Listening. Learning. Leading.®

Research Report
ETS RR–15-02

Stereotype Threat, Inquiring About Test Takers' Race and Gender, and Performance on Low-Stakes Tests in a Large-Scale Assessment

Lawrence J. Stricker

Donald A. Rock

Brent Bridgeman

June 2015

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Stereotype Threat, Inquiring About Test Takers' Race and Gender, and Performance on Low-Stakes Tests in a Large-Scale Assessment

Lawrence J. Stricker, Donald A. Rock, & Brent Bridgeman

Educational Testing Service, Princeton, NJ

This study explores stereotype threat on low-stakes tests used in a large-scale assessment, math and reading tests in the Education Longitudinal Study of 2002 (ELS). Issues identified in laboratory research (though not observed in studies of high-stakes tests) were assessed: whether inquiring about their race and gender is related to the performance of Black and female test takers and, secondarily, whether this association is greater for test takers most identified with math and reading. After high school sophomores completed a questionnaire that included inquiries about their race and gender, only one change in test performance was consistent with expectations from stereotype-threat theory: Black test takers' math scores decreased. Their reading scores and young women's math scores did not decrease, and identification with math and reading did not moderate score decreases for Black test takers or women.

Keywords Stereotype threat; race; gender; Education Longitudinal Study; low-stakes tests

doi:10.1002/ets2.12046

Stereotype threat is a major challenge to the validity of ability and achievement tests, representing a source of construct-irrelevant variance. Stereotype threat is a concern about fulfilling a negative stereotype regarding the ability of one's group (e.g., gender, ethnic, or social class) when this ability is assessed, thereby adversely affecting the performance being evaluated (see Steele, 1997; Steele, Spencer, & Aronson, 2002). This phenomenon has been extensively investigated and repeatedly demonstrated in the laboratory, with a variety of tests and other measures of ability and achievement, but relatively little research has been done in actual testing situations, and nearly all the studies there found no support for stereotype threat (see Inzlicht & Schmader, 2012). All but one of the few investigations of the effects of stereotype threat on performance on operational tests involve high-stakes tests: *SAT*® (Cullen, Hardison, & Sackett, 2004; Cullen, Waters, & Sackett, 2006; Walker & Bridgeman, 2008); *GRE*® General Test (Walters, Lee, & Trapani, 2004); *Advanced Placement*® test and Computerized Placement Test Battery, or *ACCUPLACER*® (Stricker & Ward, 2004); Armed Services Vocational Aptitude Battery (Cullen et al., 2004); and Texas Assessment of Academic Skills, a statewide test of minimum competency (Good, Aronson, & Inzlicht, 2003). Only the latter showed a stereotype-threat effect. The single study with a low-stakes test, National Assessment of Educational Progress (Wei, 2012), obtained equivocal results. (The stereotype-threat manipulation reduced, rather than increased, gender differences on math tests, but it is unclear whether that was because young women's performance was enhanced or young men's was depressed.) The limited research about the impact of stereotype threat on such low-stakes tests in large-scale surveys is a real concern, given the widespread use and importance of these surveys.

This study, like two of those with high-stakes tests (Stricker & Ward, 2004), was built on a Steele and Aronson (1995, Study 4) experiment that showed depressed performance of Black research participants on a verbal test when they were asked about their race immediately prior to working on the test; the performance of White participants was unaffected. Merely asking about race presumably primed stereotype threat for Black participants by making their race salient. (Similar experiments have been done subsequently, substituting socioeconomic status for race, with inconsistent results: Croizet & Claire, 1998, found no stereotype-threat effect, and Spencer & Cantano, 2007, found an effect.) In the Stricker and Ward studies, inquiring about race or ethnicity and gender did not affect the test performance of Black test takers, test takers from other racial and ethnic groups, or women.

The Education Longitudinal Study of 2002 (ELS; Ingels, Pratt, Rogers, Siegel, & Stutts, 2004) provides an opportunity for expanding this line of stereotype-threat research to low-stakes operational tests. The ELS is a longitudinal study of a

Corresponding author: L. J. Stricker, E-mail: lstricker@ets.org nationally representative cohort of students beginning in 2002 with 15,362 high school sophomores in 752 schools (public and private). The test-administration protocol in the 2002 base year (students taking similar math and reading tests before and after they were asked about their race or ethnicity and gender) makes it possible to examine the relationship between this inquiry and performance on the follow-up tests for different subgroups of test takers.

