

**Inquiring About
Examinees' Ethnicity
and Sex: Effects on AP[®]
Calculus AB Examination
Performance**

LAWRENCE J. STRICKER

Acknowledgments

Thanks are due to Walter B. MacDonald for encouraging this 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. McNerney, and Margaret L. Redman for recruiting AP classes; Geraldine Kovar, Behroz T. Manechshana, and Rick Morgan for providing AP Examination data; Donald A. Rock for advising on the statistical analysis; Thomas J. Jirele and Ting Lu for doing the computer analysis; Walter Emmerich for advising on the interpretation of the findings; and Rick Morgan, Claude M. Steele, and Gita Z. Wilder for commenting on a draft of this report.

Lawrence J. Stricker is a principal research scientist at ETS.

Researchers are encouraged to freely express their professional judgment. Therefore, points of view or opinions stated in College Board Reports do not necessarily represent official College Board position or policy.

Founded in 1900, the College Board is a not-for-profit educational association that supports academic preparation and transition to higher education for students around the world through the ongoing collaboration of its member schools, colleges, universities, educational systems and organizations.

In all of its activities, the Board promotes equity through universal access to high standards of teaching and learning and sufficient financial resources so that every student has the opportunity to succeed in college and work.

The College Board champions—by means of superior research; curricular development; assessment; guidance, placement, and admission information; professional development; forums; policy analysis; and public outreach—educational excellence for all students.

Additional copies of this report may be obtained from College Board Publications, Box 886, New York, New York 10101-0886. The price is \$15. Please include \$4 for postage and handling.

Copyright © 1998 by College Entrance Examination Board and Educational Testing Service. All rights reserved. College Board, Advanced Placement Program, AP, SAT, and the acorn logo are registered trademarks of the College Entrance Examination Board.

Printed in the United States of America.

Contents

<i>Abstract</i>	1
<i>Introduction</i>	1
<i>Method</i>	2
Samples	2
Procedure	2
Measures	3
Analysis	4
<i>Results and Discussion</i>	4
Intercorrelations and Reliability	4
Analyses of Variance	5
<i>Conclusions</i>	5
Laboratory and Real-Life Testing	9
Test Purpose and Content	9
Test Difficulty	10
Other Differences	10
Further Research	11
<i>References</i>	12
<i>Appendix</i>	13

Tables

1. Characteristics of Classes in Experimental and Control Groups	3
2. Intercorrelations of Scores	5
3. Correlations Between Same Scores on Different Multiple-Choice Sections	5
4. Summary of Analyses of Variance of Number Omitted, Number Not Reached, and Number Attempted Scores	6
5. Summary of Analyses of Variance of Number Correct Score, Percent Correct Score, and Formula Score	6
6. Summary of Analyses of Variance of Free- Response Section Score, Composite Score, and AP Grade	7
7. Mean Scores for Ethnic Groups	7
8. Mean Scores for Boys and Girls	8

This page is intentionally blank. Please continue on to the next page.

Abstract

Steele and Aronson (1995) found that the performance of African-American subjects on test items portrayed as a problem-solving task, in a laboratory experiment, was adversely affected when they were asked about their ethnicity. This outcome was attributed to “stereotype threat”: performance was disrupted by the subjects’ concerns about fulfilling the negative stereotype concerning African Americans’ intellectual ability. Extending that research, this field experiment evaluated the effects of inquiring about ethnicity and sex on the performance of examinees taking the Advanced Placement (AP) Calculus AB Examination in an actual test administration. With a minor exception, this inquiry had no statistically and practically significant effects on the test performance of African-American, female, or other subgroups of examinees.

Introduction

Recent research by Steele and Aronson (1995; Study 4) found that the performance of African-American undergraduates on difficult verbal ability items from the Graduate Record Examination (GRE) General Test (Briell, O’Neill, and Schueneman, 1993), portrayed as a verbal problem-solving task, was adversely affected when they were asked about their ethnicity immediately prior to working on the items, though the performance of white students was unaffected. African-American subjects who were asked about their ethnicity answered fewer items correctly, answered correctly a smaller percentage of the items that they attempted, and attempted fewer items than did African-American subjects who were not asked. This effect was not only replicable but also substantial (e.g., the mean difference for the number of items answered correctly by African-American subjects in the two conditions in the replication represented a d of 1.05; Cohen, 1988). Steele and Aronson attributed these results to “stereotype threat”: asking about ethnicity primes African-American subjects’ concerns about fulfilling the negative racial stereotype concerning their intellectual ability, thereby disrupting their performance. Based on a series of studies that elicited stereotype threat in other ways for African-American subjects taking verbal tests (Steele and Aronson, 1995) and for women taking quantitative tests (Spencer, Steele, and Quinn, 1997), Steele and his coworkers (Steele and Aronson, 1995; Spencer et al., 1997; and Steele, 1997) suggest that this phenomenon may help to account for the deficit on standardized tests and academic performance in school

that is observed for African-American, female, and other groups of examinees burdened by negative stereotypes about their ability.

The Steele and Aronson research on inquiring about ethnicity parallels the Advanced Placement (AP) Examination (College Board and Educational Testing Service, 1995a) administration procedure, which includes questions about ethnicity and sex that examinees are asked on the answer sheet immediately before they take the test, raising the possibility that the same phenomenon may occur on the AP Examinations. At the same time, several differences between that research and the AP Examination situation may militate against the applicability of the Steele and Aronson results. The subjects were knowing participants in a laboratory experiment, whereas AP examinees take the tests for important personal reasons—to earn college credit. Hence the two groups may not have the same motivation. A related difference is that the subjects were led to believe that they were engaged in a problem-solving task, whereas AP examinees are aware that they are taking achievement tests. Other research by Steele and Aronson indicates that stereotype threat is magnified when the task is portrayed as diagnostic of the subjects’ intellectual ability.¹ Thus the potential impact that inquiring about ethnicity can have on stereotype threat should be reduced on AP Examinations. A possible difference concerns the perceived difficulty of the test. Research by Spencer et al. indicates that an important element in stereotype threat is that test takers perceive the test as being difficult, at the limits of their ability; the items used in the Steele and Aronson research were chosen to be hard. It is unclear whether AP examinees perceive the AP Examinations as difficult. Finally, the Steele and Aronson results concern the consequences of asking about ethnicity. However, females’ performance may be similarly affected, at least on quantitative tests, by being asked about their sex, given the Spencer et al. findings about the applicability of stereotype threat to women on such tasks.

