Research Memorandum

The Challenge of Stereotype Threat for the Testing Community

Lawrence J. Stricker

July 2008

ETS RM-08-12

The Challenge of Stereotype Threat for the Testing Community

Lawrence J. Stricker

ETS, Princeton, NJ

July 2008

Abstract

This paper describes research on stereotype threat by my colleagues and me. One aim was to document its effects on operational tests in high-stakes settings. The second was to investigate possible changes in the tests that might minimize any effects of stereotype threat. Research by others on operational tests in these same kinds of settings is also described. Methodological issues in this research and alternative approaches to investigating stereotype threat are discussed.

Key words: Stereotype threat, AP® examination, Computer Placement Test Battery, SAT®, ASVAB, ethnic differences, gender differences

Acknowledgments

This is the presidential address to the Division of Evaluation, Measurement, and Statistics at the American Psychological Association meeting, San Francisco, August 2007.

Thanks are due to Isaac Bejar and Nathan Kogan for reviewing a draft of this paper.

Stereotype threat is a concern about fulfilling a negative stereotype about the ability of one's group when placed in a situation where this ability is being evaluated, such as when taking a cognitive test. These negative stereotypes exist about minorities, women, the working class, and the elderly. The result of this concern is that performance on the ability assessment is adversely affected. (See Steele, 1997.)

Research on stereotype threat began with Steele and Aronson's (1995) experiments. This phenomenon is illustrated by two classic studies. In one, Steele and Aronson (Study 2) administered a test made up of difficult verbal ability items from the Graduate Record Examinations® (GRE®) General Test to Black and White undergraduates. Stereotype threat was manipulated by describing the test in one group as a laboratory tool for studying verbal problem solving (the low threat condition); the other group was told that the test was diagnostic of intellectual ability (the high threat condition). Black participants' scores were lower in the high threat condition than in the low threat condition, while White participants' scores were unaffected (See Figure 1.).

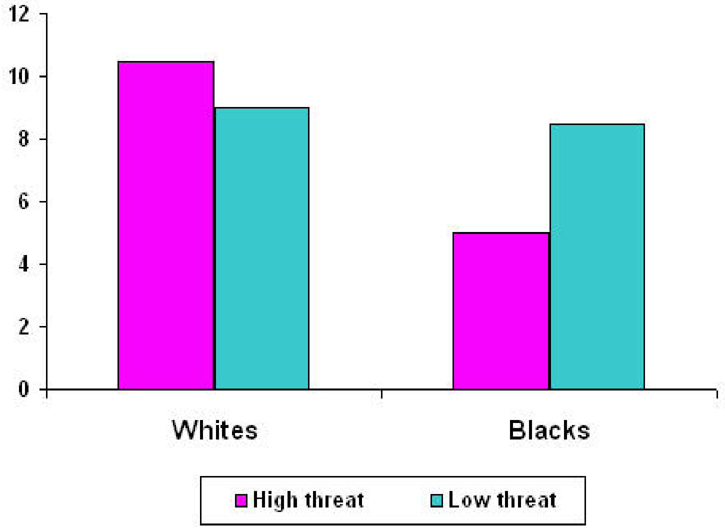

Figure 1. **Verbal score (items from GRE General Test)—covariance adjusted for SAT® scores. (Data from Steele & Aronson, 1995, Study 2)**

In the second study, Spencer, Steele, and Quinn (1999, Study 3) administered a test made up of quantitative ability items from the Graduate Management Admission Test (GMAT) to women and men undergraduates. Stereotype threat was manipulated by telling one group that there was no gender difference on the test (the no difference, low threat condition) and by telling the other group nothing (the control, high threat condition). Women’s scores were lower in the control (high threat) condition than in the no difference (low threat) condition, while men’s scores were unaffected (See Figure 2.).

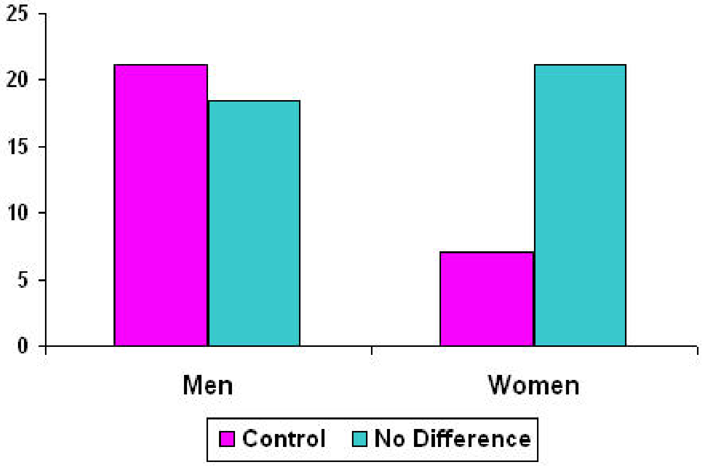

***Figure 2.* Mathematical score (items from GMAT). (Data from Spencer, Steele, & Quinn, 1999, Study 3)**

Similar effects have been amply documented in laboratory studies by a variety of investigators using a wide range of populations of research participants and a number of different ability measures. Well over a hundred studies have been conducted since the initial Steele and Aronson (1995) work, almost all of them in the laboratory. In a meta-analysis of 43 studies completed by 2001, Walton and Cohen (2003) found that the mean effect size was a *d* of .29.

Stereotype threat has understandably engendered considerable interest—and concern—among scientists and the public because of its implications for performance on operational tests. This interest is magnified because of the potential of stereotype threat in accounting for the deficits on standardized cognitive tests that are commonly observed for minority groups, women (on quantitative tests), and working-class people. And, of course, stereotype threat may have consequences for the validity of ability and achievement tests used in selection, admissions, and other high-stakes settings.

I want to describe some research that I have conducted on stereotype threat with William Ward (Stricker & Ward, 2004) and Isaac Bejar (Stricker & Bejar, 2004). Our research had two aims. One was to document the effect of stereotype threat on operational tests in high-stakes settings. The other was to investigate possible changes that could be made in tests to minimize any effects of stereotype threat. I will also describe the few studies done by others with operational tests in the same kind of settings. Finally, I will discuss methodological issues in research on stereotype threat and alternative approaches to investigating this phenomenon.

Our two studies of stereotype threat on operational tests (Stricker & Ward, 2004) were stimulated by a Steele and Aronson (1995, Study 4) experiment that found the performance of Black research participants on difficult verbal ability items from the GRE General Test was depressed when they were asked about their ethnicity immediately prior to working on the items, whereas the performance of White participants was unaffected. Prime is the asked condition; No Prime is the not asked condition (See Figure 3.). Merely asking about ethnicity presumably primed stereotype threat for Black participants by making their ethnicity salient.

