Stereotype Threat, Inquiring About Test Takers' Ethnicity and Gender, and Standardized Test Performance[1]

LAWRENCE J. STRICKER[2] AND WILLIAM C. WARD
Educational Testing Service
Princeton, New Jersey

Steele and Aronson (1995) found that the performance of Black research participants on ability test items portrayed as a problem-solving task, in laboratory experiments, was affected adversely when they were asked about their ethnicity. This outcome was attributed to stereotype threat: Performance was disrupted by participants' concerns about fulfilling the negative stereotype concerning Black people's intellectual ability. The present field experiments extended that research to other ethnic groups and to males and females taking operational tests. The experiments evaluated the effects of inquiring about ethnicity and gender on the performance of students taking 2 standardized tests—the Advanced Placement Calculus AB Examination, and the Computerized Placement Tests—in actual test administrations. This inquiry did not have any effects on the test performance of Black, female, or other subgroups of students that were both statistically and practically significant.

Research by Steele and Aronson (1995, Study 4) found that the performance of Black research participants at Stanford University on difficult verbal ability items from the Graduate Record Examinations (GRE) General Test (Briel, O'Neill, & Scheuneman, 1993) that were portrayed as a verbal problem-solving task was adversely affected when they were asked about their ethnicity immediately prior to working on the items. The performance of White participants was

[1]This research was supported in part by the College Entrance Examination Board. For Study 1, the authors thank Walter B. MacDonald for encouraging the research; Amy C. Cellini, Rick Morgan, and Gita Z. Wilder for advising on the experimental design; Amy C. Cellini for coordinating the data collection; Lorraine Emans, Tammy Haston, Kristine A. Nickerson, and Margaret L. Redman for recruiting AP classes; Geraldine Kovar, Behroz T. Manechshana, and Rick Morgan for providing AP test data; Donald A. Rock for advising on the statistical analysis; Thomas J. Jirele and Ting Lu for computer analysis; and Walter Emmerich for advising on the interpretation of the findings. For Study 2, the authors thank Central Piedmont Community College for cooperating in the study; David A. Rhoden for coordinating the data collection; Margaret L. Redman for preparing the data for analysis; and Laura M. Jenkins and Xuefei Hui for computer analysis. The authors also thank Rick Morgan, Claude M. Steele, and Gita Z. Wilder for reviewing a draft of the article. Any opinions expressed in this article are those of the authors and not necessarily of Educational Testing Service.

[2]Correspondence concerning this article should be addressed to Lawrence J. Stricker, Educational Testing Service, Rosedale Road, Princeton, NJ 08541. E-mail: lstricker@ets.org

Journal of Applied Social Psychology, 2004, **34**, 4, pp. 665-693.

unaffected. Black participants who were asked about their ethnicity answered fewer items correctly, answered correctly a smaller percentage of the items that they attempted, attempted fewer items, and spent more time working on the items than did Black participants who were not asked. This effect was not only replicable, but also substantial (e.g., the *M* difference for the number of items answered correctly by Black participants in the two conditions in the replication represented a *d* of 1.05; Cohen, 1988). The purpose of the experiments was described to the participants as nondiagnostic—to understand the psychological factors involved in solving verbal problems. Individuals' ability was not being evaluated, though they would receive feedback about their performance. Steele and Aronson attributed these results to stereotype threat: Asking about ethnicity primes Black participants' concerns about fulfilling the negative racial stereotype regarding their intellectual ability, thereby disrupting their performance.

Other research (see review by Wheeler & Petty, 2001) has elicited stereotype threat in a variety of ways for Black participants taking verbal ability tests (Blascovich, Spencer, Quinn, & Steele, 2001; Steele & Aronson, 1995), for women and young girls taking quantitative ability tests (Ambady, Shih, Kim, & Pittinsky, 2001; Brown & Josephs, 1999; Inzlicht & Ben-Zeev, 2000; Oswald & Harvey, 2000–2001; Quinn & Spencer, 2001; Shih, Pittinsky, & Ambady, 1999; Spencer, Steele, & Quinn, 1999; Walsh, Hickey, & Duffy, 1999),[3] and for working-class participants taking a verbal ability test (Croizet & Claire, 1998). These findings demonstrate that stereotype threat has a widespread effect on the test performance of groups that are targets of negative stereotypes about their intellectual ability. Indeed, Steele, his coworkers, and others (Aronson et al., 1999; Aronson, Quinn, & Spencer, 1998; Brown & Josephs, 1999; Croizet & Claire, 1998; Oswald & Harvey, 2000–2001; Quinn & Spencer, 2001, Spencer et al., 1999; Steele, 1997; Steele & Aronson, 1995) have argued that this phenomenon may help to account for the deficits on standardized tests and in academic performance in school that are observed for such groups.

Steele and Aronson's (1995) research on inquiring about ethnicity has obvious parallels with test administration procedures for widely used standardized tests that are employed in educational settings for admissions, course credit, course placement, and other purposes; and that require test takers to answer questions about their ethnicity and gender immediately before they take the tests. These parallels raise the real possibility that this practice may affect the performance of Black and female test takers on these tests.

Two highly relevant tests of this kind are the Advanced Placement (AP) Examinations (College Board & Educational Testing Service, 1995c) and the Computerized Placement Tests (CPTs; College Board, 1995; Ward, 1988). AP

[3]Cheryan and Bodenhausen (2000) is an exception to other studies that found the stereotype concerning women adversely affected their performance on a quantitative ability test.

tests are course examinations taken by high school students seeking college credit or advanced standing in college after completing a college-level course in high school. The CPTs are a battery of basic skills tests (reading, writing, and mathematics) taken for course placement by community college students, and, to a lesser extent, by 4-year college students. The two tests are of special interest for several reasons:

1. Many students take AP tests and CPTs. In 1995, for instance, 504,823 examinees took AP tests (College Board & Educational Testing Service, 1995a). And in 1996, approximately 457,000 examinees took the CPTs (M. Rosenthal, personal communication, July 13, 1998).

2. Often substantial mean differences exist in the performance of White and Black students, and of males and females on some of the 29 AP tests and some of the 4 CPTs. For example, in 1995, the mean AP grades of White and Black students were 3.15 and 2.18 ($d = .87$) for the AP English Literature and Composition Examination; and the mean AP grades of boys and girls on the AP Chemistry Examination were 2.99 and 2.55 ($d = .33$; Educational Testing Service, 1995). And for 28 community colleges that provided data for CPT examinees who took the current version of the CPT (Version 5.2) in 1997 (Educational Testing Service, 1998), the means for Reading Comprehension for White and Black students were 81.61 and 64.64 ($d = .87$), respectively; and the means for Arithmetic for men and women were 60.66 and 53.46 ($d = .28$), respectively.

3. These tests of academic achievement are pertinent to the stereotype about the intellectual ability of Black people and women.

Despite the similarities between Steele and Aronson's (1995) research and the AP and CPTs test administration procedure, the two situations may also differ in significant respects:

1. Steele and Aronson's participants were taking the tests for research purposes, whereas AP and CPTs students take the tests for important personal reasons, and hence may be more motivated to do well on them.

