave the participant a dependent measure questionnaire.

Dependent Measures

This experiment used the same 25-min test used in Study 2, but in this experiment it was administered on paper. During the test, participants marked their guesses, and after the test, they indicated on 11-point scales (with end points *not at all* and *extremely*) the extent to which they guessed when they were having difficulty, expended effort on the test, persisted on problems, limited their time on problems, read problems more than once, became frustrated and gave up, and felt that the test was biased.

Participants also completed a questionnaire aimed at measuring their stereotype threat, by expressing their agreement on 7-point scales (with endpoints *strongly disagree* and *strongly agree*) with each of eight statements (e.g., "Some people feel I have less verbal ability because of my race," "The test may have been easier for people of my race," "The experimenter expected me to do poorly because of my race," "In English classes people of my race often face biased evaluations," "My race does not affect people's perception of my verbal ability").

As a measure of academic identification, nine further items explored the effect of conditions on participants' perceptions of the importance of verbal and math skills to their education and intended

career (e.g., "verbal skills will be important to my career," "I am a verbally oriented person," "I feel that math is important to me," etc.). Participants responded to these items on 11-point scales with endpoints labeled *not at all* and *extremely*.

Results

Test Performance

A 2 (race) $\times$ 2 (race prime vs. no race prime) ANCOVA on test performance with self-reported SATs as a covariate (Black mean = 591, White mean = 643) revealed a strong condition interaction in the predicted direction. As Figure 4 shows, Blacks in the race-prime condition performed worse than virtually all of the other groups, yet in the no-race-prime condition their performance equaled that of Whites, $F(1, 39) = 7.82, p < .01$. Planned contrasts on these adjusted scores revealed that, as predicted, Blacks in the race-prime condition performed significantly worse than Blacks in the no-race-prime condition, $t(39) = 2.43, p < .02$, and significantly worse than Whites in the race-prime condition, $t(39) = 2.87, p < .01$. Black participants in the race-prime condition performed worse than Whites in the no-race-prime condition, but not significantly so. Nonetheless, the comparison pitting the Black race-prime condition against the three remaining conditions was highly significant, $F(1, 39) = 8.15, p < .01$.

Accuracy

The ANCOVA for this index—the percent correct of the items attempted for each participant—with participants' SATs as the covariate revealed a significant tendency for participants in the race-prime condition to have poorer accuracy, $F(1, 39) = 4.07, p = .05$. The adjusted means for the Black and White participants in the race-prime condition were .402 and .438 respectively, while those for the Black and White participants in the no-race-prime condition were .541 and .520 respectively. Condition contrasts did not reach significance, although the difference between the Black participants in the race-prime and no-race-prime conditions was marginally significant, $p < .08$. Again, these data suggest that lessened accuracy is part of the process through which stereotype threat impairs performance.

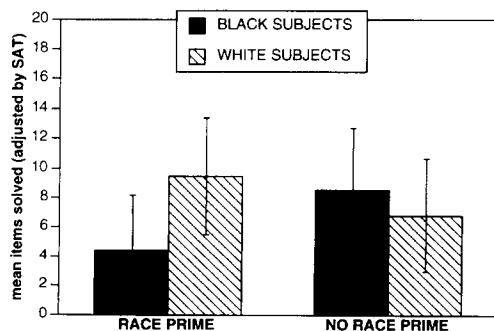


Figure 4. Mean test performance Study 4.

Number of Items Completed

An ANCOVA (again with SATs removed as a covariate) revealed only a significant Race $\times$ Race Prime interaction for the number of test items participants completed, $F(1, 39) = 12.13, p < .01$. In the race-prime condition Blacks completed fewer items than Whites, $t(39) = 3.83, p < .001$. The adjusted means were 11.58 and 20.15 respectively. In the no-race-prime condition, however, Blacks and Whites answered roughly the same number of problems. The adjusted means were 15.32 and 13.03, respectively.

Performance-Relevant Measures

Although participants' postexam ratings revealed no differences in the degree to which they thought they guessed on the test ($F < 1$), the ANCOVA performed on the actual number of guesses participants indicated on their test sheet revealed a Race $\times$ Race Prime interaction, $F(1, 39) = 5.56, p < .03$. Black participants made fewer guesses when race was primed ($M = 1.99$) than when it was not ($M = 2.74$), whereas White participants tended to guess more when race was primed ($M = 4.23$) than when it was not ($M = 1.58$). No significant condition effects emerged for participants' self-reported effort where, on an 11-point scale with 11 indicating *extremely hard* work, the overall mean was 8.84.