Accordingly, the aim of this investigation was to assess the association between inquiring about race and gender and performance on the ELS math and reading tests. The major hypotheses were that these inquiries (a) are associated with depressed performance of the Black young men on both tests, given the negative stereotype about Black people's intellectual ability in general and (b) are associated with depressed performance of the young women on the math test, given the negative stereotype about females' quantitative ability. A secondary hypothesis was that these relationships will be greater for the test takers who are *identified* with the relevant academic domain (math for the math test and reading for the reading test), staking their self-image on their ability in that domain (Aronson et al., 1999), and hence concerned about the stereotype discrediting a key self-concept. Identification with the domain is an important component of stereotype-threat theory (Steele, 1997; Steele et al., 2002) and an established moderator of the effects of this threat in laboratory experiments (e.g., Aronson et al., 1999; Cadinu, Maass, Frigerio, Impagliazzo, & Latinotti, 2003) though not in the SAT field studies (Cullen et al., 2006; Walker & Bridgeman, 2008).

Method

Sample

The sample for this study, drawn from the ELS base year, consisted of approximately 8,260 Black and White sophomores, from approximately 670 schools, with usable data: about 760 Black young women, 730 Black young men, 3,430 White young women, and 3,340 White young men.[1] Excluded were students who did not take all the tests without special accommodations or were not given the standard questionnaire that included items asking about race or ethnicity and gender.[2]

Tests and Test-Administration Protocol

The tests were given in a group administration to approximately 26 students in each school (Ingels et al., 2004). All students were given the same routing tests: a 15-item math test and a 14-item reading test. After the tests, they completed a questionnaire (described below). The students then took different second-stage tests, depending on their performance on the corresponding routing tests: a low-, middle-, or high-difficulty math test (with 25 to 27 items) and a low-, middle-, or high-difficulty reading test (with 15 to 17 items). For this study, the number-right score for each test was obtained. The coefficient alpha reliability was .84 for the math routing test and .77, .70, and .66 for the low-, middle-, and high-difficulty math second-stage tests, respectively. The corresponding reliabilities for the reading tests were .76, .60, .70, and .65.

Race and Gender Variables

Information about race and gender was obtained from the questionnaire. In the few cases where this information was missing, race and gender were deduced by the ELS staff from other sources (e.g., school roster, student name, parental questionnaire), if possible, or else were imputed by the hot-deck method (Ingels et al., 2004).[3] The questionnaire consisted of 98 multipart items (Items 14 and 17 were gender and race, respectively). The questionnaire mainly covered school experiences and activities, as well as a variety of minor topics, including intrinsic interest and self-efficacy concerning math and reading.

Domain-Identification Measures

Domain-identification measures, modeled after those used previously (see Smith & White, 2001), were derived for this study from the intrinsic interest and self-efficacy items concerning math and reading on the questionnaire.[4] The math measure consisted of three items about the importance of math (e.g., "Because doing mathematics is fun, I wouldn't want to give it up," *strongly agree* [1][5] . . . *strongly disagree* [4]) and four items about self-efficacy concerning math (e.g., "I'm certain I can understand the most difficult material presented in math texts," *almost never* [1] . . . *almost always* [4]). The reading measure was made up of corresponding items, three about importance (e.g., "Because reading is fun, I wouldn't want to give it up," *strongly agree* [1] . . . *strongly disagree* [4]) and four on self-efficacy (e.g., "I'm certain I can understand the most difficult material presented in English texts," *almost never* [1] . . . *almost always* [4]).

Missing data for these items were imputed for this study, using data for a larger base-year sample (test takers of all races and ethnic groups, including test takers without item-level test data, N = 14,320). The rate of missing data for the items ranged from 19.1% to 24.8%. Iterative data-augmentation procedures that imputed missing values for continuous variables (Gelman et al., 2013) and dichotomous or ordinal variables (van Buuren, 2007), implemented with Stata 13 software (multiple-imputation option; StataCorp, 2013), used 34 variables (two were sets of dummy variables, race/ethnicity and geographic region) obtained from the tests, student questionnaire, school-administrator questionnaire, and ELS school-sampling data. The variables were overall score for the routing and second-stage math tests, overall score for both kinds of reading tests, race/ethnicity, gender, socioeconomic status, math and reading importance and self-efficacy, study habits, plans to attend college, academic honors, school subjects interesting, liking school, enrolled in special programs (advanced placement, remedial, and bilingual), free-lunch program in school, and school type (geographic region, urban location, and public control). A single imputation for each missing value was drawn from the imputed data set.

Separately for math and reading, total scores for both the importance and self-efficacy items were computed. The two total scores were standardized to give them equal weight, the score for importance was reflected so that high scores for importance and self-efficacy represented high identification with the domain, and the two total scores were summed, yielding a composite score for use in the analysis. The reliability of the (unweighted) composite score (Mosier, 1943), based on the coefficient alpha reliability of each of the component scores, was .90 for math and .91 for reading.