2. The experimental task in Steele and Aronson's research was portrayed to the participants as innocuous problem solving,

whereas AP and CPTs test takers are aware that they are taking tests that reflect their mastery of specific course content or important academic skills. Steele and Aronson (Study 3) also found that stereotype threat was heightened when the experimental task was described as diagnostic of participants' intellectual ability. Thus, inquiring about ethnicity (and gender) may have a limited impact on AP tests and CPTs insofar as stereotype threat is already elevated by test takers' perceptions of these tests as diagnostic. In fact, Steele found in an unpublished pilot study that inquiring about ethnicity did not affect Black participants' performance when the task was described as diagnostic of their ability (C. M. Steele, personal communication, May 21, 1997), in contrast to the substantial effect of inquiring when the task was described as nondiagnostic. On the other hand, Croizet and Claire (1998) found that inquiring about socioeconomic status (parents' occupation and education) had no effect, regardless of whether the task was described as diagnostic or nondiagnostic.

3. Research by Aronson et al. (1999) suggests that stereotype threat only affects test takers who identify with the subject matter being tested. The Stanford University students in Steele and Aronson's research presumably have this identification. Although AP test takers also may be very involved with the academic material being tested, CPTs test takers are probably less involved. However, regardless of their level of identification, both groups may be ego-involved in the outcome of the tests—obtaining college credit for AP students, or avoiding remedial courses for CPTs students—making the outcome equally susceptible to stereotype threat, as Aronson et al. speculated.

4. Work by Spencer et al. (1999) implies that an important element in the operation of stereotype threat is test takers' perceptions of the items as difficult, at the limits of their ability. Data on these perceptions are unavailable from Steele and Aronson's research, and it is unclear how AP and CPT test takers perceive the tests.

Accordingly, the aim of the present research is to extend Steele and Aronson's (1995) work on inquiring about the ethnicity of research participants in laboratory experiments to inquiring about both the ethnicity and the gender of students taking operational tests (AP tests and CPTs). This research thereby assesses the generalizability of the laboratory findings to real life and evaluates the practical consequences of routine inquiries about ethnicity and gender in standardized testing. The hypotheses are as follows:

Hypothesis 1. Asking about ethnicity and gender will depress the performance of Black students on all of the tests.

Hypothesis 2. Asking about ethnicity and gender will depress the performance of females on the quantitative tests.

Hypothesis 3. Asking about ethnicity and gender will have no effect on the performance of White students and males on any of the tests.

No hypotheses are advanced about the performance of other ethnic groups, either because there were few members of these groups in the test-taking populations (Hispanics and American Indians) or because these negative stereotypes about the group's intellectual ability are not prevalent (Asians).[4]

The two studies reported here are alike in altering the standard test administration for some students by eliminating the usual questions about ethnicity and gender (experimental group), and contrasting their performance with the performance of comparable students who were asked these questions in the course of the standard test administration (control group). The two studies differ primarily in the test employed (its content and purposes) and in the test-taking population.

Study 1 used the AP Calculus AB Examination (College Board, 1994). This particular AP test was chosen because (a) it is taken by relatively large numbers of White and Black students, and boys and girls (e.g., 67,863 White and 4,020 Black students, and 52,465 boys and 47,275 girls took the test in 1995; College Board & Educational Testing Service, 1995a); (b) substantial mean differences exist in the test performance of White and Black students, as well as males and females (e.g., in 1995, for White and Black examinees, AP grades = 2.82 vs. 1.87, $d = .73$; for boys and girls, AP grades = 2.93 vs. 2.62, $d = .24$; College Board & Educational Testing Service, 1995a, 1995b); and (c) the subject matter is pertinent to the stereotype about females' quantitative ability as well as the stereotype about Black people's ability in general. Study 2 used the CPTs.

Study 1

This study modified the test administration of the AP Calculus AB Examination for a random sample of AP classes across the country by masking background questions on the standard answer sheet (the experimental group). The test performance of test takers in these classes was then compared with the

[4]Shih et al. (1999), Cheryan and Bodenhausen (2000), and Ambady et al.'s (2001) studies, reported after this research was completed, provide conflicting suggestions about whether stereotype threat would improve or degrade the performance of Asian students on quantitative ability tests, in view of the positive stereotype in the United States concerning Asians' quantitative ability.

performance of test takers in a random sample of classes that used the standard answer sheet (the control group).[5]

Method

Sample

The sampling had four objectives:

1. To obtain an appreciable sample of Black AP Calculus AB Examination test takers efficiently, given that Black students are enrolled in only a fraction of AP Calculus AB courses (19.9% of 8,222 classes had Black test takers in 1995; College Board & Educational Testing Service, 1995b; B. T. Maneckshana, personal communication, March 21, 1996), by restricting the classes in the sample to those that previously had Black test takers.

2. To secure test takers who were first asked about their ethnicity and gender in the AP test administration when they filled out the answer sheet for the AP Calculus AB Examination immediately before taking the test. This objective was accomplished by excluding:
   a. Classes that were provided with an earlier preadministration session before the test was taken in which students completed background information on the answer sheet, including answering the ethnicity and gender questions. Some schools provide these sessions to expedite the test administration.
   b. Students who took a previously administered AP test in the same 2-week testing period.

3. To obtain test takers who resided in the United States to ensure that they had been exposed to the prevalent negative stereotypes in this country about the intellectual ability of Black people and females. This objective was accomplished by excluding classes in other countries.

4. To ensure that the experimental and control groups were comparable by stratifying the classes in the sample on relevant variables: size, ethnic composition, and previous AP Calculus AB Examination performance.

[5]AP tests are administered to intact classes, precluding the use of different answer sheets with test takers in the same class.

AP Calculus AB classes taking part in the May 1996 test administration were drawn from the 1,639 classes with one or more Black students taking the AP test in 1995. For the experimental group, a stratified random sample of 181 classes (11% of the total) was drawn, stratified on 1995 data (B. T. Maneckshana, personal communication, March 21, 1996) for size (15 students or fewer, 16 or more), percentage of Black test takers (11% or less, 12% or more), and percentage of AP grades of 3 or higher on the AP Calculus AB Examination (57% or less, 58% or more; grades range from 1 to 5, and a grade of 3 or higher is considered passing; College Board & Educational Testing Service, 1995c). Of the 181 classes, 82 actually participated in the study. Most of the others did not because they were unwilling to eliminate their preadministration sessions for the test. Of the 82 classes, 77 were used in the analysis; the 5 others were excluded because all of their test takers had taken one or more of the nine previously administered AP tests in the same testing period.

For the control group, a stratified random sample of 181 classes (plus an oversample of 36 classes) was drawn. A total of 133 classes were eligible to participate in the study; most of the others were ineligible because they used a preadministration session. Of the 133 classes, 14 were excluded because all of their test takers had taken a previous AP test during the same testing period. Of the remaining 119 classes, 77 were used in the analysis. They were randomly selected from the same strata and with the same frequency as the 77 classes in the experimental group.

The classes in the experimental and control groups were similar in total number of test takers (M = 18.96 and 20.95), ethnicity (percentage of White test takers, M = 62.25 and 61.06; percentage of Black test takers, M = 10.15 and 11.21; percentage of Asian test takers, M = 13.58 and 12.17; percentage of test takers of other ethnicity, M = 7.50 and 8.37; and percentage of test takers with omitted ethnicity, M = 6.52 and 7.19),[6] and gender (percentage of boys, M = 50.72 and 50.23).[7] The two groups were also similar in their performance on the AP Calculus AB Examination in 1995 (M = 49.61 and 49.99, for percentage of test takers with AP grades of 3 or higher).