Participants' estimates of how well they had performed, taken after the test, showed no condition effects (the overall mean was 7.4 items). Neither were there condition effects on participants' ratings (made during the postexperimental debriefing) of how much having to indicate their ethnicity bothered them during the test (or *would* have bothered them in the case of participants in the no-race-prime condition). The overall mean was 3.31 on an 11-point scale for which 11 indicated the most distraction. Participants often stated in postexperimental interviews that they found recording their race unnoteworthy because they had to do it so often in everyday life. Of the items bearing on participants' experience taking the test, only one effect emerged: Black participants reported reading test items more than once to a greater degree than did White participants, $F(1, 39) = 8.62, p < .01$.

Stereotype Threat and Academic Identification Measures

A MANOVA of the stereotype threat scale revealed that Black participants felt more stereotype threat than White participants, $F(9, 31) = 8.80, p < .01$. No other effects reached significance. Analyses of participants' responses to questions regarding the personal importance of math, verbal skills, and athletics revealed that Black participants reported valuing sports less than Whites, $F(1, 39) = 4.11, p < .05$. As in Study 3, this result may reflect Black participants distancing themselves from the stereotype of the academically untalented Black athlete. Correlations between participants' numerical performance estimates and their ratings of the importance of sports, showed that for Blacks, the worse they believed they performed, the more they devalued sports—in the no-race-prime condition ($r = .56$), and particularly in the race-prime condition ($r = .70$).

Discussion

Priming racial identity depressed Black participants' performance on a difficult verbal test even when the test was not presented as diagnostic of intellectual ability. It did this, we assume, by directly making the stereotype mentally available and thus creating the self-threatening predicament that their performance could prove the stereotype self-characteristic. In Studies 1, 2 and 3, the stereotype was evoked indirectly by describing the test as diagnostic of an ability to which it was relevant. What this experiment shows is that mere cognitive availability of the racial stereotype is enough to depress Black participants' intellectual performance, and that this is so even when the test is presented as not diagnostic of intelligence. Also—because we know from Study 3 that the diagnosticity manipulation strongly affects participants' willingness to record their race—this finding shows that the performance-depressing effect of the diagnosticity manipulation in the earlier experiments was, or could have been, mediated by the effect of that manipulation on stereotype threat—as opposed to some other aspect of the manipulation.

Still, we had expected Black participants in the race-prime condition to show more stereotype threat (as measured by the stereotype threat and stereotype avoidance measures) than Black participants in the no-race-prime condition—reflecting the effect of the manipulation. Instead, while Blacks showed more stereotype threat than Whites, Blacks in the race-prime condition showed no more stereotype threat than Blacks in the no-race-prime condition. Nor did these groups differ on the identification measures. This may have happened for several reasons. These measures came after the test in this experiment, not before it as in Study 3. Thus, after experiencing the difficult, frustrating exam, all Black participants may have been somewhat stereotype threatened and stereotype avoidant (more so than the White participants) regardless of their condition. Also, the lack of a condition difference between Black participants on the stereotype threat and identification items may have occurred because these items asked participants to respond in reference to settings (e.g., English classes) and attitudes (e.g., about how one's race is generally regarded) that are beyond their immediate experience in the experiment.

Compared to participants in the other conditions, Black participants in the race-prime condition did not report expending less effort on the test; they were not more disturbed at having to list their race; and they did not guess more than other participants. Also, Black participants in both conditions reread the test items more than White participants. Such findings do not fit the idea that these participants underperformed because they withdrew effort from the experiment.

To establish the replicability of the race-prime effect and to explore the possible mediational role of anxiety, we conducted a two-condition experiment which randomly assigned only Black participants to either the race-prime or no-race-prime conditions described in Study 4. We also administered the test on computer to enable a measure of the time participants spent on the items, and gave participants an anxiety measure at the end of the experiment. Replicating Study 4, race-prime participants got significantly fewer items correct ($M = 4.4$) than no-race-prime participants ($M = 7.7$), $t(18) = 2.34$, $p < .04$; they were

marginally less accurate ($M = .334$) than no-race-prime participants ($M = .395$), $p = .10$; and they answered fewer items ($M = 13.2$) than no-race-prime participants ($M = 20.1$), $t(18) = 2.89$, $p < .01$. Race-prime participants spent more time on the first five test items (the number which all participants completed) ($M = 79$ s) than no-race-prime participants ($M = 61$ s), $t(18) = 2.27$, $p < .04$, and they were significantly more anxious than no-race-prime participants, $t(18) = 2.34$, $p < .04$. The means on the STAI were 48.5 and 40.5 respectively, on a scale that ranged from 20 (indicating *low anxiety*) to 80 (*extreme anxiety*). These results show that a race prime reliably depresses Black participants' performance on this difficult exam, and that it causes reactions that could be a response to stereotype threat—namely, an anxiety-based perseveration on especially the early test items, items that, as reading comprehension items, required multiple steps.