Analysis

The analysis was done separately for the test takers who took the low-, middle-, and high-difficulty second-stage math and reading tests (Ns appear in Table 1). Scores for each routing and second-stage test were standardized within the low-, middle-, and high-difficulty groups to enable comparisons of test takers' relative performance on the routing and second-stage tests. The equivalent of a repeated-measures analysis of variance (ANOVA) was carried out, with three between-subjects variables: gender, race, and domain identification (math or reading, dichotomized at the median for the total sample), and one within-subjects variable, pretest (routing) score – posttest (second-stage) score. Regression analyses with robust standard errors that adjusted for clustering of students within schools (Williams, 2000), implemented with Stata 13 software (cluster option; StataCorp, 2013), were used to obtain the within-subjects portion of the ANOVA and associated tests of simple effects for the pretest – posttest scores.[6]

Results

Math

The ANOVA results for the math scores are summarized in Table 2. The focus is on the within-subjects interaction of race and gender with pretest – posttest scores; these interactions represent differences in pretest – posttest score trends (substantively, changes in relative performance on the tests) associated with these test-taker characteristics.

Low-Difficulty Group

Two interactions were significant. One was for race ($F = 5.58, p = .018, \eta^2_p = .009$). However, neither Black nor White test takers' scores significantly changed between pretest and posttest: for Black test takers ($M = -.25$ and $M = -.32$), $F = 3.39$, $p = .066, \eta^2_p = .006$; for White test takers ($M = .13$ and $M = .16$), $F = 1.49, p = .224, \eta^2_p = .002$). See Figure 1.

The other interaction was for gender ($F = 9.37, p = .002, \eta^2_p = .015$). Young men's scores significantly decreased ($M = -.01$ and $M = -.09$), $F = 5.57, p = .018, \eta^2_p = .009$, but young women's scores did not change ($M = -.12$ and $M = -.07$), $F = 2.19, p = .140, \eta^2_p = .004$). See Figure 2.

Table 1 Sample Composition

	Math difficulty			Reading difficulty		
Group	Low	Middle	High	Low	Middle	High
			Low domain identification			
Black						
Young women	260	90	10	130	190	30
Young men	180	90	10	150	200	40
White						
Young women	720	910	380	210	870	500
Young men	520	690	330	280	950	670
			High domain identification			
Black						
Young women	250	120	30	100	240	80
Young men	260	140	50	70	190	80
White						
Young women	280	590	550	80	710	1,060
Young men	360	670	790	100	470	870
Total	2,830	3,300	2,130	1,120	3,810	3,330

Notes: The component *N*s may not sum to the total *N* because of rounding error (see Footnote 1). The number of schools in the low-, middle-, and high-difficulty math groups were 600, 620, and 530, respectively. The corresponding *N*s for the reading groups were 430, 630, and 590.

Table 2 Summary of Within-Subjects ANOVAs of Math Pretest and Posttests

	Difficulty								
	Low			Middle			High		
Source	F	p	η^2_p	F	p	η^2_p	F	p	η^2_p
Pretest – posttest (P – P) × Race	5.58	.018	.009	11.39	.001	.018	6.00	.015	.011
P – P × Gender	9.37	.002	.015	1.72	.191	.003	1.59	.208	.003
P – P × Domain Identification (DI)	.21	.646	.000	.25	.614	.000	1.45	.230	.003
P – P × Race × Gender	.37	.543	.001	1.36	.244	.002	.13	.721	.000
P – P × Race × DI	.07	.794	.000	.01	.941	.000	.02	.896	.000
P – P × Gender × DI	.59	.441	.001	.01	.923	.000	.04	.841	.000
P – P × Race × Gender × DI	.00	.993	.000	.44	.508	.001	.11	.736	.000

Notes: The *df* is 1 for each interaction in the three groups, and 602, 620, and 526 for the error terms for the low-, middle-, and high-difficulty groups, respectively. None of the *F* ratios is significantly lower ($p > .05$) than 1 (Snedecor, 1946).

Middle-Difficulty Group

A single interaction was significant, again for race ($F = 11.39$, $p = .001$, $\eta^2_p = .018$). Black test takers' scores decreased ($M = -.38$ and $M = -.53$), $F = 9.67$, $p = .002$, $\eta^2_p = .015$, but White test takers' scores did not change ($M = .07$ and $M = .10$), $F = 1.82$, $p = .182$, $\eta^2_p = .003$. See Figure 3.