The test takers in the analysis consisted of students from the classes in the experimental and control groups who had not taken a previously administered AP test in the same testing period. The experimental group consisted of 755 students (429 White, 52 Black, 151 Asian, 61 of other ethnicity, and 62 with omitted ethnicity; 407 boys, 348 girls), while the control group consisted of 897 students

[6]Other ethnic groups were pooled in the study because of their small size. The largest of these ethnic groups (Hispanic) accounted for a mean percentage of 4.27 and 4.07 of the classes (or 64 and 65 test takers) in the experimental and control groups, respectively. (The number of Hispanic students in the analysis was 35 in the experimental group and 27 in the control group.)

[7]Data for students who were enrolled in the classes but did not take the AP test were not available.

(555 White, 70 Black, 152 Asian, 54 of other ethnicity, and 66 with omitted ethnicity; 515 boys, 382 girls).

Procedure

Experimental group. The AP Calculus AB classes in the experimental group were recruited by telephoning the AP coordinators, who are high school staff members responsible for administering the AP tests, and asking their schools to participate. The AP coordinators were told that the value of modifying AP test administration procedures was being studied, specifically how and when students fill out background information on the answer sheet.[8] The AP coordinators were told that the study involved changes in the answer sheets, and they were asked (a) not to give the AP Calculus AB Examination in the same room as the AP Calculus BC Examination because of the altered test administration procedures for the former; and (b) not to offer a preadministration session for students for whom the AP Calculus AB Examination was their first AP test in the testing period. AP coordinators who agreed to participate were sent a modified version of the general instructions for administering AP tests, revised to conform to the changes in the answer sheets, and a supply of special answer sheets.

At the test administration, the AP coordinators gave students instructions for taking the test, which included this modification:

> Some directions for this exam differ slightly from those for other AP exams being given this month because ETS is trying out changes in the answer sheet for this exam. You will be given a special answer sheet before the exam and the regular answer sheet after the exam. This is the only change in how the exam is given. It will not delay your grade report.

The first answer sheet, given to examinees before the test, consisted of Side 1 of the regular answer sheet, containing identifying information and space for answers to the test, plus Side 2 of the answer sheet with everything masked except space for answers to the test. The second answer sheet, given to students after the test, was a regular answer sheet. Students were asked to complete only the identifying information on Side 1 and all of the background questions on Side 2, including ethnicity, gender, and date of birth.[9] No other changes were made in the test administration.

[8]The origins of the study and the specific research questions being investigated were not described.

[9]The questions (in order) were address, Social Security number, gender, present grade level, date of birth, expected date of college entrance, enrollment in student search service, intention to apply for advanced standing in college, high school attended, and names of colleges to receive AP grades. A copy of the answer sheet appears in Stricker (1998).

After the test administration, the AP coordinators were telephoned to determine that they had followed the special test administration procedures and had not offered a preadministration session. Classes that did not comply were eliminated from the experimental group.

Control group. The eligibility of AP Calculus AB classes in the control group was determined by telephoning the AP coordinators after the test administration and asking if they had offered a preadministration session for examinees taking the AP Calculus AB Examination. Classes for which a preadministration session was offered were eliminated.

Measures

The AP Calculus AB Examination consists of 40 multiple-choice items in two separately timed sections (Part A has 25 items, calculators cannot be used, and the time limit is 50 min; Part B has 15 items, graphing calculators can be used, and the time limit is 40 min) and six free-response questions (graphing calculators can be used, and the time limit is 90 min). There is a penalty for guessing on the multiple-choice items.

Six scores were obtained:

1. The three scores for multiple-choice items used by Steele and Aronson (1995):
   a. Number Attempted (i.e., Number Correct and Number Wrong).
   b. Number Correct.
   c. Accuracy (i.e., Number Correct/[Number Correct and Number Wrong]).

2. A score for multiple-choice items routinely used in standard test analyses: Formula Score (Number Correct corrected for guessing).

3. Two special AP scores:
   a. Free-Response Section Score (total score on free-response questions).
   b. AP Grade (an equally weighted composite of the Formula Score and the Free-Response Section Score reported to students and school officials).

Ethnicity and gender were determined from AP files that included students' responses on the answer sheet for the test or on answer sheets for other AP tests taken subsequently during the testing period.

Analysis

Data were pooled across classes for students in the experimental (no inquiry) group and for students in the control (inquiry) group. All analyses used unweighted means because of the unequal *N*s in the cells. A series of 2 × 5 × 2 (Condition: Experimental vs. Control × Ethnicity: White, Black, Asian, Other, Omitted × Gender) factorial ANOVAs were carried out using the least squares method (Model I error term; Overall & Spiegel, 1969) to deal with unequal *N*s.[10]

Planned comparisons of simple effects of the experimental versus control group factor for each ethnic group (e.g., Black students in the experimental group vs. Black students in the control group) and each gender (e.g., girls in the experimental group vs. girls in the control group) were conducted (Howell, 1997). These focused tests of the same ethnic group or the same gender in the experimental and control groups are essential, given the specific hypotheses about Black and White students and females and males, and the need to compare the present findings with those of Steele and Aronson (1995). Steele and Aronson statistically adjusted for preexisting differences in ability in the subgroups, whereas the present analysis does not. Hence, differences between the experimental and control groups for the same ethnic group or same gender in this study can be compared directly with the corresponding differences in Steele and Aronson's research, but differences between ethnic groups or between genders cannot be compared. Post hoc multiple comparisons of ethnic groups means were made by Tukey's honestly significant difference (HSD) test.

Both statistical and practical significance were considered in evaluating the results. For statistical significance, an .05 alpha level was used in all analyses (.05 was the familywise alpha level for the planned comparisons of simple effects, using the Bonferroni procedure,[11] and for multiple comparisons with Tukey's test). For practical significance, a partial η (Cohen, 1973) of .10 in the ANOVAs and a *d* of .20 in the multiple comparisons were used. An η of .10 and a *d* of .20 represent Cohen's (1988) definition of a small effect size, accounting for 1% of the variance.[12] The assessment of practical significance is especially important because of the large sample size involved (Cohen, 1994). More relaxed levels of statistical and practical significance (α =.10, η = .05, *d* = .10) are reported in the tables and footnotes for comprehensiveness.

[10]Students with Asian ethnicity, other ethnicity, or omitted ethnicity were included in the analysis for completeness, even though no hypotheses were advanced about these groups.

[11]Separate familywise alpha levels were determined for the planned comparisons of the ethnic groups and of the genders.

[12]In a survey of 322 meta analyses in social psychology, about 70% of the 474 mean effect sizes were above that level (Richard, Bond, & Stokes, 2001).