General Discussion

The existence of a negative stereotype about a group to which one belongs, we have argued, means that in situations where the stereotype is applicable, one is at risk of confirming it as a self-characterization, both to one's self and to others who know the stereotype. This is what is meant by stereotype threat. And when the stereotype involved demeans something as important as intellectual ability, this threat can be disruptive enough, we hypothesize, to impair intellectual performance.

In support of this reasoning, the present experiments show that making African American participants vulnerable to judgment by negative stereotypes about their group's intellectual ability depressed their standardized test performance relative to White participants, while conditions designed to alleviate this threat, improved their performance, equating the two groups once their differences in SATs were controlled. Studies 1 and 2 produced this pattern by varying whether or not the test was represented as diagnostic of intellectual ability—a procedure that varied stereotype threat by varying the relevance of the stereotype about Blacks' ability to their performance. Study 3 provided direct evidence that this manipulation aroused stereotype threat in Black participants by showing that it activated the racial stereotype and stereotype-related self-doubts in their thinking, that it led them to distance themselves from African American stereotypes. Study 4 showed that merely recording their race—presumably by making the stereotype salient—was enough to impair Black participants' performance even when the test was not diagnostic of ability. Taken together these experiments show that stereotype threat—established by quite subtle instructional differences—can impair the intellectual test performance of Black students, and that lifting it can dramatically improve that performance.

Mediation: How Stereotype Threat Impairs Performance

Study 3 offers clear evidence of what being stereotype threatened is like—as well as demonstrating that the mere prospect of a difficult, ability-diagnostic test was enough to do this to our sample of African American participants. But how precisely did this state of self-threat impair performance, through what mechanism or set of mechanisms did the impairment occur?

There are a number of possibilities: distraction, narrowed attention, anxiety, self-consciousness, withdrawal of effort, over-effort, and so on (e.g., Baumeister, 1984). In fact, several such mechanisms may be involved simultaneously, or different mechanisms may be involved under different conditions. For example, if the test were long enough to solidly engender low performance expectations, then withdrawal of effort might play a bigger mediational role than, say, anxiety, which might be more important with a shorter test. Such complexities notwithstanding, our findings offer some insight into how the present effects were mediated.

Our best assessment is that stereotype threat caused an inefficiency of processing much like that caused by other evaluative pressures. Stereotype-threatened participants spent more time doing fewer items more inaccurately—probably as a result of alternating their attention between trying to answer the items and trying to assess the self-significance of their frustration. This form of debilitation—reduced speed and accuracy—has been shown as a reaction to evaluation apprehension (e.g., Geen, 1985); test anxiety (e.g., Wine, 1971; Sarason, 1972); the presence of an audience (e.g., Bond, 1982); and competition (Baumeister, 1984). Several findings, by suggesting that stereotype-threatened participants were both motivated and inefficient, point in this direction. They reported expending as much effort as other participants. In those studies that included the requisite measures—Study 2 and the replication study reported with Study 4—they actually spent more time per item. They did not guess more than non-stereotype-threatened participants, and, as Black participants did generally, they reported rereading the items more. Also, as noted, these participants were strong students, and almost certainly identified with the material on the test. They may even have been more anxious. Stereotype threat increased Black participants' anxiety in the replication study, although not significantly in Study 2. Together then, these findings suggest that stereotype threat led participants to try hard but with impaired efficiency.