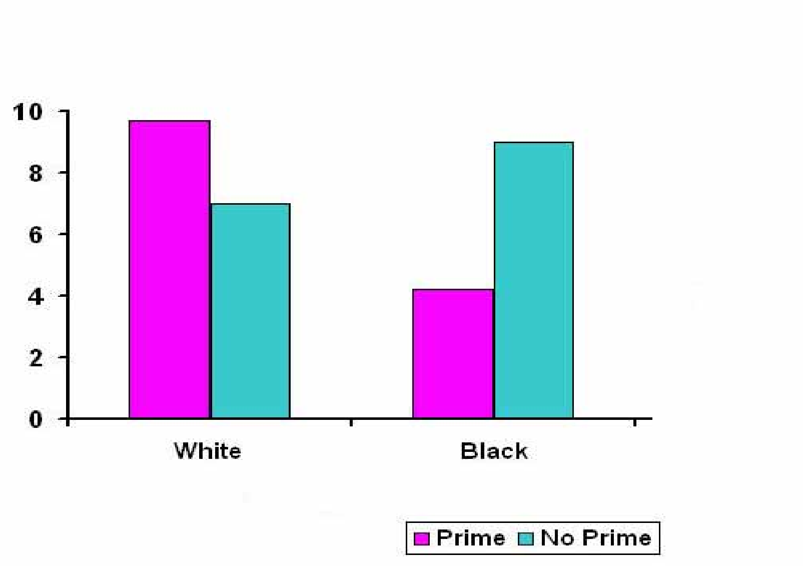

***Figure 3.* Verbal score (items from GRE General Test)—covariance adjusted for SAT scores. (Data from Steele & Aronson, 1995, Study 4)**

Inquiring about ethnicity has obvious parallels with the test administration procedures for widely used standardized tests that routinely ask examinees about their ethnicity and gender just before they take the tests. These tests include the Advanced Placement Program® (AP®) examinations and the Computerized Placement Tests (CPTs). (The CPTs are a battery of basic skills tests covering reading, writing, and math that college students take for course placement.)

Our experiments extended the Steele and Aronson (1995) work on inquiring about ethnicity in the laboratory to inquiring about ethnicity and gender in taking the AP examinations and CPTs in actual test administrations. Our hypotheses were that inquiring about ethnicity depresses the performance of Black students on all of the tests and that inquiring about gender depresses the performance of women on the quantitative tests.

Both experiments altered the standard test administration for some students by eliminating the usual questions about ethnicity and gender (the no prime group) and contrasted their test performance with the performance of comparable students who were asked these questions in the course of the standard test administration (the prime group).

The AP experiment used the Calculus AB examination. The no prime condition had a sample of 77 AP classes with 755 students; the prime condition had 77 classes with 897 students. The means for the AP grades are shown in Figure 4. (The bars are for the .95 confidence limits.) None of the differences in the means between the no prime and prime conditions was statistically significant for any ethnic group or gender.

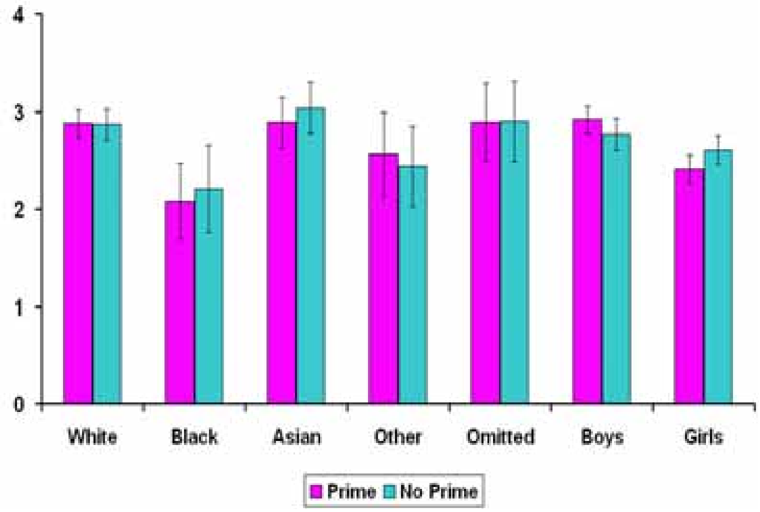

***Figure 4.* AP grade. (Data from Stricker & Ward, 2004)**

The CPTs experiment was based on all incoming students at a community college who routinely took this test battery during four weeks before the school year began. The no prime condition had 632 students who took the battery during a two-week period; the prime condition had 709 students who took the battery during the two adjacent weeks. The means for the four test scores are shown in Figures 5 to 8. One of the differences in the means between the no prime and prime conditions was statistically significant but not practically significant: Men in the no prime condition underperformed men in the prime condition on Elementary Algebra. None of the differences was statistically significant for any subgroup on the three other tests.