Table 1

Summary of Overall ANOVAs of Scores on AP Calculus AB Examination, Study 1

Source	*df*	*F*					
		Number attempted	Number correct	Accuracy	Formula score	Free-response section score	AP grade
Experimental-Control (E-C)	1	1.46	0.00	0.21	0.04	0.32	0.14
Ethnicity	4	3.44[a]***	13.09[b]***	10.13[b]***	12.75[b]***	14.57[b]***	11.79[b]***
Gender	1	9.58[a]***	18.32[b]***	11.59[a]***	16.44[b]***	10.35[a]***	14.77[a]***
E-C × Ethnicity	4	0.97[a]	0.41	0.59	0.43	1.05[a]	0.37
E-C × Gender	1	1.59	4.05[a]**	6.77[a]***	5.60[a]**	2.32	4.16[a]**
Ethnicity × Gender	4	1.22[a]	1.39[a]	1.10[a]	1.30[a]	0.35	0.80
E-C × Ethnicity × Gender	4	0.44	2.22[a]*	2.80[a]**	2.44[a]**	1.62[a]	2.21[a]*

Note. *df* for Error = 1,632. *MSE* = 31.79 for Number Attempted, 50.59 for Number Correct, 346.09 for Accuracy, 69.03 for Formula Score, 104.49 for Free-Response Section Score, and 1.57 for AP Grade.

[a]$\eta > .05$. [b]$\eta > .10$.

*$p < .10$. **$p < .05$. ***$p < .01$.

Table 2

Summary of Simple Effects of Experimental Versus Control Group in ANOVAs of Scores on AP Calculus AB Examination, Study 1

	F					
Source	Number attempted	Number correct	Accuracy	Formula score	Free-response section score	AP grade
Ethnicity						
White	0.08	0.39	0.50	0.38	0.73	0.00
Black	0.04	0.70	1.36	0.87	0.42	0.32
Asian	4.35[a]	0.00	0.80	0.17	2.70	1.09
Other	0.24	0.14	0.38	0.10	0.30	0.25
Omitted	0.12	0.57	0.24	0.56	0.07	0.00
Gender						
Male	0.00	2.13	2.37	2.46	0.48	1.44
Female	2.95	1.93	4.53[a]*	3.16	2.11	2.81

Note. $df = 1$ for each simple effect. df for Error and the MSE for each analysis appear in Table 1.

[a] $\eta > .05$.

*$p < .10$ (familywise).

Results[13]

Overall ANOVAs of the six scores are summarized in Table 1, and the analyses of the simple effects of experimental versus control group are summarized in Table 2. The means appear in Table 3. Note that most of the scores were highly correlated. Correlations ranged from .14 to .95 ($N = 755$, $p < .01$) for the experimental group and from .12 to .96 ($N = 897$, $p < .01$) for the control group. The lowest correlation in both groups was for Accuracy versus Number Attempted, and the highest was for Accuracy versus Formula Score.

[13] In multiple comparisons of mean differences for the ethnic groups ($\alpha = .10$, $d = .10$), no additional differences for Number Attempted, Free-Response Section Score, and AP Grade were significant, both statistically and practically. Two additional differences for Number Correct, Accuracy, and Formula Score were significant: White students' mean was higher than the mean of students with other ethnicity and lower than Asian students' mean.

Number Attempted

None of the main effects (condition, ethnicity, and gender) or interactions in the ANOVA of this variable were significant, both statistically ($p < .05$) and practically ($\eta > .10$). In the multiple comparisons of mean differences for the ethnic groups, two of the differences were significant: Black students attempted fewer items than did White students and Asian students. None of the simple effects for ethnicity (e.g., Black students in the experimental group vs. Black students in the control group) or for gender (e.g., girls in the experimental group vs. girls in the control group) were significant.

Number Correct

The main effects for ethnicity and gender were significant. Girls underperformed boys. In the multiple comparisons, four of the mean differences were significant: Black students underperformed White students, Asian students, and students with omitted ethnicity; and Asian students outperformed students with other ethnicity. None of the simple effects for ethnicity or for gender were significant.

Accuracy

The main effect for ethnicity was significant. In the multiple comparisons, four of the mean differences were significant: The pattern of differences was the same as for Number Correct. None of the simple effects for ethnicity or for gender were significant.

Formula Score

The main effects for ethnicity and gender were significant. Girls underperformed boys. In the multiple comparisons, four of the mean differences were significant: The pattern of differences was the same as for Number Correct. None of the simple effects for ethnicity or for gender were significant.

Free-Response Section Score

The main effect for ethnicity was significant. In the multiple comparisons, six of the mean differences were significant. Black students underperformed White students, Asian students, and students with omitted ethnicity. White students, Asian students, and students with omitted ethnicity outperformed students with other ethnicity. None of the simple effects for ethnicity or for gender were significant.

Table 3

Mean Scores on AP Calculus AB Examination, Study 1

Group	White			Black		
	Boys	Girls	Total	Boys	Girls	Total
Number attempted (*SD* = 5.64)						
Experimental	33.68	31.82	32.75	31.67	30.00	30.83
Control	32.91	32.39	32.65	30.86	31.24	31.05
Number correct (*SD* = 7.11)						
Experimental	19.63	18.79	19.21	15.67	15.71	15.69
Control	19.81	18.04	18.92	17.69	11.49	14.59
Accuracy (*SD* = 18.60)						
Experimental	58.17	58.30	58.24	49.53	52.94	51.24
Control	59.77	55.00	57.38	55.99	38.49	47.24
Formula score (*SD* = 8.31)						
Experimental	16.12	15.54	15.83	11.67	12.14	11.90
Control	16.54	14.45	15.49	14.40	6.55	10.47
Free-response section score (*SD* = 10.22)						
Experimental	19.72	17.72	18.72	12.50	13.11	12.80
Control	20.88	17.71	19.29	15.07	8.10	11.58
AP grade (*SD* = 1.25)						
Experimental	2.99	2.76	2.87	2.21	2.21	2.21
Control	3.06	2.70	2.88	2.55	1.61	2.08

Note. *SD*s calculated from the *MSE*s in the ANOVAs.

AP Grade

The main effect for ethnicity was significant. In the multiple comparisons, five of the mean differences were significant. Black students underperformed White students, Asian students, and students with omitted ethnicity. White students and Asian students outperformed students with other ethnicity. None of the simple effects for ethnicity or for gender were significant.

Summary

Ethnic group differences in test performance were extensive, and gender differences were limited. However, the experimental manipulation of inquiring about

Asian			Other			Omitted			Total	
Boys	Girls	Total	Boys	Girls	Total	Boys	Girls	Total	Boys	Girls
32.64	32.34	32.49	34.14	30.18	32.16	32.59	31.88	32.24	32.94	31.25
33.98	33.73	33.85	33.67	31.71	32.69	33.21	31.97	32.59	32.92	32.21
20.48	20.31	20.40	19.86	14.09	16.97	17.81	18.96	18.39	18.69	17.57
21.17	19.54	20.35	18.58	16.38	17.48	21.18	17.52	19.35	19.69	16.59
61.63	62.34	61.98	58.65	45.30	51.97	55.07	59.00	57.04	56.61	55.58
62.08	58.05	60.06	55.80	52.48	54.14	63.17	54.16	58.67	59.36	51.63
17.44	17.30	17.37	16.29	10.07	13.18	14.11	15.73	14.92	15.13	14.16
17.97	15.99	16.98	14.80	12.55	13.68	18.17	13.90	16.04	16.38	12.69
20.15	19.45	19.80	17.18	11.94	14.56	18.38	19.72	19.05	17.59	16.39
18.83	16.89	17.86	15.88	15.38	15.63	20.67	16.48	18.58	18.26	14.91
3.05	3.03	3.04	2.86	2.03	2.44	2.76	3.04	2.90	2.77	2.61
3.00	2.77	2.89	2.70	2.43	2.56	3.27	2.52	2.89	2.92	2.41

ethnicity and gender did not have any differential effect on the various ethnic groups or on boys and girls that were both statistically and practically significant.