Accordingly, the aim of this study, an extension of the Steele and Aronson research, was to investigate the effects on AP Examination performance of asking examinees about their ethnicity and sex. The AP Calculus AB Examination (College Board, 1994) was chosen for this purpose because (1) it is taken by a relatively large number of African-American and female examinees (4,020 African Americans and 47,275 girls in 1995;

¹Steele found in an unpublished pilot study that inquiring about ethnicity did not affect African-American subjects’ performance when the task was described as diagnostic of their ability (C. M. Steele, personal communication, May 21, 1997).

College Board and Educational Testing Service, 1995c), (2) it has substantial mean differences for white and African-American and for male and female examinees (AP grades of 2.82 versus 1.87, a *d* of .73, for white and African-American test takers; and 2.93 versus 2.62, a *d* of .24, for boys and girls in 1995; College Board and Educational Testing Service, 1995b, 1995c), and (3) its subject matter is pertinent to the negative stereotypes about females' quantitative ability and about African Americans' intellectual ability in general.

Method

Samples

The sampling had four objectives:

1. Obtain a sample of African-American AP Calculus AB examinees efficiently, given that African-American students are enrolled in only a fraction of AP Calculus AB courses (19.9 percent of 8,222 classes had African-American examinees in 1995; College Board and Educational Testing Service, 1995c; B. T. Maneckshana, personal communication, March 21, 1996), by restricting the classes in the sample to those that previously had African-American test takers.

2. Secure examinees who were first asked about their ethnicity and sex in the AP Examination 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 (1) classes that were provided with an earlier preadministration session before the test was taken in which examinees completed background information on the answer sheet, including answering the ethnicity and sex questions; and (2) examinees who took a previously administered AP Examination in the same testing period.

3. Select examinees who resided in the United States to ensure that they were exposed to the negative stereotypes about the intellectual ability of African Americans and females. This objective was accomplished by excluding classes offered in other countries.

4. 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.

AP Calculus AB classes taking part in the May 1996 examination administration were drawn from the 1,639 classes with one or more African-American examinees taking the AP Examination in 1995. A random sample of 181 classes (11 percent of the total), stratified on 1995 data for size (15 students or fewer, 16 or more),

percentage of African-American examinees (11 percent or less, 12 percent or more), and percentage of AP Calculus AB grades of 3 or higher (57 percent or less, 58 percent or more)² was drawn for the experimental group. Eighty-two of these classes actually participated in the study; most of the others did not because they were unwilling to eliminate their preadministration sessions for the AP Calculus AB Examination. Seventy-seven of these classes were used in the analysis; the five others were excluded because all their examinees had taken one or more of the nine previously administered AP Examinations in the same two-week testing period.³

A stratified random sample of 181 classes, plus an oversample of 36 classes, was drawn for the control group. A total of 133 classes were eligible to participate in the study; most of the others were ineligible because they used a preadministration session. Fourteen of the 133 were excluded because all their examinees had taken a previous AP Examination 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 characteristics of the classes in the experimental and control groups are summarized in Table 1, using data for 1996 AP examinees and 1995 AP Examination performance. The two groups were similar in total number of examinees (means of 18.96 and 20.95), sex (means of 50.72 and 50.23 for percentage of boys), and ethnicity (means of 62.25 and 61.06 for percentage of white examinees, 10.15 and 11.21 for percentage of African-American examinees, 13.58 and 12.17 for percentage of Asian examinees, and 7.50 and 8.37 for percentage of other examinees).⁴ (Data for students who were enrolled in the classes but did not take the AP Examination were not available.) The two groups were also similar in their performance on

²An AP grade of 3 or higher is considered "qualified" or passing (College Board and Educational Testing Service, 1995a).

³The nine AP Examinations were: French Language, Physics B, Physics C: Mechanics, Physics C: Electricity and Magnetism, Spanish Language, Latin: Vergil, Latin Literature, English Literature and Composition, and Music Theory.

⁴Other ethnic groups were pooled in the study because of their small size. The largest of these ethnic groups, Hispanic (combining Puerto Rican, Mexican American, and other South American), accounted for a mean percentage of 4.27 ($N = 64$) and 4.07 ($N = 65$) of the classes in the experimental and control groups.

The actual ethnic group categories provided to examinees to describe themselves were: American Indian or Alaskan native; Black or African American; Mexican American or Chicano; Asian, Asian American, or Pacific Islander; Puerto Rican; South American, Latin American, Central American, or other Hispanic; White; and Other.

TABLE 1

Characteristics of Classes in Experimental and Control Groups

Variable	Experimental (N=77)		Control (N=77)	
	Mean	S.D.	Mean	S.D.
Total Number of Examinees	18.96	20.11	20.95	18.69
Percent Male Examinees	50.72	17.98	50.23	22.76
Percent Female Examinees	49.28	17.98	49.77	22.76
Percent White Examinees	62.25	28.43	61.06	27.85
Percent African-American Examinees	10.15	17.74	11.21	20.19
Percent Asian Examinees	13.58	17.17	12.17	15.76
Percent Other Examinees	7.50	11.32	8.37	15.62
Percent Examinees with Omitted Ethnicity	6.52	9.34	7.19	8.20
Percent AP Grades of 3 or Higher in 1995	49.61	31.30	49.99	34.10

the AP Calculus AB Examination in 1995 (means of 49.61 and 49.99 for percentage of examinees with AP grades of 3 or higher).