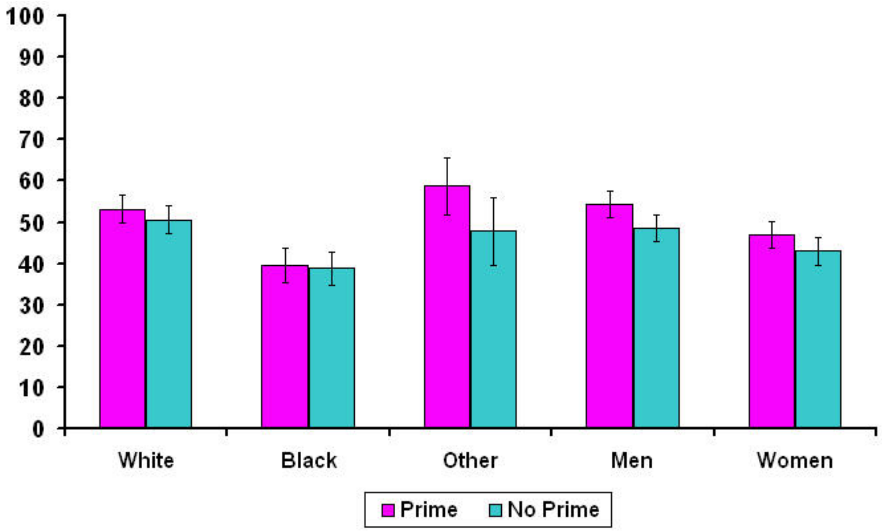

Figure 5. **CPT Elementary Algebra score. (Data from Stricker & Ward, 2004)**

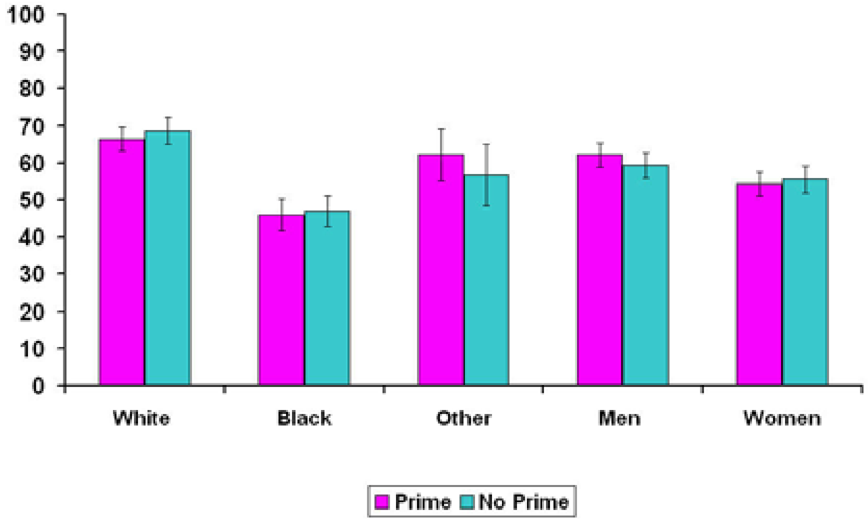

Figure 6. **CPT Arithmetic score. (Data from Stricker & Ward, 2004)**

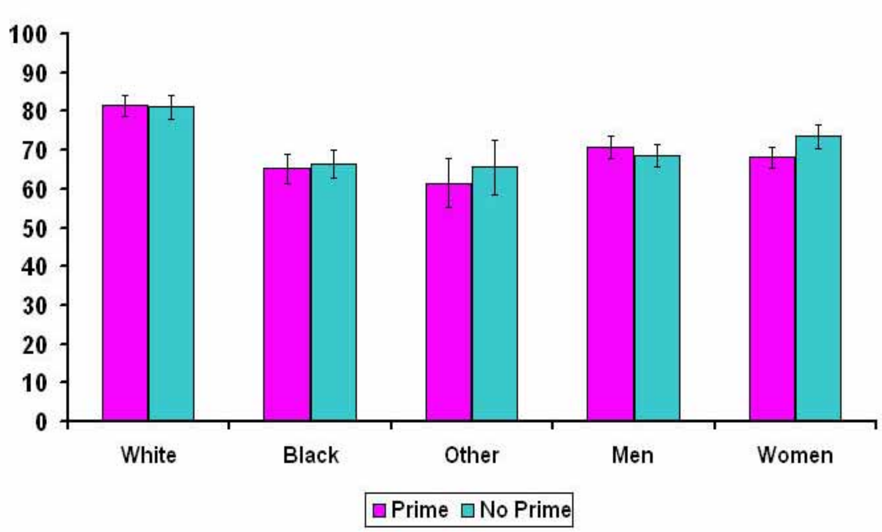

Figure 7. **CPT Reading Comprehension score. (Data from Stricker & Ward, 2004)**

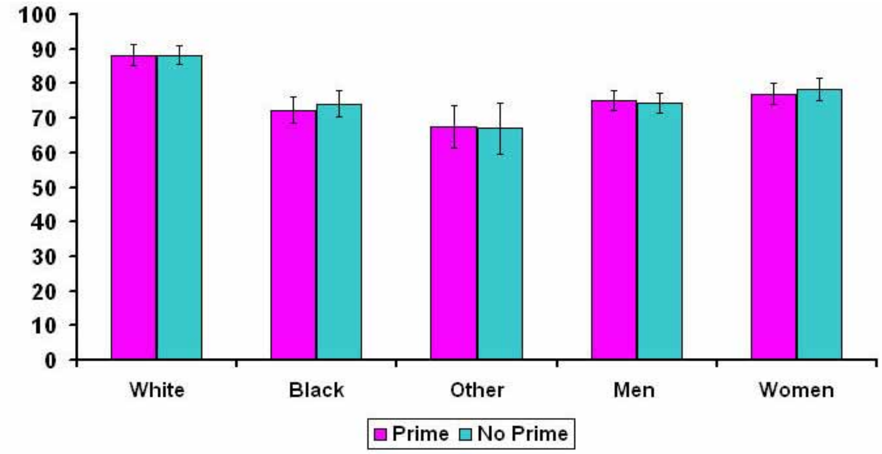

Figure 8. **CPT Sentence Skills score. (Data from Stricker & Ward, 2004)**

In short, in the two experiments, inquiring about ethnicity and gender did not affect the level of test performance on the AP examination and the CPTs for Black students or women. The congruence in the results for the studies supports their generalizability. These outcomes fail to confirm the hypotheses about the adverse effects for Black and women students based on the Steele and Aronson (1995) findings. Our findings, from a scientific perspective, shed some light on the generalizability of the Steele and Aronson results to operational tests. And from a practical standpoint, our results suggest that the common practice of making these inquiries does not degrade the performance of test takers.

Stereotype threat on operational tests has also been investigated in several correlational studies by Cullen and his colleagues (Cullen, Hardison, & Sackett, 2004; Cullen, Waters, & Sackett, 2006) and in an experiment by Good, Aronson, and Inzlicht (2003). Cullen et al. (2004) examined large data sets that had operational test scores and real-world criteria. One data set, from the College Board, consisted of SAT scores and freshman grade point average (GPA) for 49,374 college students from 13 universities. The other data set, from the Army's Project A, consisted of Armed Services Vocational Aptitude Battery (ASVAB) scores and two on-the-job criteria for 5,397 soldiers from 13 military occupational specialties. One criterion was core technical proficiency (on the central tasks specific to the particular job). The other criterion was general soldiering proficiency (on tasks common across all jobs). The criteria were derived from supervisory ratings, work samples, and job knowledge tests.

The analyses focused on differential prediction by gender or by ethnicity in predicting the criteria from the test scores. Regressions were examined for nonlinearity or discontinuity for women or for Black test takers that might be expected from the operation of stereotype threat. Several kinds of nonlinearity or discontinuity were anticipated; all were based on the idea that the test scores should underestimate the performance of test takers from stereotyped groups at the upper end of the test score distribution. This idea came from Steele's (1997) suggestion that stereotype threat should have its greatest effect on the higher achievers because they are more strongly identified with the domain being evaluated. Panel a in Figure 9 shows the same regression line for the stereotyped and nonstereotyped groups over the entire score range and hence no overprediction or underprediction for the stereotyped group. Panels b–d in Figures 9 show discontinuity (Panel b) or nonlinearity (Panels c and d) in the regression line for the stereotype group in the upper end of the score range and hence underprediction for this group in that range. One analysis was done for the SAT-Mathematical scores and English GPA for women and men. (English GPA rather than mathematics GPA or overall GPA was used as the criterion because no negative stereotypes exist about women's performance in English, and hence English GPA should be unaffected by stereotype threat.) The empirically fitted regression curves for women and men are shown in Figure 10. The regression was linear for women across the entire test score range. However, their GPA was underpredicted. Although this underprediction is consistent with expectations from stereotype threat theory, additional analyses using SAT-Verbal scores, which are not susceptible to stereotype threat for women (no negative stereotypes exist about women's

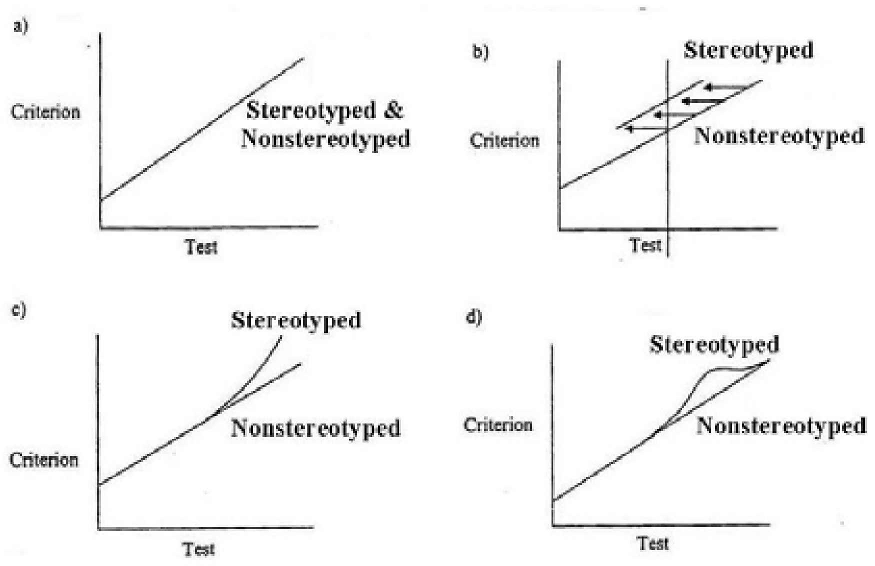

Figure 9. **Hypothesized test-criterion relationships.**

Note: Adapted from “Using SAT-Grade and Ability-Job Performance Relationships to Test Predictions Derived From Stereotype Threat Theory,” by M. J. Cullen, C. M. Hardison, and P. R. Sackett, 2004, *Journal of Applied Psychology, 89,* p. 221. Copyright 2004 by the American Psychological Association. Adapted with permission.