Study 2

This study modified the test administration of the CPTs at a community college by eliminating the initial computer screens that present background questions for all students taking the tests during a 2-week period (the experimental group). The test performance of these students was then compared with the performance of all students taking the tests in a standard administration during two adjacent weeks (the control group).

Method

Sample

The sample consisted of all incoming students at Central Piedmont Community College, Charlotte, NC, who took the CPTs for the first time during a 4-week period, from August 12 to September 7, 1996. The 4 weeks were assigned to the experimental and control conditions using an A-B-B-A counterbalanced order to minimize any trends over time.

The total experimental group consisted of 632 students who took the CPTs during the two middle weeks (those of August 20 and August 26): 329 White, 241 Black, and 62 other ethnicities; 296 men and 336 women. The total control group consisted of 709 students who took the tests the first week (that of August 12) or the last week (that of September 3): 395 White, 227 Black, and 87 other ethnicities; 351 men and 358 women. (One student in the experimental group whose ethnicity could not be ascertained and 7 students in the control group who took the CPTs with the test administration procedures for the experimental group were excluded.)

The size of the experimental and control groups for each CPT varies because students did not necessarily take all of the CPTs. The experimental and control groups, respectively, consisted of 561 and 615 students for Elementary Algebra, 582 and 656 for Arithmetic, 487 and 557 for Reading Comprehension, and 488 and 585 for Sentence Skills.

The total experimental and control groups were comparable in ethnicity (52.0% and 55.7% White, 38.1% and 32.0% Black, and 9.8% and 12.3% other ethnicity),[14] gender (46.8% and 49.5% men), age (45.6% and 51.5% 19 years old or under), intended program in college (27.8% and 32.4% associate's degree in arts and science, 34.3% and 36.0% associate's degree in a vocational field, 5.2% and 3.4% diploma in a vocational field, 24.1% and 21.7% undecided, and 8.5% and 6.5% not ascertained), and CPTs taken (88.8% and 86.7% Elementary Algebra, 92.1% and 92.5% Arithmetic, 77.1% and 78.6% Reading Comprehension, and 77.2% and 82.5% Sentence Skills).

Procedure

Students routinely scheduled to take the CPTs at the college's test center, before beginning their course work in the Fall 1996 semester, were directed to the 16 personal computers regularly used to administer the CPTs. For the experimental group, the initial computer screens containing the background questions,

[14]Other ethnic groups were pooled in the study because of their small size. The largest of these ethnic groups (Asian) consisted of 23 students in the experimental group and 33 in the control group.

including ethnicity, gender, date of birth, and parents' education, were eliminated on all computers, and a paper-and-pencil questionnaire with these questions was administered after the CPTs were completed.[15] No other changes were made in the test administration. For the control group, all of the regular test administration procedures were followed, including the presentation on all computers of the initial computer screens with the background questions.

Measures

CPTs. The CPTs consist of four tests: Elementary Algebra (12 items), Arithmetic (17 items), Reading Comprehension (20 items), and Sentence Skills (20 items covering sentence-level skills underlying writing). The CPTs are computer-adaptive tests, and the same number of items (12 to 20, depending on the test) is administered to all test takers. Test takers are required to attempt every item presented to them, the tests have no time limits, and there is no penalty for guessing. Version 4.5 of the CPTs was used.

One score was obtained for each test: the Total Right Score (reported to test takers and school officials). This score is an estimate of the number of items that the student would answer correctly if his or her test consisted of all 120 items in the original pool for each test. The Total Right Score corresponds to the Number Correct score in Study 1 and in Steele and Aronson's (1995) research. Because test takers must answer all items, the Accuracy and Number Attempted scores in Study 1 and in Steele and Aronson's research were not available.

Other variables. Ethnicity, gender, and other background variables were obtained from the CPTs' electronic records or the paper-and-pencil questionnaire. In cases where ethnicity and gender were not reported, this information was obtained from college records.

Analysis

The analyses of the four scores were identical to the analyses in Study 1, except that Asian and other ethnic groups were combined. In brief, a series of 2 × 3 × 2 (Condition: Experimental vs. Control × Ethnicity: White, Black, Other × Gender) ANOVAs were carried out. Planned comparisons of simple effects of the experimental versus control group factor for each ethnic group and each gender were conducted. And post hoc multiple comparisons of ethnic group means were made by Tukey's honestly significant difference test.

[15]The questions (in order) were: name, Social Security number, date of birth, date of test, years of English in high school, years of mathematics in high school, algebra course taken in high school, years since studied mathematics, gender, ethnicity, first language, disabilities, father's (or male guardian's) education, and mother's (or female guardian's) education. A copy of the questionnaire appears in Stricker and Ward (1998).

Table 4

Summary of Overall ANOVAs of Total Right Scores on Computerized Placement Tests, Study 2

Source	*df*	*F*			
		Elementary algebra	Arithmetic	Reading comprehension	Sentence skills
Experimental–Control (E–C)	1	6.43[a]***	0.15	1.04	0.05
Ethnicity	2	34.86[c]***	84.32[c]***	74.58[c]***	75.32[c]***
Gender	1	11.68[b]***	9.32[a]***	0.45	3.08[a]*
E–C × Ethnicity	2	1.84[a]	1.09	0.53	0.23
E–C × Gender	1	0.27	1.17	5.02[a]**	0.38
Ethnicity × Gender	2	1.49[a]	1.61[a]	1.03	0.08
E–C × Ethnicity × Gender	2	0.26	0.11	0.56	0.93

Note. Elementary Algebra, *df* for Error = 1,164, *MSE* = 629.53; Arithmetic, *df* = 1,226, *MSE* = 676.50; Reading Comprehension, *df* = 1,032, *MSE* = 426.51; Sentence Skills, *df* = 1,061, *MSE* = 465.81.
[a]$\eta > .05$. [b]$\eta > .10$. [c]$\eta > .24$.
*$p < .10$. **$p < .05$. ***$p < .01$.

Results[16]

Overall ANOVAs of the four scores are summarized in Table 4, and the analyses of the simple effects of experimental versus control group are summarized in Table 5. The means are reported in Table 6. Note that some of the scores were highly correlated. Correlations ranged from .37 to .78 (N = 440-561, $p < .01$) for the experimental group and from .24 to .73 (N = 499-615, $p < .01$) for the control group (N = 440-461, $p < .01$). The lowest correlation in both groups was for Reading Comprehension versus Elementary Algebra, and the highest was for Reading Comprehension versus Sentence Skills.

[16]In multiple comparisons of mean differences for the ethnic groups ($\alpha = .10$, $d = .10$), no additional differences were significant, both statistically and practically, for the four scores.

Table 5

Summary of Simple Effects of Experimental Versus Control Group in ANOVAs of Total Right Scores on Computerized Placement Tests, Study 2

	F			
Source	Elementary algebra	Arithmetic	Reading comprehension	Sentence skills
Ethnicity				
White	1.64	1.20	0.03	0.00
Black	0.11	0.14	0.33	0.60
Other	5.48[a]*	1.30	0.99	0.02
Gender				
Male	5.47[a]**	1.26	0.85	0.10
Female	1.77	0.21	4.74[a]*	0.29

Note. $df = 1$ for each simple effect. df for Error and the *MSE* for each analysis appear in Table 4.