The examinees in the analysis consisted of students from the classes in the experimental and control groups who had not taken a previously administered AP Examination in the same testing period. The experimental group consisted of 755 examinees: 407 boys and 348 girls; 429 white, 52 African-American, 151 Asian, 61 other ethnicities, and 62 with omitted ethnicity. The control group consisted of 897 examinees: 515 boys and 382 girls; 555 white, 70 African-American, 152 Asian, 54 other ethnicities, and 66 with omitted ethnicity.

Procedure

Experimental Group

The AP Calculus AB classes in the experimental group were recruited by telephoning the AP coordinators, high school staff members responsible for administering the AP Examinations, and asking their schools to participate. The AP coordinators were told that the value of modifying AP Examination administration procedures was being studied, specifically how and when students fill out background information on the answer sheet. They were told that the study involved changes in the answer sheets and were asked (1) not to give the Calculus AB Examination in the same room as the Calculus BC Examination because of the altered test administration procedures for the former; and (2) not to offer a preadministration session for examinees for whom Calculus AB was their first AP Examination in the testing period. AP coordinators who agreed to par-

ticipate were sent a modified version of the general instructions for administering AP Examinations, revised to be in line with the changes in the answer sheets, and a supply of special answer sheets.

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 examinees after the test, was a regular answer sheet. Examinees were asked to complete only the identifying information on Side 1 and all the background information on Side 2, including ethnic group, sex, date of birth, and present grade level. (Both answer sheets appear in the Appendix.)

Examinees were instructed that:

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.

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 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 and calculators cannot be used; Part B has 15 items and graphing calculators can be used) and six free-response questions (graphing calculators can be used). A variety of scores were obtained for the multiple-choice items (for each section and for the entire test) and for the free-response questions:

1. All the scores for multiple-choice items used by Steele and Aronson (1995); Number Correct; "accuracy," Percent Correct: Number Correct/(Number Correct and Number Wrong); and Number Attempted (Number Correct and Number Wrong).

2. All scores for multiple-choice items routinely used in test analyses: Number Omitted, Number Not Reached, and Formula Score.
3. Special AP scores: Free-Response Section Score; Composite Score, based on the Formula Score and Free-Response Section Score; and AP grade.

The 21 specific scores were:

- | | |
|-------------------------------|---------------------------------|
| 1. Number omitted: Part A | 12. Number correct: Total |
| 2. Number omitted: Part B | 13. Percent correct: Part A |
| 3. Number omitted: Total | 14. Percent correct: Part B |
| 4. Number not reached: Part A | 15. Percent correct: Total |
| 5. Number not reached: Part B | 16. Formula score: Part A |
| 6. Number not reached: Total | 17. Formula score: Part B |
| 7. Number attempted: Part A | 18. Formula score: Total |
| 8. Number attempted: Part B | 19. Free-response section score |
| 9. Number attempted: Total | 20. Composite score |
| 10. Number correct: Part A | 21. AP grade |
| 11. Number correct: Part B | |

Ethnicity and sex were determined from AP files that included examinees' responses on the answer sheet for the AP Calculus AB Examination or on answer sheets for AP Examinations taken subsequently during the testing period.

Analysis

Data were pooled across classes for examinees in the experimental group and for examinees in the control group. The product-moment intercorrelations of the scores were computed separately for the experimental and control groups. Reliability was estimated by coefficient alpha for all scores, except Percent Correct, Composite Score, and AP Grade. For the Percent Correct score, the stepped-up split-half reliability was computed. For the Composite Score, the reliability of a weighted composite of the Formula Score: Total and Free-Response Section Score, was computed. The reliability of the AP Grade was not computed; its reliability should be close to that of the Composite Score, because the AP Grade is a direct function of the Composite Score.

A series of 2 (Experimental versus Control) x 5 (Ethnicity: White, African-American, Asian, Other, Omitted) x 2 (Sex) factorial analyses of variance were carried out using the least squares method (Model II error term; Overall and Spiegel, 1969) to deal with unequal *N*s. Planned comparisons of simple main effects of the experimental versus control group factor for each ethnic group (e.g., African-American examinees in the experimental group versus African-American examinees in the control group) and each sex (e.g., girls in the experimental group versus girls in the control group) were also conducted (Howell, 1997). Note that all these analyses used unweighted means. Effect sizes were assessed by the correlation ratio (η).

Both statistical and practical significance were considered in evaluating the results. A .05 significance level and an η of .10 (Cohen's, 1988, definition of a "small" effect size) were employed throughout (including the planned comparisons of simple main effects; Keppel, 1982).

Results and Discussion

Intercorrelations and Reliability

The intercorrelations of the various test scores (except scores for the multiple-choice sections) for the experimental and control groups are reported in Table 2; the correlations between the same scores on the different multiple-choice sections for the two groups are given in Table 3. The reliabilities of the scores are also shown in these tables.

The reliability was generally similar for the two groups, with the exception of consistently lower reliability for Number Not Reached scores for the control group (.83 to .88 versus .73 to .74). All the reliabilities were above .60, except for Percent Correct: Part B (.55 to .58), and all the reliabilities of the total scores were above .80, except for Number Not Reached for the control group (.77) and Free-Response Section Score for both groups (.79).