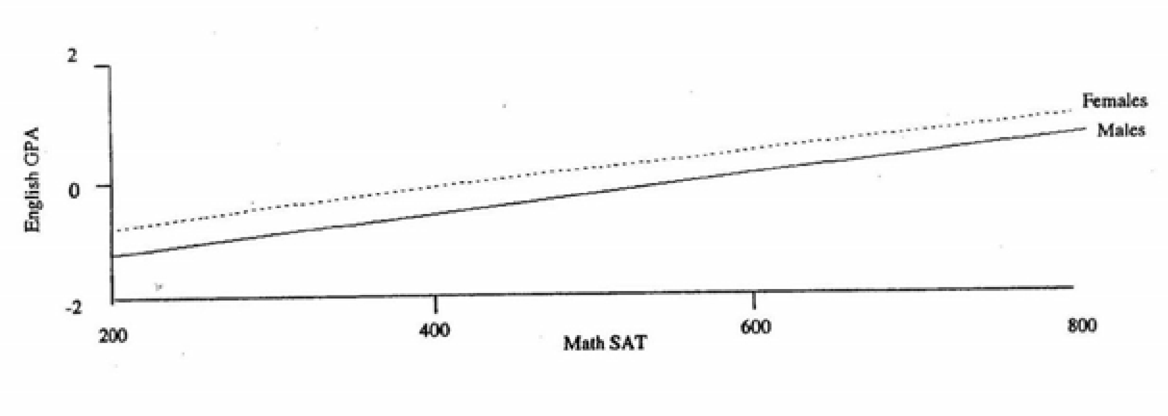

Figure 10. **SAT-Mathematical versus English GPA by gender.**

Note: From “Using SAT-Grade and Ability-Job Performance Relationships to Test Predictions Derived From Stereotype Threat Theory,” by M. J. Cullen, C. M. Hardison, and P. R. Sackett, 2004, *Journal of Applied Psychology, 89*, p. 225. Copyright 2004 by the American Psychological Association. Reprinted with permission.

verbal ability), found the same underprediction. This suggests that the underprediction for SAT-Mathematical is not attributable to stereotype threat.

A follow-up analysis (Cullen et al., 2006) was limited to 2,687 White women and men who either identified with mathematics (as indicated by their stated intention to major in quantitatively oriented fields in college) or did not identify with mathematics. This restriction was imposed to deal more directly with the already mentioned possibility that stereotype threat is most potent for individuals identified with the domain being assessed. The sample was limited to White students to eliminate any confounding with ethnicity. The focus was on gender differences in the relationships of SAT-Mathematical for mathematics-identified and nonmathematics-identified women and men.

The regressions for SAT-Mathematical against English GPA for mathematics-identified and nonmathematics-identified women are shown in Figure 11. (English GPA is actually residualized, partialing out SAT-Verbal scores and amount of course work and related activities.) English GPA was overpredicted for the mathematics-identified women.

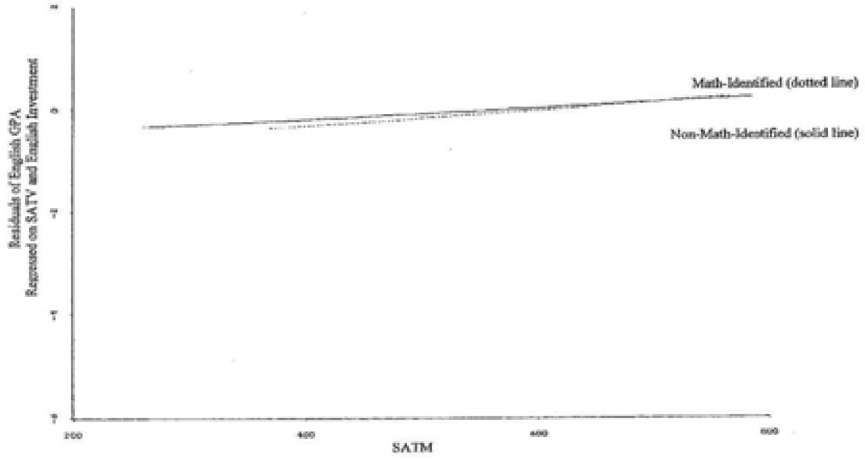

Figure 11. **SAT-Mathematical versus English GPA (residualized) for women.**

Note: From “Testing Stereotype Threat Theory Predictions for Math-Identified and Non-Math Identified Students by Gender,” by M. J. Cullen, S. D. Waters, and P.R. Sackett, 2006, *Human Performance, 19,* p. 437. Copyright 2006 by Lawrence Erlbaum Associates. Reprinted with permission.

The corresponding regressions for men are shown in Figure 12. English GPA was also overpredicted for the mathematics-identified men. The key point is that the difference in the regressions for the mathematics-identified and nonmathematics- identified students were the same for the women and the men, not a greater difference for mathematics-identified and nonmathematics-identified women.

Analyses were also done for SAT-Mathematical and SAT-Verbal scores against overall GPA for Black and White students (Cullen et al. 2004). (Cullen et al. noted that this GPA criterion was problematic because course examinations may have been affected by stereotype threat for Black students, just like the SAT itself.) The SAT-Mathematical regressions for Black and White students are shown in Figure 13. The slope changed in the top half of the score range for Black students, but it changed in the same way for White students, too. Furthermore, the GPA of Black students was generally overpredicted.