[a]$\eta > .05$.

*$p < .10$ (familywise). **$p < .05$ (familywise).

Elementary Algebra

Of the main effects (condition, ethnicity, and gender) and interactions in the ANOVA of this variable, the main effects for ethnicity and gender were significant, both statistically ($p < .05$) and practically ($\eta > .10$). Women underperformed men. In the multiple comparisons of the mean differences for the ethnic groups, two of the differences were significant. Black students underperformed White students and students with other ethnicity. None of the simple effects for ethnicity or for gender were significant.

Arithmetic

The main effect for ethnicity was significant. In the multiple comparisons, three of the mean differences were significant. Black students and students with other ethnicity underperformed White students, and Black students underperformed students with other ethnicity. None of the simple effects for ethnicity or for gender were significant.

Table 6

Mean Total Right Scores on Computerized Placement Tests, Study 2

Group	White			Black			Other			Total	
	Men	Women	Total	Men	Women	Total	Men	Women	Total	Men	Women
Elementary algebra (*SD* = 25.09)											
Experimental	53.96	47.29	50.62	39.25	38.33	38.79	52.20	43.35	47.78	48.47	42.99
Control	56.01	50.39	53.20	41.54	37.71	39.62	65.22	52.31	58.76	54.26	46.80
Arithmetic (*SD* = 26.01)											
Experimental	72.36	64.85	68.61	47.38	46.57	46.98	58.25	55.24	56.75	59.33	55.56
Control	71.16	61.57	66.36	48.37	43.70	46.03	66.89	57.41	62.15	62.14	54.23
Reading comprehension (*SD* = 20.65)											
Experimental	80.29	82.01	81.15	63.84	69.12	66.48	61.89	69.48	65.69	68.67	73.54
Control	83.59	79.28	81.43	68.38	62.06	65.22	60.19	62.92	61.55	70.72	68.09
Sentence skills (*SD* = 21.58)											
Experimental	85.81	90.67	88.24	70.15	78.01	74.08	67.16	66.83	66.99	74.37	78.50
Control	87.71	88.77	88.24	72.33	72.30	72.31	65.19	70.10	67.64	75.08	77.06

Note. *SD*s calculated from MSEs in ANOVAs.

Reading Comprehension

The main effect for ethnicity was significant. In the multiple comparisons, two of the mean differences were significant. Black students and students with other ethnicity underperformed White students. None of the simple effects for ethnicity or for gender were significant.

Sentence Skills

The main effect for ethnicity was significant. In the multiple comparisons, three of the mean differences were significant. Black students and students with other ethnicity underperformed White students, and Black students outperformed students with other ethnicity. None of the simple effects for ethnicity or for gender were significant.

Summary

The results were similar to those for Study 1. Ethnic group differences in test performance were pervasive, gender differences were limited, and the inquiring manipulation had no differential effects for ethnic groups and genders that were both statistically and practically significant.

General Discussion

Overview

A clear and consistent finding in these two studies was the general absence of effects of inquiring about ethnicity and gender on performance on the two operational tests: the AP Calculus AB Examination, and the CPTs. No effects, negative or positive, that were both statistically and practically significant occurred, regardless of whether the students were Black, female, or from any other ethnic or gender group.

The convergence between the two studies, which differed markedly in tests and test-taking populations, supports the generalizability of these outcomes. These results fail to confirm the hypotheses about the adverse effects for Black and female students based on Steele and Aronson's (1995) findings for Black research participants and the implications of this result for the performance of females on quantitative tests. The present results are congruent with the findings of Croizet and Claire (1998) about the absence of effects of inquiring about socioeconomic status (SES). Of course, their research on SES and working-class French students is clearly less relevant to the present studies than is Steele and Aronson's research on ethnicity and Black students in this country.

Major Differences From Previous Research

Steele and Aronson's (1995) research and the present studies differed in some respects, already mentioned, that may account for the divergent findings. One major difference is that Steele and Aronson's research employed participants in laboratory studies, whereas the present studies used students taking operational tests with real-life consequences. Motivation to perform well was probably heightened in the high-stakes settings of the present studies. Whether this elevated motivation increases susceptibility to stereotype threat, as Aronson et al. (1999) suggest or, instead, overrides its harmful effects is unknown. The reduced motivation of participants taking achievement tests in research settings has been documented extensively (e.g., Brown & Walberg, 1993; Marsh, 1994; O'Neil, Sugrue, & Baker, 1995-1996).

Another difference is that Steele and Aronson's (1995) participants believed that they were working on a mundane task in an experiment, whereas the students in the present studies knew that they were being tested for their achievement or skills. It is conceivable that the level of stereotype threat is already so high on the tests in the present studies that questions about ethnicity and gender cannot increase it further. But this is a conjecture, for nothing is actually known about the ambient level of stereotype threat on these tests or any other tests when they are used operationally. Indeed, as Wheeler and Petty (2001) pointed out, it is not clear whether stereotype threat is always present when members of stigmatized groups are tested or whether it needs to be stimulated by other variables. A survey of students who took the Graduate Management Admission Test (Hecht & Schrader, 1986)—a test of verbal ability, quantitative ability, and analytical writing used for admission to graduate business schools—suggests that stereotype threat on operational tests may not be as pervasive as supposed. The students who took the test in May 1999 were asked immediately after the test about how other people evaluate the test takers' verbal and quantitative ability. No appreciable ethnic group or gender differences were found in the percentage reporting that their ability was underestimated, and all of the percentages were below 20% (B. Bridgeman, personal communication, July 7, 1999).[17]

[17]In response to the question, "When other people evaluate your verbal ability, do you think their estimates are *much too high, a little too high, about right, a little too low,* or *much too low*?" 11.2% of White students (N = 5,202), 13.2% of Black students (N = 607), 12.6% of Hispanic students (N = 477), 18.1% of Asian students (N = 1,065), 13.5% of men (N = 4,605), and 11.3% of women (N = 3,306) reported *a little too low* or *much too low.* Similarly, in response to a parallel question about quantitative ability, 18.3% of White students (N = 5,177), 19.0% of Black students (N = 605), 16.9% of Hispanic students (N = 478), 15.2% of Asian students (N = 1,062), 18.6% of men (N = 4,590), and 16.5% of woman (N = 3,290) reported *a little too low* or *much too low.*

Other Differences From Previous Research

Other differences between Steele and Aronson's (1995) research and the present studies deserve mention, although they are unlikely to account for the divergent results:

1. Steele and Aronson's (1995) participants were Stanford undergraduates, whereas the students in the present studies were high school students enrolled in AP courses across the country or students entering a community college. The AP students were probably very similar to the Stanford students in ability and ego involvement in the subject matter of the tests, and the community college students were probably less able and less ego involved. However, both groups in the present studies were probably highly ego involved in the outcome of the tests, as mentioned earlier. This point is supported by personal interviews with eight AP Calculus AB teachers in public and private high schools across the country, and with the director of the test center at the community college in Study 2.[18]

2. The difficulty of the test items is potentially important insofar as stereotype threat is enhanced when the test is seen to be hard. No data are available about students' perceptions of difficulty in Steele and Aronson's (1995) research or in the present studies. However, objectively the level of difficulty is roughly similar in Steele and Aronson's research and in these studies. Steele and Aronson's research and the AP test in Study 1 used a conventional testing approach, with all test takers being given the same items. Mean accuracy was 47.52% (i.e., 48% of attempted items were answered correctly) in Steele and Aronson's research, and 57.48% (i.e., 57% were answered correctly) in Study 1. In Study 2, the CPTs, because they are computer adaptive, are geared to administer items at each test taker's ability level, with the result that he or she should be able to answer about 60% correctly (allowing for chance success).