High-Difficulty Group

Once more, race was the only significant interaction ($F = 6.00$, $p = .015$, $\eta^2_p = .011$). Black test takers' scores decreased ($M = -.27$ and $M = -.64$), $F = 6.50$, $p = .011$, $\eta^2_p = .012$, but White test taker' scores did not change ($M = -.03$ and $M = -.04$), $F = .18$, $p = .676$, $\eta^2_p = .000$. See Figure 4.

Summary

Changes in math scores were limited to decreases in Black test takers' scores in the middle- and high-difficulty groups and a decrease in young men's scores in the low-difficulty group.

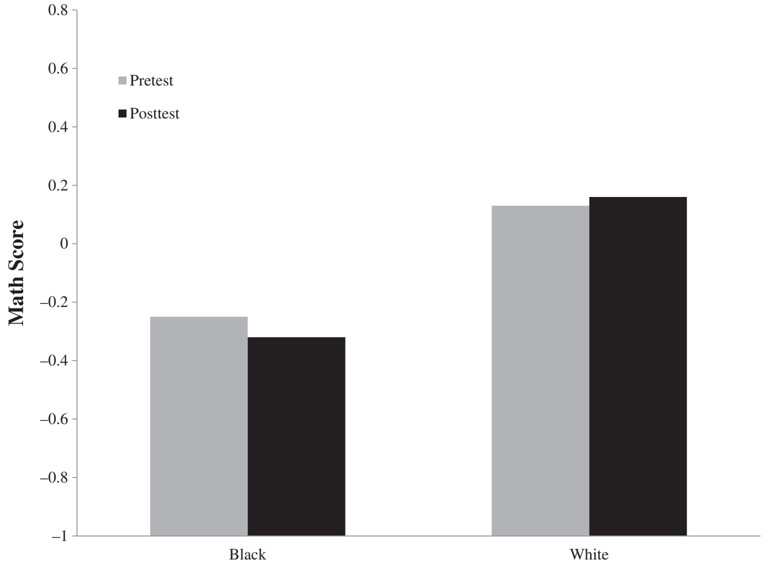

Figure 1 Mean scores on math pretest and posttest for Black and White test takers, low-difficulty group.

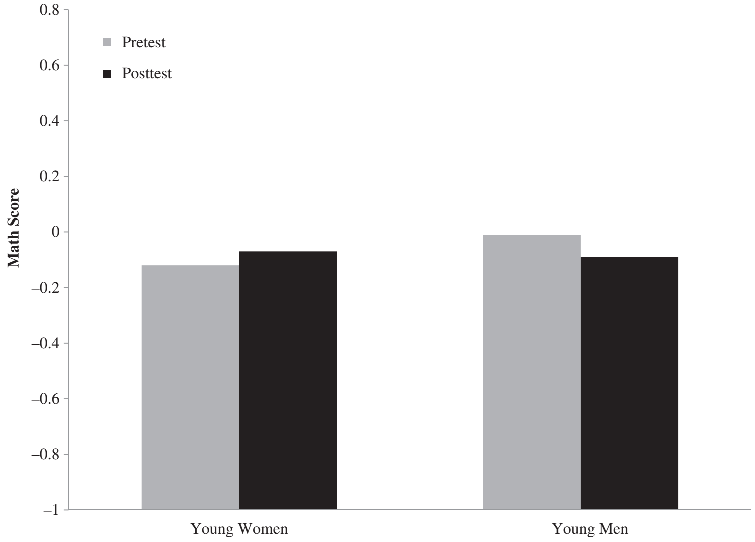

Figure 2 Mean scores on math pretest and posttest for young women and men, low-difficulty group.

Reading

The ANOVA results for the reading scores are summarized in Table 3. Again, the focus is on the within-subjects interactions of race and gender with pretest–posttest scores.

Low-Difficulty Group

Two interactions were significant. One was a second-order interaction of race and gender ($F = 6.79$, $p = .010$, $\eta^2{}_p = .016$). White young women's scores increased ($M = .14$ and $M = .40$), F = 9.55, $p = .002$, $\eta^2{}_p = .022$, but the other subgroups' scores did not change ($p > .05$). (See Figure 5.)

This interaction was qualified by a third-order interaction of race, gender, and domain identification ($F = 7.30$, $p = .007$, $\eta^2{}_p = .017$). Scores of White young women with high domain identification increased ($M = .11$ and $M = .67$), F = 14.14, $p = <.00$ l, $\eta^2{}_p = .032$, but the other subgroups' scores did not change ($p > .05$). (See Figure 6.)