Still, we note that lower expectations may have also been involved, especially in real-life occurrences of stereotype threat. As performance falters under stereotype threat, and as the stereotype frames that faltering as a sign of a group-based inferiority, the individual's expectations about his or her ability and performance may drop—presumably faster than they would if the stereotype were not there to credit the inability interpretation. And lower expectations, as the literature has long emphasized (e.g., Bandura, 1977, 1986; Carver, Blaney, & Scheier, 1979; Pyszczynski & Greenberg, 1983) can further undermine performance by undermining motivation and effort. It is precisely a process of stereotype threat fostering low expectations in a domain that we suggest leads eventually to disidentification with the domain. We assume that this process did not get very far in the present research because the tests were short, and because our participants, as highly identified students, were unlikely to give up on these tests—as their self-reports tell us. But we do assume that lower expectations can play a role in mediating stereotype threat effects.

There is, however, strong evidence against one kind of expectancy mediation. This is the idea that lowered performance or self-efficacy expectations alone mediated the effects of stereotype

threat. Conceivably, the stereotype threat treatments got Black participants to expect that they would perform poorly on the test—presumably by getting them to accept the image of themselves inherent in the racial stereotype. The stereotype threat condition did activate participants' self-doubts. This lower expectation, then, outside of any experience these participants may have had with the test itself, and outside of any apprehension they may have had about self-confirming the stereotype, may have directly weakened their motivation and performance. Of course it would be important to show that stereotype threat effects are mediated in African American students by expectations implicit in the stereotype, expectations powerful enough to more or less automatically cause their underperformance.

But there are several reasons to doubt this view. For one thing, it isn't clear that our stereotype threat manipulations led Black participants to accept lower expectations and then to follow them unrevisedly to lower performance. For example, they resisted the self-applicability of the stereotype. But most important, as noted, it is almost certain that any expectation formed prior to the test would be superseded by the participants' actual experience with the test items; rising with success and falling with frustration. In fact, another experiment in our lab offered direct evidence of this by showing that expectations manipulated before the test had no effect on performance. Its procedure followed, in all conditions, that of the standard diagnostic condition used in Studies 1 and 2—with the exception that it directly manipulated efficacy and performance expectations before participants took the test. After being told that the test was ability diagnostic, and just before taking the test, the experimenter (an Asian woman) asked participants what their SAT scores were. After hearing the score, in the positive expectation condition, she commented that the participant should have little trouble with the test. In the negative expectation condition, this comment indicated that the participant would have trouble with the test, and nothing was said in a no-expectation condition. Both White and Black participants were run in all three expectation conditions. While the experiment replicated the standard effect of Whites outperforming Blacks under these stereotype threat conditions (participants' SATs were again used as a covariate) $F(1, 32) = 5.12, p < .03$, this personalized expectation manipulation had no effect on the performance of either group. For Blacks, the means were 4.32, 6.38, and 6.55, for the positive, negative and no-expectations conditions, respectively, and for Whites, for the same conditions, they were 8.24, 9.25, and 11.23, respectively. Thus in an experiment that was sensitive enough to replicate the standard stereotype threat effect, expectations explicitly manipulated before the test had no effect on performance. They are unlikely, then, to have been the medium through which stereotype threat affected performance in this research.

Finally, participants in all conditions of these experiments were given low performance expectations by telling them that they should expect to get few items correct due to the difficulty of the test. Importantly, this instruction did not depress the performance of participants in the non-stereotype-threat conditions. Thus it is not likely that a low performance expectation, implied by the stereotype, would have been powerful enough, by itself, to lower performance among these participants when a direct manipulation of the expectation could not.

The Emerging Picture of Stereotype Threat

In the social psychological literature there are other constructs that address the experience of potential victims of stereotypes. For clarity's sake, we briefly compare the construct of stereotype threat to these.

"Token" Status and Cognitive Functioning

Lord & Saenz (1985) have shown that token status in a group—that is, being the token minority in a group that is otherwise homogeneous—can cause deficits in cognitive functioning and memory, presumably as an outgrowth of the self-consciousness it causes. Although probably in the same family of effects as stereotype threat, token status would be expected to disrupt cognitive functioning even when the token individual is not targeted by a performance-relevant stereotype, as with, for example, a White man in a group of women solving math problems. Nor do stereotype threat effects require token status, as was shown in the present experiments. In real life, of course, these two processes may often co-occur, as for the Black in an otherwise non-Black classroom. They are nonetheless, distinct processes.