In general, the correlations were similar for the two groups. The same scores for the different multiple-choice sections correlated substantially (.59 to .65 and .59 to .64), except for Number Not Reached (.39 and .20). Corresponding total scores also correlated substantially (Number Omitted and Number Attempted, -.86 and -.94; Number Not Reached and Number Attempted, -.65 and -.67; Number Correct and Percent Correct, .90 and .92; Number Correct and Formula Score, .99 for both groups; and Percent Correct and Formula Score, .95 and .96). The total scores for correctness on the multiple-choice items (Number Correct, Percent Correct, and Formula Score) correlated highly with the Free-Response Section Score (.80 for both groups for Number Correct, .73 and .77 for Percent Correct, and .80 and .81 for Formula Score). And these correctness scores and the Free-Response Section Score correlated highly with the Composite Score (.94 for both groups for Number Correct, .89 and .91 for Percent Correct, .95 for both groups for Formula Score, and .94 and .95 for Free-Response Section Score) and with the AP Grade (.91 for both groups for Number Correct, .86 and .89 for Percent Correct, .92 for both groups for Formula Score, and .92 and .91 for both groups for Free-Response Section Score). Note that these are part-whole correlations. The Composite Score

TABLE 2

Intercorrelations of Scores

	Score									Reliability	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	Exp	Con
1. Number Omitted: Total		.16	-.86	-.44	-.07	-.32	-.31	-.33	-.31	.81	.80
2. Number Not Reached: Total	.39		-.65	-.33	-.16	-.24	-.15	-.21	-.19	.88	.77
3. Number Attempted: Total	-.94	-.67		.51	.14	.37	.31	.36	.34	.87	.86
4. Number Correct: Total	-.45	-.30	.47		.90	.99	.80	.94	.91	.86	.86
5. Percent Correct: Total	-.11	-.16	.12	.92		.95	.73	.89	.86	.82	.84
6. Formula Score: Total	-.33	-.21	.34	.99	.96		.80	.95	.92	.85	.86
7. Free-Response Section Score	-.24	-.20	.27	.80	.77	.81		.94	.92	.79	.79
8. Composite Score	-.30	-.22	.32	.94	.91	.95	.95		.97	.84	.83
9. AP Grade	-.30	-.21	.31	.91	.89	.92	.91	.96		-	-

Note: Correlations for the experimental group appear above the diagonal; correlations for the control group appear below it. Correlations of .07 and .09 are significant at the .05 and .01 levels (two-tail), respectively, for both the experimental and control groups.

TABLE 3

Correlations Between Same Scores on Different Multiple-Choice Sections

Score	Correlation		Reliability ^a			
	Exp	Con	Exp		Con	
			A	B	A	B
Number Omitted	.63	.59	.72	.65	.73	.62
Number Not Reached	.39	.20	.83	.88	.74	.73
Number Attempted	.59	.64	.82	.77	.80	.71
Number Correct	.65	.64	.82	.64	.83	.66
Percent Correct	.62	.60	.79	.55	.80	.58
Formula Score	.65	.64	.82	.62	.83	.64

^a Internal-consistency reliability.

and AP Grade also correlated highly (.97 and .96). Again, note that the AP Grade is a function of the Composite Score. The pattern of correlations for the scores on the multiple-choice sections was similar to that for the total scores.

In summary, the reliability and intercorrelations of the various test scores were similar for the experimental and control groups. The reliability of the scores was generally substantial, particularly the total scores. The same scores for different multiple-choice sections correlated appreciably. The corresponding total scores correlated substantially. And the correctness scores for multiple-choice items and the Free-Response Section Score correlated highly with each other and with the Composite Score and AP Grade. Hence, the various scores had adequate reliability for the main analyses of means, but many of the scores were highly related and not independent of each other.

Analyses of Variance

The analyses of variance of the 21 scores as well as the related planned comparisons, are summarized in Tables 4 to 6; the corresponding means for the subgroups in the experimental and control groups are shown in Tables 7 and 8.

Because the focus is on differences between the experimental and control groups for each ethnic group and sex, the main effects for experimental versus control group, ethnicity, and sex, and the interactions between ethnicity and sex, will not be described. None of the 21 two-way interactions of experimental versus control group with ethnicity, none of the 21 two-way interactions of experimental versus control group with sex, and none of the 21 three-way interactions of experimental versus control group with ethnicity and sex were both statistically and practically significant.

None of the 105 simple main effects for ethnicity (White, African American, Asian, Other, Omitted) were significant. One of the 42 simple main effects for sex was significant ($F = 17.86$, $p < .01$, $\eta = .10$): girls in the experimental group had a higher mean on Number Not Reached: Part B (1.21) than girls in the control group (.73).

In summary, in analyses of variance of the means for the various test scores, none of the interactions of experimental versus control group with ethnicity or sex, and only one of the simple main effects for experimental versus control group for ethnicity or sex, was significant. In short, the test scores for an ethnic group or sex were generally unaffected by the experimental manipulation.

TABLE 4

Summary of Analyses of Variance of Number Omitted, Number Not Reached, and Number Attempted Scores

Source	df	F								
		Number Omitted			Number Not Reached			Number Attempted		
		Part A	Part B	Total	Part A	Part B	Total	Part A	Part B	Total
Experimental-Control (E-C)	1	.03	.05	.05	.17	11.22**	5.35*	.10	4.54*	1.46
Sex	1	5.88*	11.63**	9.94**	1.55	.93	1.84	6.15**	10.23**	9.58**
E-C x Sex	1	.88	1.68	1.46	.00	1.39	.48	.53	2.86	1.59
Male	1	.62	1.27	1.07	.24	5.13**	2.87	.18	.21	.00
Female	1	1.09	1.99	1.77	.11	17.86**a	7.86**	.95	12.74**	5.32*
Ethnicity	4	4.29**a	.75	3.01*	2.58*	3.12*	4.23**a	4.01**	1.93	3.44**
E-C x Ethnicity	4	.66	.33	.53	1.21	2.73*	1.28	.87	1.81	.97
White	1	.62	.16	.48	.26	.21	.36	.17	.00	.08
African American	1	.12	.35	.00	2.83	5.29*	.15	1.00	3.26	.04
Asian	1	1.62	.65	1.44	.72	10.67**	6.42*	1.92	6.46*	4.35*
Other	1	.07	.16	.13	1.28	2.84	3.01	.07	.46	.24
Omitted	1	.20	.02	.05	.35	.00	.15	.38	.01	.12
Ethnicity x Sex	4	.70	1.26	1.01	1.48	.83	.55	1.23	.82	1.22
E-C x Ethnicity x Sex	4	.56	.78	.65	.56	.39	.09	.53	.40	.44
Error	1632	(8.35)	(4.21)	(19.74)	(2.43)	(2.43)	(6.42)	(13.03)	(7.06)	(31.79)

Note: Values enclosed in parentheses represent mean square errors. * $p < .05$; ** $p < .01$; $\eta^2 > .10$.