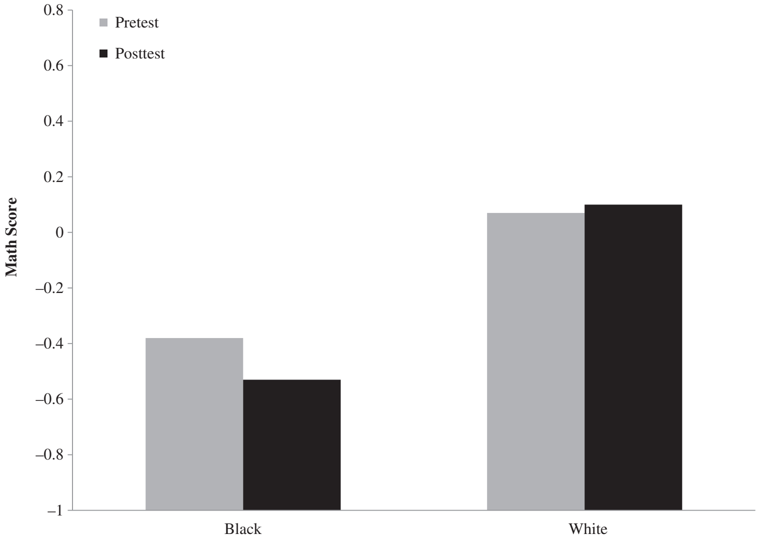

Figure 3 Mean scores on math pretest and posttest for Black and White test takers, middle-difficulty group.

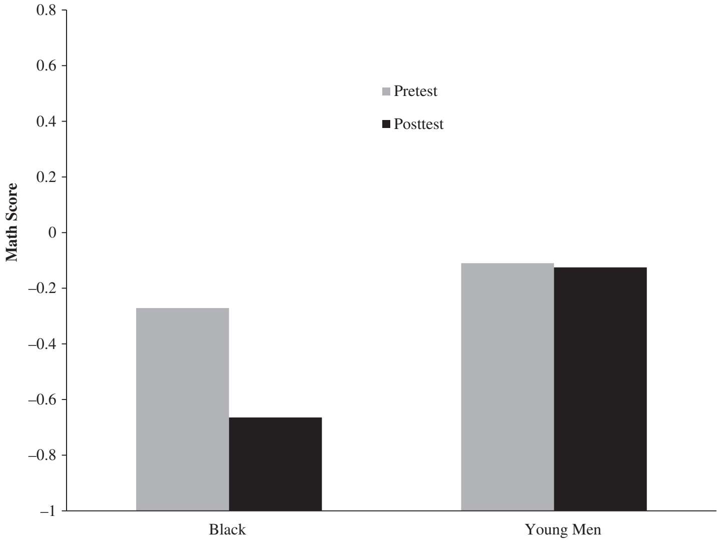

Figure 4 Mean scores on math pretest and posttest for Black and White test takers, high-difficulty group.

Middle-Difficulty Group

One interaction was significant: gender ($F = 45.65$, $p < .00\,l$, $\eta^2_p = .067$). Young women's scores increased ($M = -.11$ and $M = .06$), $F = 22.00$, $p < .001$, $\eta^2_p = .034$, and young men's scores decreased ($M = -.11$ and $M = -.26$), $F = 14.44$, $p < .001$, $\eta^2_p = .022$). (See Figure 7.)

High-Difficulty Group

Three interactions were significant. Again, one was for gender, $F = 7.69$, $p = .006$, $\eta^2_p = .013$). Young men's scores decreased ($M = -.13$ and $M = -.34$), $F = 12.25$, $p < .001$, $\eta^2_p = .020$, and young women's scores did not change ($M = -.18$ and $M = -.12$), $F = .88$, $p < .348$, $\eta^2_p = .001$. (See Figure 8.)

This interaction was qualified by a second-order interaction of gender and domain identification ($F = 4.01$, $p = .046$, $\eta^2_p = .007$). In the high domain-identification group, young men's scores decreased ($M = -.01$ and $M = -.14$), $F = 3.69$, $p = .055$, $\eta^2_p = .006$), and young women's scores increased ($M = -.14$ and $M = .19$), $F = 34.34$, $p < .001$, $\eta^2_p = .055$.