Attributional Ambiguity

Another important theory, and now extensive program of research by Crocker and Major (e.g., Crocker & Major, 1989; Crocker, Voelkl, Testa, & Major, 1991) examined how people contend with the self-evaluative implications of having a stigmatized identity. Both their theory and ours focus on the psychology of contending with social devaluation and differ most clearly in which aspect of this psychology they attend to. The work of Crocker and Major focused on the implications of this psychology for self-esteem maintenance (for example, the strategies available for protecting self-esteem against stigmatized status) and we have focused on its implications for intellectual performance. There is also a conceptual difference. Attributional ambiguity refers to the confusion a potential target of prejudice might have over whether or not he is being treated prejudicially. Stereotype threat, of course, refers to his apprehension over confirming, or eliciting the judgment that the stereotype is self-characteristic. Again, the two processes can co-occur—as for the woman who gets cut from the math team, for example—but are distinct.

The Earlier Research of the Katz Group

We also note that stereotype threat may explain the earlier findings of Katz and his colleagues. They found in the 1960s that the intellectual performance of Black participants rose and fell with conditions that seemed to vary in stereotype threat—for example, whether the test was represented as a test of intelligence or as one of psychomotor skill. A stereotype threat interpretation of these findings was foiled, however, by the lack of White participant control groups. Thus, the finding that manipulations very similar to Katz's depressed Black participants' performance while not depressing White participants' performance makes stereotype threat a parsimonious account of all these findings.

Test Difficulty and Racial Differences in Standardized Test Performance

The test used in these experiments is quite difficult, as the low performance scores indicate. As we argued, it may have to be at least somewhat demanding for stereotype threat to be occasioned. But acknowledging this parameter raises a question: Does stereotype threat significantly undermine the performance of Black students on the SAT? And if it does, is it appropriate to use the SAT as the standard for equating Black and White participants on skill level within our experiments? The answer to the first question has to be that it depends on how much frustration is experienced on the SAT. If the student perceives that a significant portion of the test is within his or her competence, it may preempt or override stereotype threat by proving the stereotype inapplicable. When the student cannot gain this perception, however, the group stereotype becomes relevant as an explanation and may undermine performance. Thus we surmise that over the entire range of Black student test takers, stereotype threat causes a significant depression of scores.

And, of course, this point holds more generally. An important implication of this research is that stereotype threat is an underappreciated source of classic deficits in standardized test performance (e.g., IQ) suffered by Blacks and other stereotype-threatened groups such as those of lower socioeconomic status and women in mathematics (Herrnstein, 1973; Jensen, 1969, 1980; Spencer & Steele, 1994). In addition to whatever environmental or genetic endowments a person brings to the testing situation, this research shows that this situation is not group-neutral—not even, quite possibly, when the tester and test content have been accommodated to the test-taker's background. The problem is that stereotypes afoot in the larger society establish a predicament in the testing situation—aside from test content—that still has the power to undermine standardized test performance, and, we suspect, contribute powerfully to the pattern of group differences that have characterized these tests since their inception.

But, for several reasons, we doubt that this possibility compromises the interpretation of the present findings. First, it is unlikely that stereotype threat had much differential effect on the SATs of our Black and White participants since both groups, as highly selected students, are not likely to have experienced very great frustration on these tests. Second, even if our Black participants' SATs were more depressed in this way, using such depressed scores as a covariate in the present analyses would only adjust Black performance more in the direction of reducing the Black-White difference in the stereotype threat conditions. Thus, while a self-threateningly difficult test is probably a necessary condition for stereotype threat, and while stereotype threat may commonly depress the standardized test performance of Black test takers, these facts are not likely to have compromised the present results.

In conclusion, our focus in this research has been on how social context and group identity come together to mediate an important behavior. This approach is Lewinian; it is also hopeful. Compared to viewing the problem of Black underachievement as rooted in something about the group or its societal conditions, this analysis uncovers a social psychological predicament of race, rife in the standardized testing situation, that is amenable to change—as we hope our manipulations have illustrated.