Whether a test is speeded—that is, test takers have insufficient time to complete it—can also contribute to the perception of a test's difficulty. The set of

[18]The AP teachers' consensus was that most students who take the AP Calculus AB Examination are highly motivated: Everyone takes the AP course voluntarily, most want to secure college credit, and in most high schools they also take the test voluntarily and pay their own test fees (G. Duering, D. Kennedy, S. Kornstein, D. Lotesto, J. Mechura, M. Montgomery, N. Stephenson, and M. White, personal communications, May 8-10, 2001). The test center director reported that almost all students who take the CPTs are highly motivated. Virtually everyone wants to avoid being placed in remedial courses, which require the same investment of time and tuition as regular courses, but earn no academic credit toward a degree, are not included in the GPA, and are seen as blemishes on the academic record (D. A. Rhoden, personal communication, January 17, 2001).

items in Steele and Aronson's (1995) research and the AP test in Study 1 had time limits, while the CPTs in Study 2 did not. It is unknown whether Steele and Aronson's set of items was speeded; the AP test was relatively unspeeded.[19]

3. In both studies, the two questions about ethnicity and gender were embedded in a set of other questions asking for various kinds of information, raising the possibility that the others mitigated the effects of the ethnicity and gender questions. In Study 1, only one question (date of birth) was demographic. In Study 2, three were demographic (date of birth, and two about parents' education). These age and SES questions are unlikely to offset the ethnicity and gender questions. Age is not an issue in either study, for the students were in high school in Study 1 and 90.4% were under 35 years of age in Study 2. SES is relevant in Study 2, but Croizet and Claire (1998) found that inquiring about this characteristic failed to affect test performance.

4. Unlike Steele and Aronson's (1995) participants, who were tested individually, the students in Study 1 took the test in a group administration. Group administrations are more depersonalized and test takers have greater anonymity, potentially ameliorating stereotype threat, but the group sizes in Study 1 were relatively small ($M = 18.96$ for the experimental group; the mean for the control group was unknown because the AP Calculus AB Examination was administered in some schools with another test, the AP Calculus BC Examination, to students from both AP courses), and depersonalization and anonymity were correspondingly limited. Furthermore, testing was done individually in Study 2 (the testing was done by computer in that study and in the Steele and Aronson replication), with the same outcome as Study 1, suggesting that mode of test administration is not important.

5. Another feature of Study 1 deserves comment. A small number of students in the experimental and control groups (8.2% and 7.4%, respectively) omitted their ethnicity. Steele and Aronson (1995, Study 3) found that many Black participants omitted their ethnicity when the experimental task was described

[19]Steele and Aronson's (1995) 25-min time limit for 27 items is slightly shorter than the 30-min time limit for 30 items on the GRE Quantitative test, the source of Steele and Aronson's items. Steele and Aronson's items were chosen because they were more difficult, and for that reason, the items should take longer to complete. With regard to the AP test, Educational Testing Service general guidelines concerning speededness are that (a) virtually all test takers should reach at least 75% of the items on a test, and (b) at least 80% of test takers should reach the last item (Swineford, 1974). For Part A of the AP test, only 0.2% of students in this study did not reach 19 of the 25 items, and 29.6% did not reach Item 24 (the number reaching the last item cannot be ascertained). For Part B, only 0.7% of students did not reach 12 of the 15 items, and only 15.6% did not reach Item 14.

as diagnostic, suggesting that these participants were especially affected by stereotype threat. Hence, insofar as those who omitted their ethnicity in the present study were largely Black students, the Black students who did report their ethnicity and made up the Black sample may have been atypically resistant to threat, explaining the lack of effects for them. This possibility cannot be ruled out, but it seems unlikely that most of those who omitted their ethnicity in Study 1 were Black students, judging from their test performance. In the experimental group, unaffected by the inquiry about ethnicity and gender, the performance of students with omitted ethnicity was substantially better than the performance of Black students (e.g., M accuracy = 57.04 vs. 51.24, $d = .31$). In any event, ethnicity was known for all but 1 student in Study 2, but the results were similar to Study 1, suggesting that the occurrence of omitted ethnicity in Study 1 may not be a serious concern.

6. As noted earlier, no adjustment for preexisting differences in ability was made, unlike Steele and Aronson's (1995) research, which eliminated ability differences by covarying on the participants' self-reported SAT Verbal score (Donlon, 1984). The focus of the analysis in the present studies was on comparisons of the performance of each ethnic group or gender in the experimental and control conditions, not comparisons of one subgroup (e.g., Black students) with another subgroup (e.g., White students). This analytical strategy not only makes control for preexisting differences unnecessary, but also avoids the interpretive complexities involved in using ability or achievement test scores as covariates, given the possibility that the covariate is also affected by the same phenomenon (stereotype threat) represented in the independent variable (Brown & Josephs, 1999; Sackett, Schmitt, Ellingson, & Kabin, 2001; Spencer et al., 1999; Steele & Aronson, 1995). Hence, these within-ethnic-group and within-gender contrasts in both studies are directly comparable to those in Steele and Aronson's research. The interactions between ethnic group or gender and experimental versus control group in the present studies are not comparable to those in Steele and Aronson's research. Nonetheless, these analyses are informative in describing the actual effects of the experimental manipulations on the test performance of the AP and CPT test takers.

7. In contrast to Steele and Aronson's (1995) research, assignment of test takers to experimental and control conditions was not strictly random. In Study 1, the initial assignment of schools to the experimental and control groups was random, but there were unavoidable differences between the two groups of schools that actually participated in the study. Some schools in the experimental group (about one fourth) normally offered a preadministration session for the AP Calculus AB Examination but agreed to eliminate it for the study; none of the schools in the control group offered such a

preadministration session. It is unlikely that this difference between the groups had an impact, given that the two sets of schools were matched on highly pertinent stratification variables. In Study 2, randomization was approximated by assigning students tested at different time periods to conditions, with adjacent weeks assigned to conditions in a counterbalanced order to minimize time trends. The experimental and control groups were observed to be similar on a number of key background variables. It is still possible, of course, that other relevant differences exist between the two groups.

8. The samples in both studies (N = 1,652 students in Study 1, including 122 Black students and 730 girls; N = 1,341 students in Study 2, including 468 Black students and 694 women) were substantially larger than the samples in Steele and Aronson's (1995) experiments (N = 44 in Study 4, including 22 Black participants; N = 20 Black participants in the replication). Hence, the statistical power to detect mean differences was appreciable in the present studies. In both studies, the power was approximately .99 or more for all effects in the ANOVAs (main effects, interactions, and simple effects), using the .05 alpha level and a medium ($\eta > .24$) effect size (Cohen, 1988).