Table 3 Summary of Within-Subjects ANOVAs of Reading Pretest and Posttest

	Difficulty								
	Low			Middle			High		
Source	F	p	η^2_p	F	p	η^2_p	F	p	η^2_p
Pretest – Posttest (P – P) × Race	1.22	.270	.003	.01	.915	.000	.96	.327	.002
P – P × Gender	1.86	.173	.004	45.65	<.001	.067	7.69	.006	.013
P – P × Domain Identification (DI)	8.07	.005	.018	6.79	.009	.011	19.70	<.001	.032
P – P × Race × Gender	6.79	.010	.016	.01	.919	.000	.03	.853	.000
P – P × Ethnicity × DI	1.13	.288	.003	1.76	.186	.003	1.82	.177	.003
P – P × Gender × DI	.18	.674	.000	.00	.973	.000	4.01	.046	.007
P – P × Ethnicity × Gender × DI	7.30	.007	.017	.00	.989	.000	3.68	.056	.006

Notes. The *df* is 1 for each interaction in the three groups, and 431, 633, and 594 for the error terms for the low-, middle-, and high-difficulty groups, respectively. None of the F ratios is significantly lower ($p > .05$) than 1.

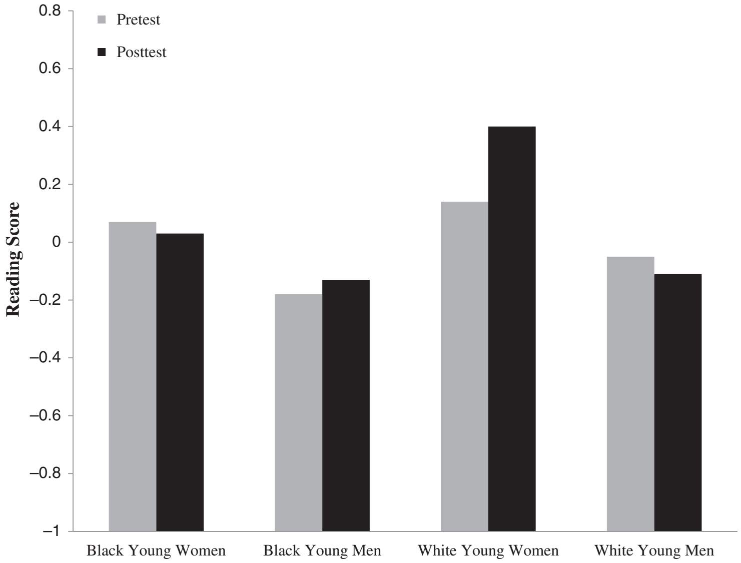

Figure 5 Mean scores on reading pretest and posttest for Black and White young women and men, low-difficulty group.

In the low domain-identification group, young men's scores decreased ($M = -.26$ and $M = -.53$), $F = 8.70$, $p = .003$, $\eta^2_p = .014$, but young women's scores were unchanged ($M = -.23$ and $M = -.43$), $F = 2.92$, $p = .088$, $\eta^2_p = .005$. (See Figure 9.)

This interaction of gender and domain identification was qualified, in turn, by a third-order interaction of race, gender, and domain interaction ($F = 3.68$, $p = .056$, $\eta^2_p = .006$). In the high domain-identification group, scores of Black young women increased ($M = -.34$ and $M = .09$), $F = 15.92$, $p < .001$, $\eta^2_p = .026$, as did those of White young women ($M = .06$ and $M = .30$), $F = 45.16$, $p < .001$, $\eta^2_p = .071$. In the low domain-identification group, scores of White young men decreased ($M = -.12$ and $M = -.41$), $F = 36.48$, p .001, $\eta^2_p = .057$. The other subgroups' scores did not change ($p > .05$). (See Figure 10.)

Summary

Changes in reading scores were complicated but limited. These changes consistently involved score increases for young women: White young women with high domain identification in the low-difficulty group, young women of both races in the middle-difficulty group, and young women of both races with high domain identification in the high-difficulty group. In addition, scores decreased for young men of both races in the middle-difficulty group and for White young men with low domain identification in the high-difficulty group.

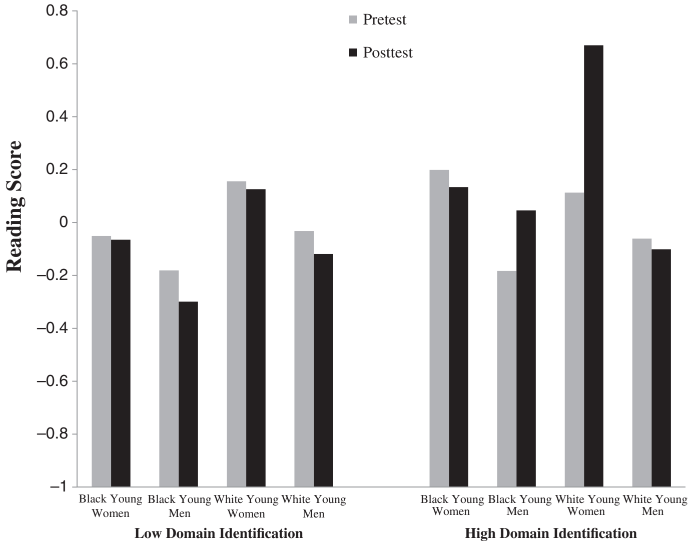

Figure 6 Mean scores on reading pretest and posttest for Black and White young women and men with high and low domain identification, low-difficulty group.