TABLE 5

Summary of Analyses of Variance of Number Correct Score, Percent Correct Score, and Formula Score

Source	df	F								
		Number Correct			Percent Correct			Formula Score		
		Part A	Part B	Total	Part A	Part B	Total	Part A	Part B	Total
Experimental-Control (E-C)	1	.11	.39	.00	.36	.00	.21	.16	.06	.04
Sex	1	9.93**	27.41**a	18.32**a	5.06*	21.65**a	11.59**	8.83**	25.12**a	16.44**
E-C x Sex	1	3.99*	2.39	4.05*	4.24*	5.18*	6.77**	5.02*	4.08*	5.60*
Male	1	3.01	5.13*	4.46*	2.31	5.68*	4.97*	3.66	5.55*	5.15*
Female	1	4.74*	.74	3.48	6.17*	4.48*	8.18**	6.10*	2.76	5.70*
Ethnicity	4	14.14**a	8.83**a	13.09**a	11.25**a	7.43**a	10.13**a	13.80**a	8.63**a	12.75**a
E-C x Ethnicity	4	.73	.13	.41	.79	.38	.59	.67	.12	.43
White	1	.71	.01	.39	.67	.08	.50	.70	.01	.38
African American	1	1.72	.04	.70	2.33	.09	1.36	1.54	.02	.87
Asian	1	.06	.08	.00	.55	.63	.80	.22	.04	.17
Other	1	.02	.49	.14	.24	.90	.38	.01	.38	.10
Omitted	1	1.05	.01	.57	.45	.00	.24	.99	.02	.56
Ethnicity x Sex	4	1.39	1.02	1.39	1.11	1.10	1.10	1.40	.91	1.30
E-C x Ethnicity x Sex	4	1.70	2.73*	2.22	2.20	2.52*	2.80*	1.86	2.97*	2.44*
Error	1632	(24.79)	(7.94)	(50.59)	(427.46)	(430.46)	(346.09)	(34.15)	(10.61)	(69.03)

Note: Values enclosed in parentheses represent mean square errors. * $p < .05$; ** $p < .01$; $\eta^2 > .10$.

TABLE 6

Summary of Analyses of Variance of Free-Response Section Score, Composite Score, and AP Grade

Source	df	F		
		Free-Response Section Score	Composite Score	AP Grade
Experimental				
-Control (E-C)	1	.32	.15	.14
Sex	1	10.35**	14.59**	14.77**
E-C x Sex	1	2.32	4.29*	4.16*
Male	1	.99	3.10	3.02
Female	1	3.80	5.25*	5.07*
Ethnicity	4	14.57**a	14.21**a	11.79**a
E-C x Ethnicity	4	1.05	.45	.37
White	1	.73	.01	.00
African American	1	.42	.57	.32
Asian	1	2.70	1.16	1.09
Other	1	.30	.15	.25
Omitted	1	.07	.08	.00
Ethnicity x Sex	4	.35	.66	.80
E-C x Ethnicity x Sex	4	1.62	2.24	2.21
Error	1632	(104.49)	(408.81)	(1.57)

Note: Values enclosed in parentheses represent mean square errors.

* $p < .05$; ** $p < .01$; ^a $\eta > .10$.

TABLE 7

Mean Scores for Ethnic Groups

Score	Ethnicity										S.D. ^a	
	White		African American		Asian		Other		Omitted			
	Exp	Con	Exp	Con	Exp	Con	Exp	Con	Exp	Con		
Number Omitted												
Part A	3.24	3.39	3.82	4.00	2.97	2.54	3.14	3.29	3.72	3.48	2.89	
Part B	2.39	2.45	2.74	2.52	3.56	2.16	2.33	2.49	2.40	2.45	2.05	
Total	5.63	5.83	6.56	6.52	5.32	4.70	5.47	5.77	6.12	5.94	4.44	
Number Not Reached												
Part A	.86	.80	1.06	1.54	.99	.84	1.13	.80	.89	.73	1.56	
Part B	.76	.72	1.55	.89	1.19	.60	1.23	.74	.75	.74	1.56	
Total	1.62	1.52	2.61	2.43	2.19	1.44	2.37	1.53	1.64	1.47	2.53	
Number Attempted												
Part A	20.91	20.81	20.12	19.46	21.04	21.62	20.73	20.92	20.39	20.79	3.61	
Part B	11.84	11.84	10.71	11.59	11.45	12.23	11.43	11.77	11.85	11.80	2.66	
Total	32.75	32.65	30.83	31.05	32.49	33.85	32.16	32.69	32.24	32.59	5.64	
Number Correct												
Part A	12.80	12.53	10.82	9.61	14.14	14.00	11.61	11.74	12.20	13.11	4.98	
Part B	6.41	6.40	4.88	4.98	6.26	6.36	5.37	5.74	6.19	6.24	2.82	
Total	19.21	18.92	15.69	14.59	20.40	20.35	16.97	17.48	18.39	19.35	7.11	
Percent Correct												
Part A	60.61	59.51	54.40	48.58	66.33	64.56	54.98	56.90	59.66	62.14	20.67	
Part B	53.85	53.45	44.35	43.18	53.38	51.48	45.66	49.38	52.37	52.17	20.75	
Total	58.24	57.38	51.24	47.24	61.98	60.06	51.97	54.14	57.04	58.67	18.60	

(continued on page 8)

TABLE 7 (continued)

Mean Scores for Ethnic Groups

Score	Ethnicity										S.D. ^a	
	White		African American		Asian		Other		Omitted			
	Exp	Con	Exp	Con	Exp	Con	Exp	Con	Exp	Con		
Formula Score												
Part A	10.77	10.45	8.49	7.15	12.41	12.09	9.33	9.44	10.15	11.19	5.84	
Part B	5.06	5.04	3.42	3.32	4.96	4.89	3.85	4.23	4.77	4.85	3.26	
Total	15.83	15.49	11.90	10.47	17.37	16.98	13.18	13.68	14.92	16.04	8.31	
Free-Response Section Score	18.72	19.29	12.80	11.58	19.80	17.86	14.56	15.63	19.05	18.58	10.22	
Composite Score	40.11	40.27	28.90	26.10	43.30	40.78	32.62	34.09	39.34	40.36	20.22	
AP Grade	2.87	2.88	2.21	2.08	3.04	2.89	2.44	2.56	2.90	2.89	1.25	

^a Calculated from the mean square errors in the analyses of variance.