References

- Allport, G. (1954). *The nature of prejudice*. New York: Addison-Wesley.
- American Council on Education. (1990). *Minorities in higher education*. Washington, DC: Office of Minority Concerns.
- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84, 191-215.
- Bandura, A. (1986). Fearful expectations and avoidant actions as coeffects of perceived self-inefficacy. *American Psychologist*, 41, 1389-1391.
- Baumeister, R. F. (1984). Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of Personality and Social Psychology*, 46, 610-620.
- Bond, C. F. (1982). Social facilitation: A self-presentational view. *Journal of Personality and Social Psychology*, 42, 1042-1050.
- Carter, S. L. (1991). *Reflections of an affirmative action baby*. New York: Basic Books.
- Carver, C. S., Blaney, P. H., & Scheier, M. F. (1979). Reassertion and giving up: The interactive role of self-directed attention and outcome expectancy. *Journal of Personality and Social Psychology*, 37, 1859-1870.
- Cleary, T. A., Humphreys, L. G., Kendrick, S. A., & Wesman, A. (1975). Educational uses of tests with disadvantaged students. *American Psychologist*, 30, 15-41.
- Crocker, J., & Major, B. (1989). Social stigma and self-esteem: The self-protective properties of stigma. *Psychological Review*, 96, 608-630.
- Crocker, J., Voelkl, K., Testa, M., & Major, B. (1991). Social stigma: The affective consequences of attributional ambiguity. *Journal of Personality and Social Psychology*, 60, 218-228.
- Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, 56, 5-18.
- Dovidio, J. F., Evans, N., & Tyler, R. B. (1986). Racial stereotypes: The contents of their cognitive representations. *Journal of Experimental Social Psychology*, 22, 22-37.
- Easterbrook, J. A. (1959). The effect of emotion on cue utilization and the organization of behavior. *Psychological Review*, 66, 183-201.
- Geen, R. G. (1985). Evaluation apprehension and response withholding in solution of anagrams. *Personality and Individual Differences*, 6, 293-298.
- Geen, R. G. (1991). Social motivation. *Annual Review of Psychology*, 42, 377-399.
- Gilbert, D. T., & Hixon, J. G. (1991). The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology*, 60, 509-517.
- Goffman, I. (1963). *Stigma*. New York: Simon & Shuster, Inc.
- Herrnstein, R. (1973). *IQ in the meritocracy*. Boston: Little Brown.
- Higgins, E. T. (1989). Knowledge accessibility and activation: Subjectivity and suffering from unconscious sources. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended Thoughts* (pp. 75-123). New York: Guilford.
- Jensen, A. R. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review*, 39, 1-123.
- Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.
- Katz, I. (1964). Review of evidence relating to effects of desegregation on the intellectual performance of Negroes. *American Psychologist*, 19, 381-399.
- Katz, I., Epps, E. G., & Axelson, L. J. (1964). Effect upon Negro digit symbol performance of comparison with Whites and with other Negroes. *Journal of Abnormal and Social Psychology*, 69, 963-970.
- Katz, I., Roberts, S. O., & Robinson, J. M. (1965). Effects of task difficulty, race of administrator, and instructions on digit-symbol performance of Negroes. *Journal of Personality and Social Psychology*, 2, 53-59.
- Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research*, 43, 139-161.
- Lord, C. G., & Saenz, D. S. (1985). Memory deficits and memory surfeits: Differential cognitive consequences of tokenism for tokens and observers. *Journal of Personality and Social Psychology*, 49, 918-926.
- Lord, C. G., Saenz, D. S., & Godfrey, D. K. (1987). Effects of perceived scrutiny on participant memory for social interactions. *Journal of Experimental Social Psychology*, 23, 498-517.
- Nettles, M. T. (1988). *Toward undergraduate student equality in American higher education*. New York: Greenwood.
- Pyszczynski, T., & Greenberg, J. (1983). Determinants of reduction in effort as a strategy for coping with anticipated failure. *Journal of Research in Personality*, 17, 412-422.
- Sarason, I. G. (1972). Experimental approaches to test anxiety: Attention and the uses of information. In C. D. Spielberger (Ed.), *Anxiety: Current trends in theory and research* (Vol. 2). New York: Academic Press.
- Seta, J. J. (1982). The impact of coactors' comparison processes on task performance. *Journal of Personality and Social Psychology*, 42, 281-291.
- Spencer, S. J., & Steele, C. M. (1994). *Under suspicion of inability: Stereotype vulnerability and women's math performance*. Unpublished manuscript, State University of New York at Buffalo and Stanford University.
- Stanley, J. C. (1971). Predicting college success of the educationally disadvantaged. *Science*, 171, 640-647.
- Steele, C. M. (1992, April). Race and the schooling of black Americans. *The Atlantic Monthly*.
- Steele, S. (1990). *The content of our character*. New York: St. Martin's Press.
- Tulving, E., Schacter, D. L., & Stark, H. A. (1982). Priming effects in word-fragment completion are independent of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8, 336-342.
- Wine, J. (1971). Test anxiety and direction of attention. *Psychological Bulletin*, 76, 92-104.

Received August 9, 1994

Revision received May 9, 1995

Accepted May 18, 1995 ■