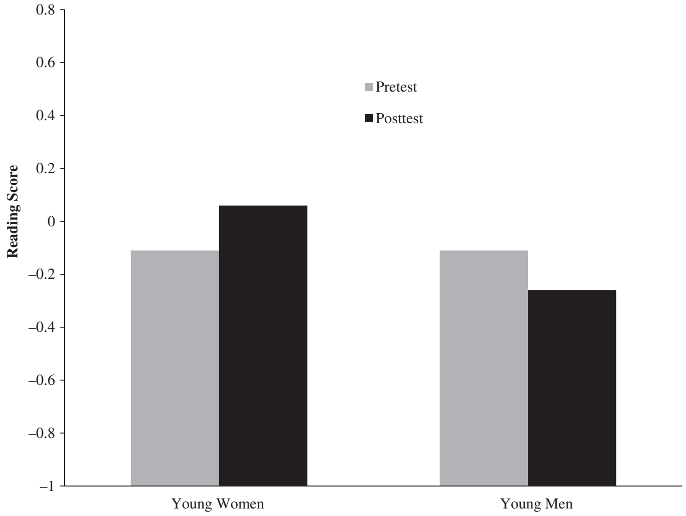

Figure 7 Mean scores on reading pretest and posttest for young women and men, middle-difficulty group.

Discussion

The findings about changes in relative performance on the pretests and posttests are only partly consistent with the expectations stemming from stereotype-threat theory about the adverse effects of inquiring about race and gender on the test performance of Black and female test takers and the moderation of these effects by domain identification. On the one hand, Black test takers' math scores decreased in two of the three ability-level groups after the inquiry. On the other hand, the math scores of young women were not depressed; the reading scores of Black test takers were also not depressed, even improving for young women in two groups; and domain identification did not moderate decreases in math or reading scores for Black test takers or decreases in math scores for young women. The absence of a heightened decrease in Black young women's performance, relative to Black young men, on the math test is especially striking given their *double-minority status* (Gonzalez, Blanton, & Williams, 2002) as targets of negative stereotypes associated with their race (intellectual ability in general) as well as their gender (quantitative ability). The decrease in Black test takers' math scores, in two of the groups, in accord with stereotype-threat theory, is notable. But if stereotype threat was the source of the drop, it is curious that these test takers' reading scores were not also depressed.

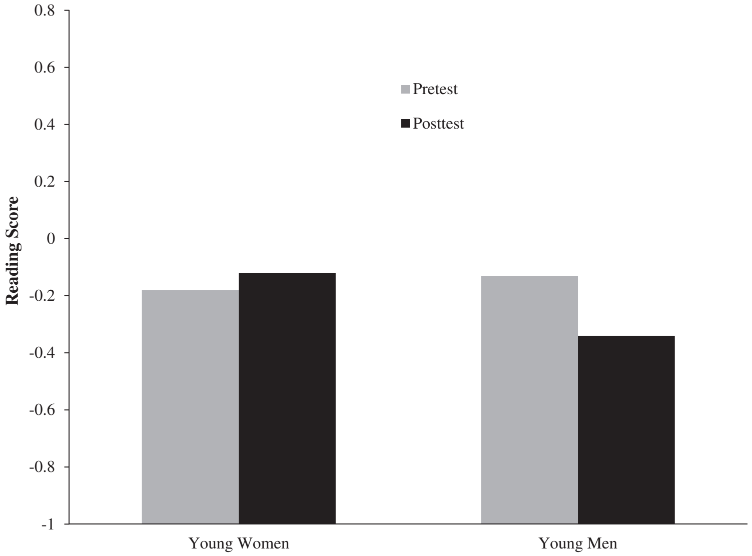

Figure 8 Mean scores on reading pretest and posttest for young women and men, high-difficulty group.