TABLE 8

Mean Scores for Boys and Girls

Score	Sex				S.D. ^a
	Boys		Girls		
	Exp	Con	Exp	Con	
Number Omitted					
Part A	3.04	3.19	3.71	3.49	2.89
Part B	2.11	2.26	2.78	2.56	2.05
Total	5.15	5.45	6.49	6.05	4.44
Number Not Reached					
Part A	.92	.87	1.05	1.01	1.56
Part B	.98	.75	1.21	.73	1.56
Total	1.91	1.62	2.26	1.74	2.53
Number Attempted					
Part A	21.04	20.94	20.24	20.50	3.61
Part B	11.91	11.99	11.01	11.71	2.66
Total	32.94	32.92	31.25	32.21	5.64
Number Correct					
Part A	12.51	13.08	12.11	11.31	4.98
Part B	6.18	6.60	5.46	5.28	2.82
Total	18.69	19.69	17.57	16.59	7.11
Percent Correct					
Part A	59.33	61.41	59.06	55.26	20.67
Part B	51.63	54.91	48.22	44.96	20.75
Total	56.61	59.36	55.58	51.63	18.60
Formula Score					
Part A	10.38	11.12	10.08	9.01	5.84
Part B	4.75	5.26	4.08	3.67	3.26
Total	15.13	16.38	14.16	12.69	8.31
Free-Response Section Score	17.59	18.26	16.39	14.91	10.22
Composite Score	38.08	40.44	35.63	32.20	20.22
AP Grade	2.77	2.92	2.62	2.41	1.25

^a Calculated from the mean square errors in the analyses of variance.

Conclusions

A clear and consistent finding of this study was the general absence of effects, negative or positive, of inquiring about ethnicity and sex on performance on the AP Calculus AB Examination for examinees who were African-American, female, or from any other ethnic or gender group, as reflected in a variety of correctness scores for multiple-choice items, scores for free-response questions, and test-taking behavior scores for multiple-choice items (i.e., omitting, not reaching, or attempting items). This outcome contrasts with the Steele and Aronson (1995) finding that asking about ethnicity adversely affected the test performance of African-American subjects and with the implications of this result for the performance of females on quantitative tests. (The absence of effects for white, Asian, and male examinees was not unexpected; negative stereotypes about intellectual ability are not prevalent for white people and males, and a positive stereotype about quantitative ability exists for Asian people.)

Laboratory and Real-Life Testing

It is impossible to account for the differences between the two sets of findings at this juncture because the present study and the Steele and Aronson (1995) research varied in a number of respects that might be relevant. An obvious difference, noted previously, is that the Steele and Aronson results were derived from subjects who were participating in laboratory experiments whereas the present findings came from examinees who were voluntarily taking a test with important real-life consequences. Differences in the motivation of the two groups are likely. Motivation to perform well was probably heightened in the high-stakes setting of the present

study, conceivably overriding any harmful effects of stereotype threat. The reduced motivation of subjects taking achievement tests in research settings has been extensively documented (e.g., Brown and Walberg, 1993; Marsh, 1984; O'Neil, Sugrue, and Baker, 1995/1996).

One possible indication of differences in motivation is that the number of multiple-choice items omitted or not reached was substantially greater in the Steele and Aronson research. In the initial Steele and Aronson study on inquiring about ethnicity (Study 4), the mean for these items, pooling the data for the 22 African-American and 22 white subjects in the two experimental conditions, was 11.98, or 44.4 percent of the 27 items. (Omitted and not reached items were not distinguished from one another in the Steele and Aronson research.) In the present study, the mean was 13.18, or 33 percent of the 40 items, pooling and equally weighting the data for African-American and white examinees only in the two conditions. Furthermore, in the replication of Study 4, limited to African-American subjects, one or more subjects completed only five items, taking 300 seconds per item compared, on average, to 70 seconds per item for all subjects.

Instructions about guessing should have led to more omitted answers in the present study because the examinees were penalized for guessing incorrectly; subjects in the Steele and Aronson research were not discouraged from guessing. On the other hand, the time limit in the Steele and Aronson research may have been too short (25 minutes for 27 items, or 56 seconds per item), leading to a large number of not reached items because of the subjects' inability to finish before the deadline. The time limit for the verbal sections of the GRE General Test, from which the items were obtained, is less generous (30 minutes for 38 items, or 47 seconds per item), but virtually all examinees complete these sections (the median percentage of examinees completing these sections was 95.7 percent on seven recent forms of the test; Educational Testing Service, 1990, 1992; E. Broch, personal communication, April 12, 1997). However, these forms were composed of both easy and hard items whereas the Steele and Aronson research used only difficult items answered correctly by 30 percent or fewer of the examinees taking the General Test, a small fraction of the items (12.2 percent of the verbal items in eight recent forms of the test; Educational Testing Service, 1990, 1992), and it probably takes more time to answer difficult items. Although the number of omitted and not reached answers in the Steele and Aronson research is clearly anomalous, whether it reflects subjects' motivation, insufficient time limits for the test, or something else is uncertain.