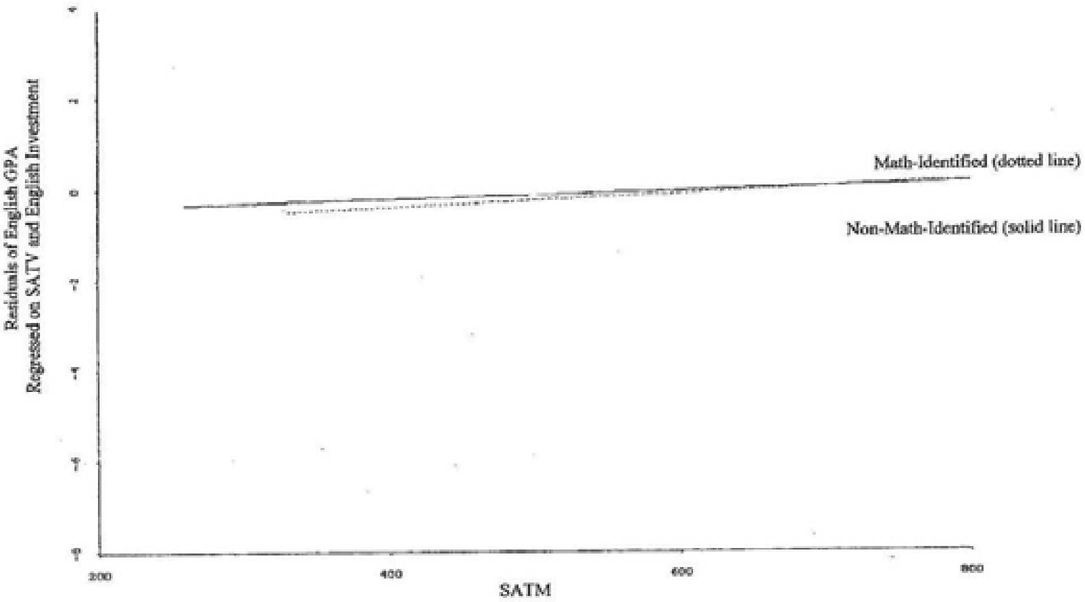

Figure 12. **SAT-Mathematical versus English GPA (residualized) for men.**

Note: From “Testing Stereotype Threat Theory Predictions from Math-Identified and Non-Math-Identified Students by Gender,” by M. J. Cullen, S. D. Waters, and P. R. Sackett, 2006, *Human Performance, 19,* p. 436. Copyright 2006 by Lawrence Erlbaum Associates. Reprinted with permission.

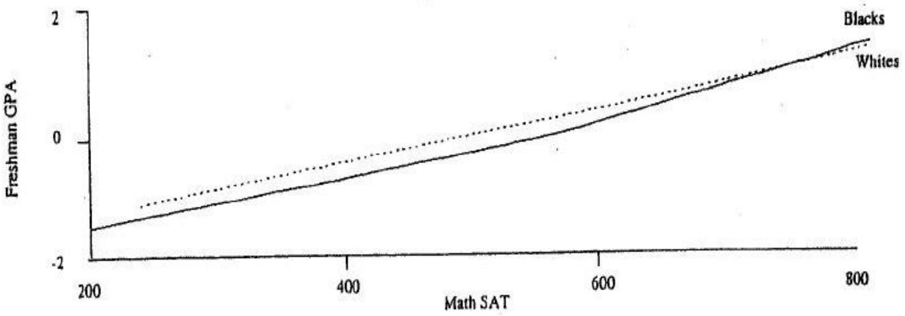

Figure 13. **SAT-Mathematical versus overall GPA by ethnicity.**
Note: From "Using SAT-Grade and Ability-Job Performance Relationships to Test Predictions Derived from Stereotype Threat Theory," by M. J. Cullen, C. M. Hardison, and P. R. Sackett, 2004 *Journal of Applied Psychology, 89,* p. 228. Copyright 2004 by the American Psychological Association. Reprinted with permission.

The SAT-Verbal regressions for Black and White students are shown in Figure 14. Contrary to stereotype threat theory, the Black students' regression was linear throughout the score range, and GPA was consistently overpredicted not underpredicted.

Similar analyses were done for the ASVAB and Project A criteria for Black and White soldiers (Cullen et al., 2004). For the ASVAB, a score on the first unrotated factor was used as a measure of general ability. (Cullen et al. cautioned that these criteria may have been affected by stereotype threat for Black soldiers because the criteria were based, in part, on tests of job knowledge.)

The regressions for the core technical proficiency criterion for Black and White soldiers are shown in Figure 15. The slope changed at the top of the score range for Black soldiers, but it also changed in the same way for White soldiers. Furthermore, the criterion performance was consistently overpredicted for Black soldiers.

The regressions for the general soldiering proficiency criterion for Black and White soldiers are shown in Figure 16. The regression was linear throughout the score range for Black soldiers, and their criterion performance was consistently overpredicted.

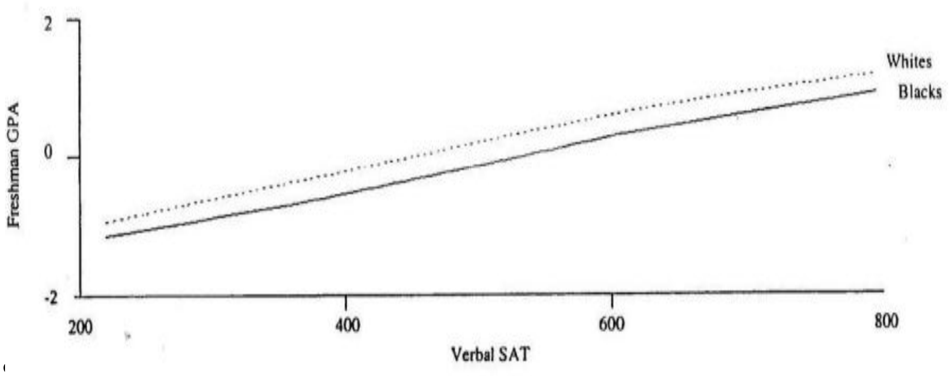

***Figure 14.* SAT-Verbal versus overall GPA by ethnicity.**

Note: From “Using SAT-Grade and Ability-Job Performance Relationships to Test Predictions Derived from Stereotype Threat Theory,” by M. J. Cullen, C. M. Hardison, and P. R. Sackett, 2004, *Journal of Applied Psychology, 89*, p. 228. Copyright 2004 by the American Psychological Association. Reprinted with permission.

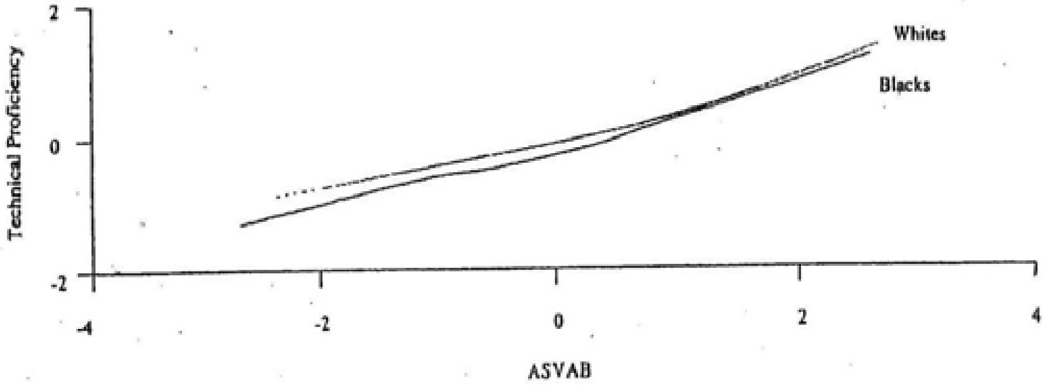

***Figure 15.* ASVAB versus core technical proficiency by ethnicity.**

Note: From “Using SAT-Grade and Ability-Job Performance Relationships to Test Predictions Derived From Stereotype Threat Theory,” by M. J. Cullen, C. M. Hardison, and P. R. Sackett, 2004, *Journal of Applied Psychology, 89*, p. 227. Copyright 2004 by the American Psychological Association. Reprinted with permission.