Conclusions

A clear limitation of these studies was that data were only available about test performance and not about its possible causes (e.g., stereotype threat) or mediators (e.g., anxiety). This limitation is close to inevitable in field experiments in operational settings that must rely on unobtusive observations.

The findings, from a scientific perspective, shed some light on the generalizability of Steele and Aronson's (1995) results to operational tests. And, from a practical standpoint, the findings suggest that the common practice of making these inquiries does not adversely affect the performance of people taking these tests.

Much of the interest and importance of stereotype threat theorizing and research, like work on test anxiety (e.g., Spielberger & Vagg, 1995; Zeidner, 1998), stems from its implications for performance on operational tests and in school. This interest and importance is magnified for stereotype threat because of its pertinence to ethnic and gender differences in academic and test performance. Precisely for these reasons, research into the impact of stereotype threat on standardized tests in operational use is critical (Brown & Josephs, 1999; Frisby, 1999; Jensen, 1998; Ryan, 2001; Sackett et al., 2001; Steele, 1998; Whaley, 1998). Stereotype threat is clearly robust and potent in the laboratory, as amply documented by the research cited earlier. How this phenomenon plays out in real life, where many influences on academic and test performance operate, remains to be seen.

References

Ambady, N., Shih, M., Kim, A., & Pittinsky, T. L. (2001). Stereotype susceptibility in children: Effects of identity activation on quantitative performance. *Psychological Science, 12*, 385-390.

Aronson, J., Lustina, M. J., Good, C., Keough, K., Steele, C. M., & Brown, J. (1999). When White men can't do math: Necessary and sufficient factors in stereotype threat. *Journal of Experimental Social Psychology, 35*, 29-46.

Aronson, J., Quinn, D. M., & Spencer, S. J. (1998). Stereotype threat and the academic performance of minorities and women. In J. K. Swim & C. Stangor (Eds.), *Prejudice—The target's perspective* (pp. 83-103). San Diego, CA: Academic Press.

Blascovich, J., Spencer, S. J., Quinn, D., & Steele, C. (2001). African Americans and high blood pressure: The role of stereotype threat. *Psychological Science, 12*, 225-229.

Briel, J. B., O'Neill, K. A., & Scheuneman, J. D. (Eds.). (1993). *GRE technical manual: Test development, score interpretation, and research for the Graduate Record Examinations program*. Princeton, NJ: Educational Testing Service.

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

Brown, S. M., & Walberg, H. J. (1993). Motivational effects on test scores of elementary students. *Journal of Educational Research, 86*, 133-136.

Cheryan, S., & Bodenhausen, G. V. (2000). When positive stereotypes threaten intellectual performance: The psychological hazards of "modest minority" status. *Psychological Science, 11*, 399-402.

Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement, 33*, 107-112.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.

College Board. (1994). *Advanced Placement course description, Mathematics, Calculus AB, Calculus BC—May 1995, May 1996*. New York, NY: Author.

College Board. (1995). *ACCUPLACER program overview: Coordinator's guide*. New York, NY: Author.

College Board and Educational Testing Service. (1995a). *Advanced placement program statistical tables, 1994-95*. Princeton, NJ: Educational Testing Service.

College Board and Educational Testing Service. (1995b). *AP national summary reports, 1995*. Princeton, NJ: Educational Testing Service.

College Board and Educational Testing Service. (1995c). *A guide to the advanced placement program, May 1996*. Princeton, NJ: Educational Testing Service.

Croizet, J.-C., & Claire, T. (1998). Extending the concept of stereotype threat to social class: The intellectual underperformance of students from low socioeconomic backgrounds. *Personality and Social Psychology Bulletin, 24*, 588-594.

Donlon, T. F. (Ed.). (1984). *The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests*. New York, NY: College Board.

Educational Testing Service. (1995). [Test analyses of 1995 Advanced Placement examinations]. Unpublished raw data.

Educational Testing Service. (1998). [Test analyses of Computerized Placement Tests, DOS 5.2 version]. Unpublished raw data.

Frisby, C. L. (1999). Culture and test session behavior: Part II. *School Psychology Quarterly, 14*, 281-303.

Hecht, L. W., & Schrader, W. B. (1986). *Technical report on test development and score interpretation for GMAT users*. Princeton, NJ: Graduate Management Admission Council and Educational Testing Service.

Howell, D. C. (1997). *Statistical methods for psychology* (4th ed.). Belmont, CA: Duxbury.

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

Jensen, A. R. (1998). *The g factor—The science of mental ability*. Westport, CT: Praeger.

Marsh, H. W. (1984). Experimental manipulations of university student motivation and their effects on examination performance. *British Journal of Educational Psychology, 54*, 206-213.

O'Neil, H. F., Jr., Sugrue, B., & Baker, E. L. (1995-1996). Effects of motivational intervention on the National Assessment of Educational Progress mathematics performance. *Educational Assessment, 3*, 135-157.

Oswald, D. L., & Harvey, R. D. (2000-2001). Hostile environments, stereotype threat, and math performance among undergraduate women. *Current Psychology: Developmental, Learning, Personality, Social, 19*, 338-356.

Overall, J. E., & Spiegel, D. K. (1969). Concerning least squares analysis of experimental data. *Psychological Bulletin, 72*, 311-322.

Quinn, D. M., & Spencer, S. J. (2001). The interference of stereotype threat with women's generation of mathematical problem-solving strategies. *Journal of Social Issues, 57*, 55-71.

Richard, F. D., Bond, C. F., Jr., & Stokes, J. J. (2001). *One hundred years of social psychology quantitatively described*. Manuscript submitted for publication.

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

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.

Shih, M., Pittinsky, T. L., & Ambady, N. (1999). Stereotype perceptibility: Identity salience and shifts in quantitative performance. *Psychological Science, 10*, 81-84.

Spencer, S. J., Steele, C. M., & Quinn, D. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35*, 29-46.

Spielberger, C. D., & Vagg, P. R. (Eds.). (1995). *Test anxiety: Theory, assessment, and treatment*. Washington: Taylor & Francis.

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

Steele, C. M. (1998). Stereotyping and its threat are real. *American Psychologist, 53*, 680-681.

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.

Stricker, L. J. (1998). *Inquiring about examinees' ethnicity and sex: Effects on AP Calculus AB examination performance* (College Board Rep. 98-1, ETS Res. Rep. 98-5). New York, NY: College Board.

Stricker, L. J., & Ward, W. C. (1998). *Inquiring about examinees' ethnicity and sex: Effects on Computerized Placement Test performance* (College Board Rep. 98-2, ETS Res. Rep. 98-9). New York, NY: College Board.

Swineford, F. (1974). *The test analysis manual* (ETS Statistical Report 74-06). Princeton, NJ: Educational Testing Service.

Walsh, M., Hickey, C., & Duffy, J. (1999). Influence of item content and stereotype situation on gender differences in mathematical problem solving. *Sex Roles, 41*, 219-240.

Ward, W. C. (1988). The College Board Computerized Placement Tests: An application of computerized adaptive testing. *Machine-Mediated Learning, 2*, 271-282.

Whaley, A. L. (1998). Issues of validity in empirical tests of stereotype threat theory. *American Psychologist, 53*, 679-680.

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

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