Test Purpose and Content

Another obvious difference, already mentioned, between the Steele and Aronson (1995) research and the present study concerns the purpose of the test employed and its content. The Steele and Aronson research used verbal ability items and described them as an innocuous verbal problem-solving task; the present study used a mathematics achievement test that explicitly reflected the students' mastery of subject matter they had just studied. The Steele and Aronson research makes it clear that stereotype threat is substantial on diagnostic tests, raising the possibility that the level of stereotype threat is already so high on the AP Examination that questions about ethnicity and sex cannot increase it any further. But nothing is actually known about whether AP Examinations are perceived by examinees as diagnostic of their intellectual ability or about the ambient level of stereotype threat on these tests.

It is also conceivable that examinees may make different attributions of the reasons for their performance on achievement tests than on ability tests, viewing their performance on achievement tests as simply reflecting the quality of the course, not their own intellectual ability (e.g., "I am able, but I did not learn anything in the course"). Such attributions would presumably eliminate stereotype threat because test performance does not reflect on the person, much less on the ethnic or gender group to which he or she belongs. This speculation is supported by the finding that students were more likely to attribute their performance on course examinations in statistics to the quality of the instruction than to their ability (Hunsley, 1987). Spencer et al. (1997), using achievement tests (GRE tests in English and mathematics), elicited stereotype threat, but unlike the AP Examination used in the present study, these tests are not linked to particular courses, making such external attributions less likely.

The connection between the AP Examination and the course in this study suggests still another possible explanation for the divergent findings. The experience of taking course examinations similar to the AP Examination and receiving feedback about performance on them may have inoculated AP examinees against stereotype threat so that it did not have an influence on the AP Examination. Other research by Steele and Aronson indicates that explicitly manipulating expectancies does not affect test performance or prevent the stereotype threat effect from occurring. However, expectations based on day-to-day experience may be more potent. Alternatively, this experience with course examinations may have caused AP examinees to "disidentify" (Steele, 1997) with the subject matter, vitiating the effects of stereotype

threat because of the test takers' lack of ego involvement in the test. This disengagement would not necessarily preclude examinees from taking the test (though some students may fail to take it for that reason), because there may be external pressures from schools, teachers, or parents to do so, such as needing a course grade to graduate.

Test Difficulty

A potential difference between the Steele and Aronson (1995) research and the present study, also previously mentioned, is the difficulty of the items, because stereotype threat is enhanced when the test is seen to be hard. Objectively, the items in the Steele and Aronson (1995) research and the present study appear to be roughly similar in difficulty. The mean Number Correct was 7.48 (27.7 percent of the items) in Steele and Aronson's initial study (Study 4), pooling the data for the two experimental conditions; the corresponding mean in the present study was 17.10 (42.8 percent of the items), pooling the data for African-American and white examinees in the two conditions. Moreover, in the same Steele and Aronson study, the mean Percent Correct was 47.52; the corresponding mean in the present study was 53.52.

More relevant are students' subjective perceptions of difficulty. No data are available on this issue for either study. Steele and Aronson did obtain data on these perceptions in a related study (Study 1), but the level of difficulty was not reported. It is conceivable that these perceptions differed in the present study, because mathematics tests may generally be seen as difficult. Indeed, attributions of success and failure to difficulty are more prevalent for tests in mathematics than in other fields (Birenbaum and Kraemer, 1995). In addition, attributions to difficulty are also more prevalent for courses in mathematics and science than in other courses (Ryckman and Peckham, 1987), and mathematics courses are perceived as more difficult than other courses (Eccles, 1984). The AP examinees' perceptions may also have been influenced, one way or the other, by their exposure to classroom examinations, similar to the AP Examination, taken during their course. If the other examinations were equal in difficulty to the AP Examination, the examinees might adapt and see the AP Examination as less difficult. But if the examinations were easier than the AP Examination, the examinees might see the AP Examination as more difficult. Adaptation of this kind has been observed for a course examination made up of multiple-choice and true-false items: when the items were presented in increasing order of difficulty, they were judged to be more difficult than when they were presented in decreasing order (Barcikowski and Olsen, 1975).

Other Differences

Other differences exist between the Steele and Aronson (1995) research and the present study that deserve mention, though they are unlikely to explain the discrepant findings. First, the Steele and Aronson subjects were undergraduates at Stanford University whereas the examinees in the present study were high school students enrolled in AP Calculus courses across the country. These two groups are probably very similar on an array of characteristics that are potentially relevant to their performance in the research, including identification with the academic skills being tested, motivation to attend college, preparation for it, and ability.

Second, the Steele and Aronson subjects were tested individually whereas the examinees in the present study took the examination in a group administration. Although group administrations are more depersonalized and the test takers have greater anonymity, potentially ameliorating stereotype threat, the group sizes in this study were relatively small (means of 18.96 and 20.95 examinees), and depersonalization and anonymity were correspondingly limited.

Third, in the present study some schools in the experimental group (about a quarter) 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 two samples had any impact, given that the two sets of schools were matched on three highly pertinent stratification variables.

Fourth, the sample in the present study was large, totaling 1,652 examinees, including 730 girls and 122 African Americans. (By contrast, the sample in the Steele and Aronson research was modest: 44 subjects in the initial study, Study 4, and 20 in the replication.) Hence, the statistical power was more than sufficient to identify mean differences for the various subgroups in the present study.

Fifth, the statistical analysis in the Steele and Aronson research controlled for subjects' self-reported SAT (Donlon, 1984) verbal scores to take into account differences in ability between the white and African-American subjects, whereas no controls were used in the present study. Using the SAT as a control variable in this situation introduces the danger that the covariate (SAT score) is affected by the same phenomenon represented in the independent variable (stereotype threat), as Steele and Aronson point out, though they argue that the relatively high SAT scores of their subjects suggest that performance on this test was unaffected. The focus of the analysis in the present study was on comparisons

of the performance of each ethnic or sex subgroup in the experimental and control groups, not comparisons of one subgroup (e.g., African-American examinees) with another subgroup (e.g., white examinees), making control for differences in ability unnecessary in view of the random assignment of classes to the two conditions and also avoiding the interpretive complexities involved in using ability or achievement test scores as covariates. Hence, these within-subgroup contrasts are directly comparable to those in the Steele and Aronson studies. The interactions of the experimental and control conditions with ethnicity and sex in the present study are not comparable to those in the Steele and Aronson research because no control for ability was used here, but these analyses are informative in describing the actual effects of the experimental manipulations on the test performance of AP examinees.