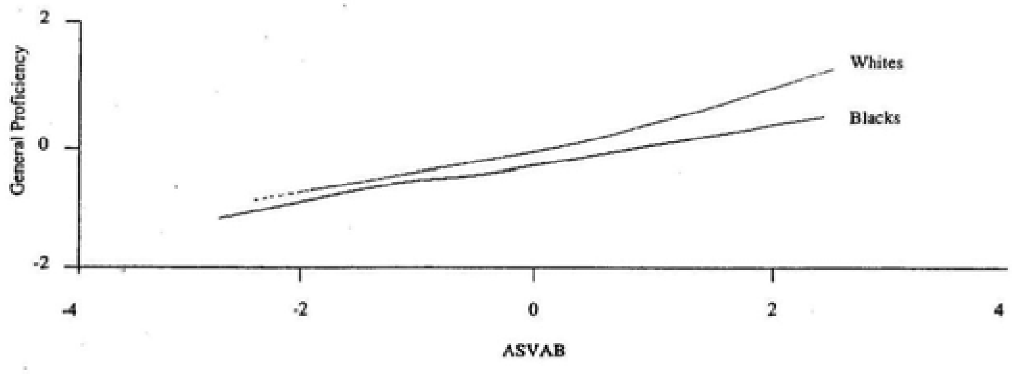

Figure 16. **ASVAB versus general soldiering proficiency by ethnicity.**
Note: From "Using SAT-Grade and Ability-Job Performance Relationships to Test Predictions Derived From Stereotype Threat Theory," by M. J. Cullen, C. M. Hardison, and P. R. Sackett, 2004, *Journal of Applied Psychology, 89,* p. 227. Copyright 2004 by the American Psychological Association. Reprinted with permission.

In brief, Cullen et al. (2004, 2006) found no evidence of stereotype threat for women and Black test takers in the regressions of SAT and ASVAB scores against academic and occupational performance criteria. These findings, like our own, bear on the generalizability of stereotype threat in the real world. They also suggest that the predictive validity of operational tests is unaffected by this phenomenon.

Good et al. (2003) studied 138 seventh grade students in a computer skills class. Most students were Hispanic or Black and were from low-income families. Students were assigned to a college-student mentor for the school year. Each mentor conveyed one of four different educational messages: intelligence is expandable and the brain forms new connections during the lifetime (the incremental condition), everyone initially experiences academic difficulties in the transition to junior high school (the attribution condition), a combination of the two messages—intelligence is expandable and the transition to junior high school is temporarily difficult (the combined condition), or drugs are dangerous (the control condition). At the end of the school year, students took the Texas Assessment of Academic Skills (TAAS), a statewide test of minimum competency then in use to determine whether students were promoted. The test's Mathematics and Reading scores were analyzed. The means for girls and boys on the Mathematics test are shown in Figure 17. Compared to the control (antidrug) condition, girls' means were significantly higher, statistically and practically, in all of the other conditions. Boys' means were only significantly higher in the incremental condition.

The means on the Reading test for the total sample, pooling girls and boys, are shown in Figure 18. Compared to the control condition, the means were significantly higher in the incremental and attribution conditions.

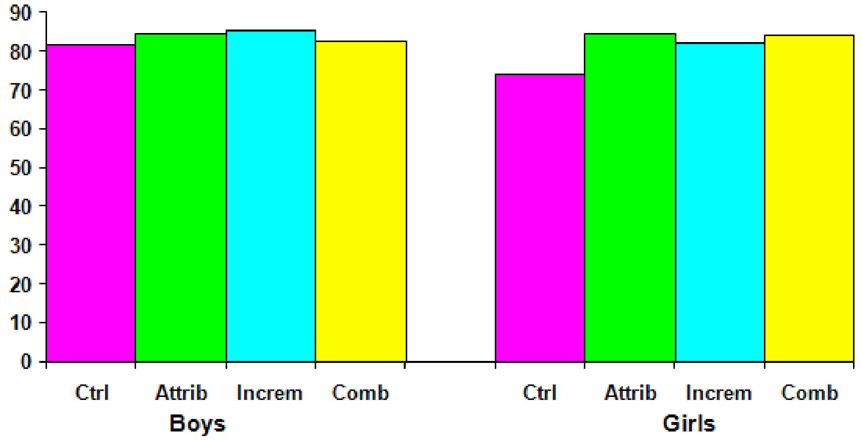

Figure 17. **TAAS Mathematics score. (Data from Good, Aronson, & Inzlicht, 2003).**
Note: Ctrl = control; Attirb = attribution; Increm = incremental; Comb = combined.

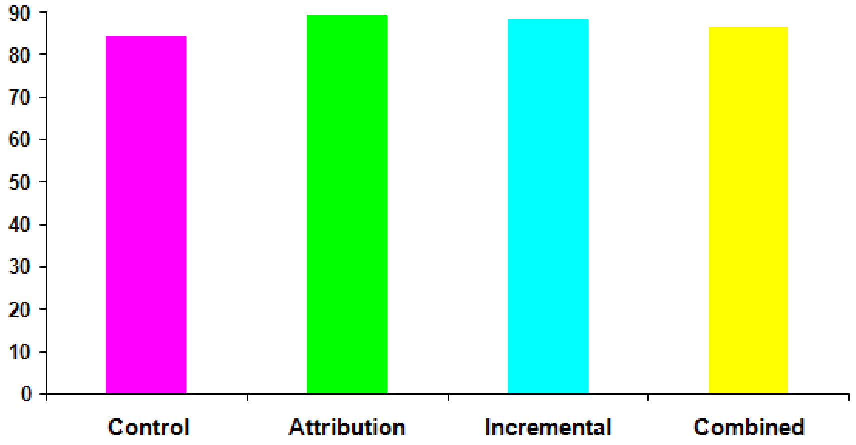

Figure 18. **TAAS Reading score. (Data from Good, Aronson, & Inzlicht, 2003).**

In sum, the mentoring interventions in the Good et al. (2003) study did affect the level of test performance, consistent with stereotype threat theory. These findings suggest that stereotype threat may be present on operational tests, unlike the results of our own experiments and those of the Cullen et al. (2004, 2006) studies. Important differences between the Good et al. study and the others may explain this divergence. On the one hand, the Good et al. manipulation was a major one (individually mentoring students over a school year) compared to the minor one in our own experiments (simply inquiring about test takers' ethnicity and gender). On the other hand, the Good et al. sample was small and atypical in contrast to the very large and diverse samples in the other studies.

This handful of investigations comprises the sum of the published research to date on stereotype threat with operational tests in high-stakes settings. This dearth is not surprising in view of the difficulty of studying stereotype threat in the real world. As Cullen et al. (2004) pointed out, pragmatic and ethical concerns preclude the powerful manipulations, often produced by deception that can be carried out in the laboratory. It isn't possible or proper to tell students taking the SAT, "This isn't really a test," or to tell those taking the AP Calculus Examination, "There are no gender differences on this test."

It is remarkable that all of the studies, with the important exception of Good et al. (2003), obtained negative results in contrast with the almost uniformly positive effects observed in the laboratory (see the reviews by Walton & Cohen, 2003; Wheeler & Petty, 2001). The most plausible explanation stems from differences in the settings. Motivation to perform well was probably heightened in the field studies because the students were taking high-stakes, operational tests, with real-life consequences. And this motivation overrode any harmful effects of stereotype threat that may otherwise be found in the laboratory.

Another explanation deserves mention. It was offered by Steele, Spencer, and Aronson (2002) to account for the negative results in our own studies: The level of stereotype threat associated with operational tests in real-world settings is so high that manipulations to lower or raise it are ineffective. Hence, Wicherts, Dolan, and Hessen (2005) questioned the internal validity of real-world experiments on stereotype threat. The same explanation is also presumably relevant to the Cullen et al. (2006) correlational studies that focus on the performance of subsets of test takers: The level of threat is so elevated that there are no appreciable differences for the subsets. But, in fact, this is all speculation, for nothing is actually known about the ambient level of stereotype threat on operational tests. And a survey of students who took the GMAT suggests that stereotype threat may not be as pervasive as supposed. The students were asked immediately after the test about how other people evaluated the test takers' verbal and quantitative ability.

The percentage reporting that their ability was underestimated did not differ appreciably for the minority and White students, or for women and men, and all of the percentages fell below 20% (B. Bridgeman, personal communication, July 7, 1999). Furthermore, the notion that stereotype threat on operational tests is immovably high is contradicted by the dramatic effects in the Good et al. (2003) study.

The general absence of effects for stereotype threat on operational tests in these studies is congruent with the absence of test bias against minority group members and women observed in numerous investigations over the years (Linn, 1992; Schmitt, 1989). (The flip side is that the striking effects in laboratory studies of stereotype threat are inconsistent with this test bias literature, as Sackett, Schmitt, Ellingson, and Kabin, 2001, pointed out.) Most of this literature consists of studies of predictive bias in regression analyses, similar to the Cullen et al. (2004, 2006) work The previous predictive bias studies, in common with the Cullen et al. (2004) research, found overprediction not underprediction for Black test takers. However, the criterion problem is particularly knotty when studies of predictive bias are used to draw inferences about the operation of stereotype threat. The special difficulty is that stereotype threat may affect the criterion as well as the predictor test. Precisely for that reason, the Cullen et al. (2004, 2006) studies of gender used English GPA as the criterion because it is presumably unaffected by stereotype threat for women. And Cullen et al. (2004) also cautioned about the uncertain validity of GPA and Project A criteria in their studies of Black test takers. This difficulty may extend beyond test criteria to nontest criteria. For instance, Steele (1997) argued that the underperformance of Black students in college is attributable to stereotype threat producing disidentification with the academic domain. It may be difficult to identify academic, occupational, and other important criteria that are unaffected by stereotype threat for Blacks and other groups that are targets of negative stereotypes in many domains.

Wicherts et al. (2005) persuasively argued that investigations of factorial invariance of tests across subgroups presumed to be affected or unaffected by stereotype threat are ideally suited to assess measurement bias produced by this phenomenon. There appears to be no published research of this kind for operational tests in high-stakes settings. However, two such studies appear in technical reports, both with negative results. Rock and Werts (1979), in an ETS and College Board report, found factorial invariance in the SAT across six ethnic groups. And

Rock, Werts, and Grandy (1982), in an ETS and GRE report, found this invariance in the GRE Aptitude Test (now the GRE General Test) across Black and White women and men.

The Good et al. (2003) study exemplifies another approach worth pursuing in investigating stereotype threat on operational tests: inoculating test takers beforehand to fend off its possible effects. This can be readily done as part of test preparation activities, delivered in written material or in courses. One such intervention, modeled after Good et al., is to advise test takers that cognitive tests are simply measures of "developed ability" or achievement, not assessments of innate ability. This point is often mentioned in the material that examinees receive from testing organizations, but it is not at all emphasized. Another intervention, successfully used in the laboratory by Johns, Schmader, and Martens (2005), is to inform examinees about stereotype threat and the anxiety it may cause in taking tests. Such interventions on operational tests do have an important and inevitable drawback. Direct checks on test takers' level of stereotype threat after such interventions are precluded. This makes it impossible to know whether any resulting changes in test performance are attributable to stereotype threat or to something else, such as test anxiety.

Modifications in the tests themselves and in the test administration procedures also need to be investigated, though they are severely circumscribed by practical considerations. Our own research on the impact of inquiring about ethnicity and gender is a case in point. Another of our efforts was a laboratory experiment (Stricker & Bejar, 2004). It was motivated by a Spencer et al. (1999, Study 1) finding that stereotype threat for women on a quantitative ability test was minimized when an easy test (items from the GRE General Test) was used instead of a hard test (items from the GRE Mathematics Test). This result supports a key hypothesis in stereotype threat theory that the threat is greatest for tasks that are at the limit of the test taker's ability. Women and men did not differ in their performance on the easy test, but they did differ on the hard test. (See Figure 19.) We modified the computer-adaptive version of the GRE General Test to produce an easier test that would yield comparable scores to the standard version. (The difficulty of the items on the Verbal and Quantitative sections of the easier test was reduced by one standard deviation; the Analytical section was not used.) The easier version or the standard version was administered to 343 Black and White women and men. They were college seniors planning to attend graduate school or first-year graduate students. The means for the GRE-Verbal and -Quantitative scores are shown in Figures 20 and 21. None of the differences between the easier and standard tests was statistically significant for any subgroup. Thus, this study failed to replicate the Spencer et al. results, suggesting that their findings may not be robust. Our outcome casts doubt on the likelihood that lowering difficulty can reduce stereotype threat on this test when it is used operationally.

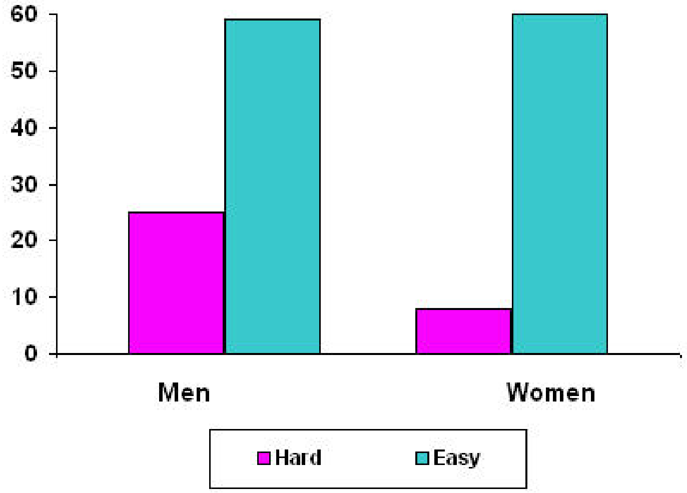

***Figure 19.* Mathematical score: easy (items from GRE General Test) and hard (items from GRE Mathematics Test). (Data from Spencer, Steele, & Quinn, 1999, Study 1)**