Further Research

A clear limitation of the present study 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 inevitable in field experiments. Follow-up research would be useful into possible causes of the differences in the findings of the Steele and Aronson (1995) research and the present study (particularly attributions of test performance, perceptions of test difficulty, experience with similar tests, and expectancies based on this experience) and into the applicability of these results to other test-taking populations and other tests, especially ability and intelligence tests, when they are used operationally.

More generally, the boundary conditions for the effects of stereotype threat on standardized test performance in real life need to be established: what is the ambient level of threat, how does it vary with the type of test, and how does it vary for different groups of examinees?

References

Barcikowski, R. S., and Olsen, H. (1975). Test item arrangement and adaptation level. *Journal of Psychology*, 90, 87-93.

Birenbaum, M., and Kraemer, R. (1995). Gender and ethnic-group differences in causal attributions for success and failure in mathematics and language examinations. *Journal of Cross-Cultural Psychology*, 26, 342-359.

Briel, J. B., O'Neill, K. A., and Schueneman, J. D. (1993). *GRE technical manual*. Princeton, NJ: Educational Testing Service.

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

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

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

College Board and Educational Testing Service. (1995a). *A guide to the Advanced Placement Program. May 1996*. Princeton, NJ: Educational Testing Service.

College Board and Educational Testing Service. (1995b). *Advanced Placement Program statistical tables, 1994-95*. Princeton, NJ: Educational Testing Service.

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

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

Eccles, J. (1984). Sex differences in achievement patterns. *Nebraska Symposium on Motivation*, 32, 97-132.

Educational Testing Service. (1990). *Practicing to take the GRE General Test—No. 8*. Princeton, NJ: Author.

Educational Testing Service. (1992). *Practicing to take the GRE General Test—No. 9*. Princeton, NJ: Author.

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

Hunsley, J. (1987). Cognitive processes in mathematics anxiety and test anxiety: The role of appraisals, internal dialogue, and attributions. *Journal of Educational Psychology*, 79, 388-392.

Keppel, G. (1982). *Design and analysis—A researcher's handbook* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

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., and Baker, E. L. (1995/1996). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. *Educational Assessment*, 3, 135-157.

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

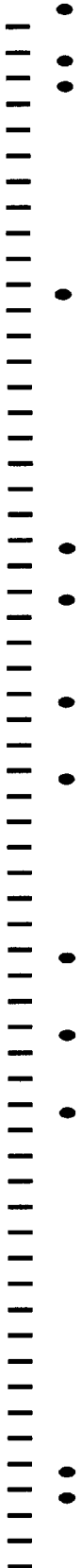
Ryckman, D. B., and Peckham, P. (1987). Gender differences in attributions for success and failure across subject areas. *Journal of Educational Research*, 81, 120-125.

Spencer, S. J., Steele, C. M., and Quinn, D. (1997). *Stereotype threat and women's math performance*. Manuscript submitted for publication.

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

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

This page is intentionally blank. Please continue on to the next page.



SIDE 2

136	A B C D E	139	A B C D E	142	A B C D E	145	A B C D E	148	A B C D E	151	A B C D E	154	A B C D E
137	A B C D E	140	A B C D E	143	A B C D E	146	A B C D E	149	A B C D E	152	A B C D E	155	A B C D E
138	A B C D E	141	A B C D E	144	A B C D E	147	A B C D E	150	A B C D E	153	A B C D E	156	A B C D E

1. YOUR MAILING ADDRESS

* YOUR GRADE REPORT WILL BE MAILED TO THIS ADDRESS IN JULY

1. YOUR MAILING ADDRESS

* USING THE ABBREVIATIONS GIVEN IN YOUR CANDIDATE PACK, FILL ADDRESS INTO BOXES PROVIDED. IF YOUR ADDRESS DOES NOT FIT, SEE ITEM 10 BELOW. * INDICATE A SPACE IN YOUR ADDRESS BY LEAVING A BLANK BOX AND FILLING IN THE CORRESPONDING DIAMOND (◇) BELOW THE BOX

Grid for mailing address with diamond indicators for spaces. Includes fields for Street, City, State, U.S. ZIP CODE, and Social Security Number.

Numbered grid for address continuation, containing numbers 136 through 156.

3. SEX (Male/Female) and 5. DATE OF BIRTH (Month/Day/Year)

4. PRESENT GRADE LEVEL (9th grade through College/Other)

8. WILL YOU BE APPLYING FOR SOPHOMORE STANDING AT COLLEGE? (Yes/No)

8. STUDENT SEARCH SERVICE OF THE COLLEGE BOARD (Complete ONLY if you are a SOPHOMORE or a JUNIOR). Includes options for wanting college board information and scholarship programs.

7. ETHNIC GROUP (American Indian or Alaskan native, Black or African American, Mexican American or Chicano, Asian, South American, or Pacific Islander, Puerto Rican, Central American, or other Hispanic, White, Other)

6. EXPECTED DATE OF COLLEGE ENTRANCE (Fall, Winter/Spring, Summer, Un-decided) with years 1996-1999

11. COLLEGE TO RECEIVE YOUR AP GRADES. Includes College Name and Address, and College Code.

10. SCHOOL YOU ATTEND. Includes School Name, City, and State, and School Code.

1b. If the address gridded above is not complete enough for delivery of your grade report, please fill in this oval and print your complete address below.

1a. COUNTRY CODE and fields for Address, City, State or Province, Country, and Zip or Postal Code.