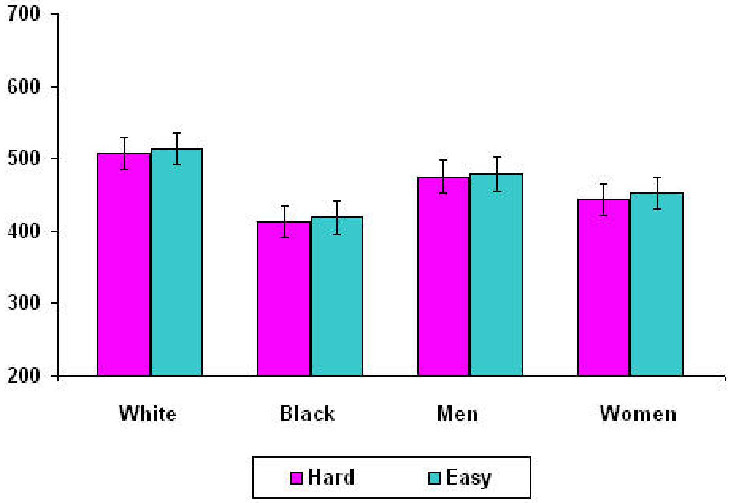

***Figure 20.* GRE Verbal score. (Data from Stricker & Bejar, 2004)**

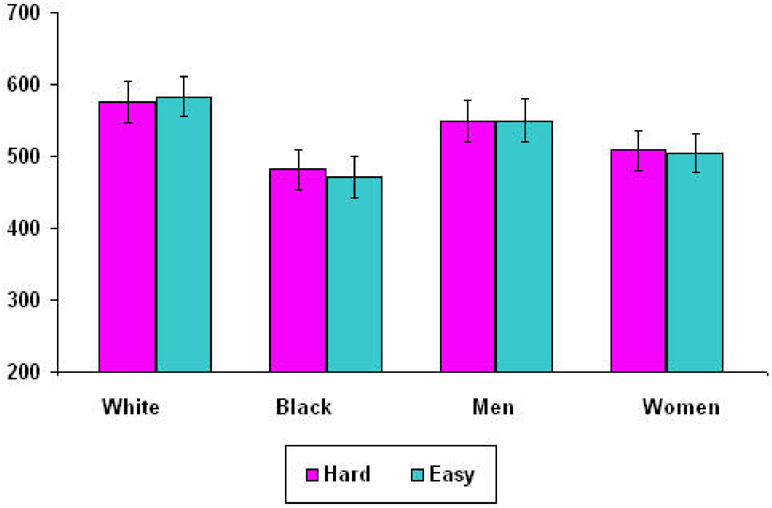

***Figure 21.* GRE Quantitative score. (Data from Stricker & Bejar, 2004)**

Research on stereotype threat needs to attend to individual differences in test takers' proneness to it. Several potentially important moderator variables have already been uncovered in laboratory experiments. A major moderator variable, identification with the domain, was already investigated by Cullen et al. (2006) in their field study. Others promising moderator variables include identification with the stereotyped group (Schmader, 2001), acceptance of stereotypes about one's group (Schmader, Johns, & Barquissau, 2004), and stigma consciousness (Brown & Pinel, 2003).

The bottom line is that it is premature to dismiss concerns about stereotype threat on the basis of the findings to date with field studies, given the limitations of such investigations, the weight of laboratory evidence, and the potential importance of this phenomenon as a threat to test validity. Research on the impact of stereotype threat on operational tests and on means of ameliorating any effects it may have continues to be critically important. In sharp contrast to the outpouring of research by social psychologists, though, this topic has largely been neglected by the testing community, with the notable exception of some industrial/organizational (I/O) psychologists. The Cullen et al. (2004, 2006) studies—published in I/O journals—have been discussed, and a special issue of *Human Performance* in 2003 was devoted to stereotype threat. But, in 2006, for example, not a single article on stereotype threat appeared in the five major journals in our field: *Applied Psychological Measurement, Educational and Psychological*

Measurement, Journal of Educational Measurement, Applied Measurement in Education, and *Educational Assessment.* Stereotype threat is a challenge that the testing community must meet.

From a broader perspective, the stereotype threat phenomenon is a welcome but belated wake-up call to us about the importance of attending to the social situation surrounding testing. Test anxiety is a long standing and major field of inquiry (Zeidner, 1998), but other critical areas are neglected. They include test-taking motivation; test takers' attitudes about tests; and the consequences of these variables for test performance and for the acceptance of tests by test takers, institutions, and the public. Again, the exception is the work on these topics by a small band of I/O psychologists, notably David Chan, Robert Ployhart, Ann Marie Ryan, and Neal Schmitt (e.g., see the review by Ryan & Ployhart, 2000). This matter is especially important for educational testing, given the explosion of statewide minimum competency tests spawned by No Child Left Behind and the ongoing importance of National Assessment of Educational Progress and of international comparisons of student achievement. My modest proposal is that the testing community shift a portion of the resources now devoted to perfecting psychometrics, refining test content, and improving test delivery systems to greater efforts to understand and deal with the social context of testing.

References

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

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

Cullen, M. J., Waters, S. D., & Sackett, P. (2006). Testing stereotype threat theory predictions for math-identified and non-math-identified students by gender. *Human Performance, 19*, 421–440.

Good, C., Aronson, J., & Inzlicht, M. (2003). Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Applied Developmental Psychology, 24,* 645–662.

Johns, M., Schmader, T., & Martens, A. (2005). Knowing is half the battle: Teaching stereotype threat as a means of improving women's math performance. *Psychological Science, 16,* 175–179.

Linn, R. (1992). Ability testing: Individual differences, prediction, and differential prediction. In A. K. Wigdor & W. R. Garner (Eds.), *Ability testing: Uses, consequences, and controversies: Part 2. Documentation section* (pp. 335–388). Washington, DC: National Academy Press.

Rock, D. A., & Werts, C. E. (1979). *Construct validity of the SAT across populations—An empirical confirmatory study* (ETS Research Rep. No. RR 79-2; College Board RDR 78-79, No. 5). Princeton, NJ: ETS.

Rock, D. A., Werts, C., & Grandy, J. (1982). *Construct validity of the GRE Aptitude Test across populations—An empirical confirmatory study* (ETS Research Rep. No. RR 81-57; GRE Board Professional Rep. No. 78-01P). Princeton, NJ: ETS.

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

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

Schmader, T. (2001). Gender identification moderates stereotype threat effects on women's math performance. *Journal of Experimental Social Psychology, 38,* 194–201.

Schmader, T., Johns, M., & Barquissau, M. (2004). The costs of accepting gender differences: The role of stereotype endorsement in women's experience in the math domain. *Sex Roles, 50*, 835–850.

Schmitt, N. (1989). Fairness in employment selection. In M. Smith & I. T. Robertson (Eds.), *Advances in selection and assessment* (pp. 134–153). Chichester, England: Wiley.

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

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

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

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and identity threat. *Advances in Experimental Social Psychology, 34*, 379–440.

Stricker, L. J., & Bejar, I. I. (2004). Test difficulty and stereotype threat on the GRE General Test. *Journal of Applied Social Psychology, 34*, 563–597.

Stricker, L. J., & Ward, W. C. (2004). Stereotype threat, inquiring about test taker's ethnicity and gender, and standardized test performance. *Journal of Applied Social Psychology, 34*, 665–693.

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

Wheeler, S. C., & Petty, R. E. (2001). The effects of stereotype activation on behavior: A review of possible mechanisms. *Psychological Bulletin, 127,* 797–826.

Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology, 89,* 696–716.

Zeidner, M. (1998). *Test anxiety: The state of the art.* New York: Plenum.