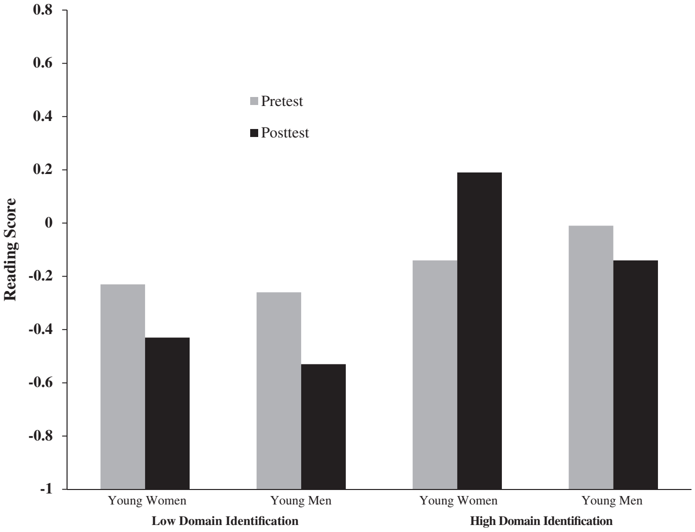

Figure 9 Mean scores on reading pretest and posttest for young women and men with high and low domain identification, high-difficulty group.

All in all, the results regarding inquiries about race and gender were more mixed than the two field experiments that consistently failed to find a stereotype-threat effect (Stricker & Ward, 2004) and most resemble the conflicting results in the laboratory experiments (Croizet & Claire, 1998; Spencer & Cantano, 2007; Steele & Aronson, 1995). And the findings about the inability of domain identification to moderate the stereotype-threat effects are consistent with the outcomes of the SAT field studies (Cullen et al., 2006; Walker & Bridgeman, 2008), not the positive results in several laboratory experiments (e.g., Aronson et al., 1999; Cadinu et al., 2003). What sets this study apart from all the others are the population of participants and the nature of the test: a cross-section of high school students taking low-stakes operational tests.

The findings are intriguing but tenuous, and important caveats about this study should be kept in mind. First, the samples were large (N = 1,120 test takers [430 clusters] to N = 3,810 test takers [630 clusters]), and hence many results were statistically significant at the conventional .05 alpha level but represent small effects that can be over interpreted. For instance, the first-order interaction for race in the low-difficulty math group ($p = .018$) accounts for only .9% of the variance ($\eta^2_p = .009$). Similarly, a simple effect for young men in the same group ($p = .018$) also accounts for .9% of the variance ($\eta^2_p = .009$).

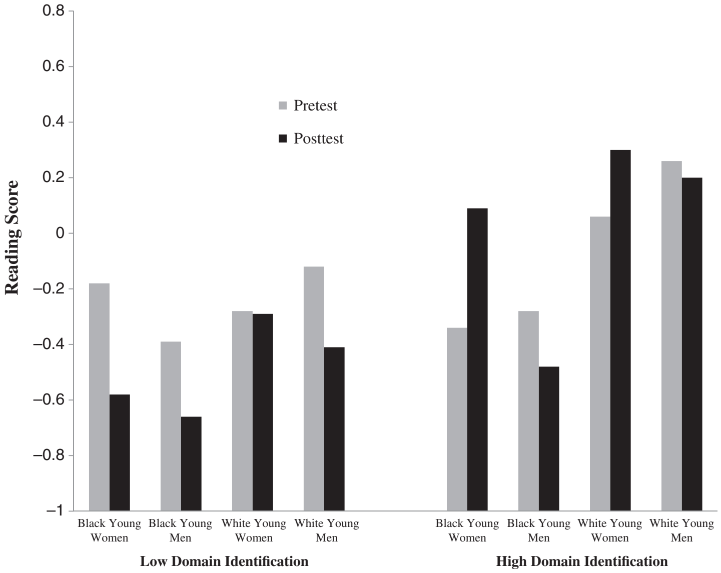

Figure 10 Mean scores of reading pretest and posttest for Black and White young women and men with high and low domain identification, high-difficulty group.

Second, some aspects of the questionnaire content, other than the items about race and gender, may have influenced test performance. Race and gender items had a prominent position at the beginning of the lengthy questionnaire, and there is some evidence of the impact of such questions in previous studies, but the effects of other questionnaire content cannot be ruled out.

Third, assessing the effect of stereotype threat from the difference between pretest and posttest performance necessarily confounds this effect with practice and fatigue effects, and the latter may not be uniform across the race and gender subgroups. Whether practice and fatigue effects were sufficiently potent to overwhelm any stereotype-threat effects in this situation is uncertain.

Clearly, settling these issues and obtaining a definitive answer about the consequences of stereotype threat for low-stakes operational tests call for experimentation. The Stricker and Ward (2004) Advanced Placement experiment, in which test takers were asked about their race and ethnicity and gender, either before or after the test, would be a useful model. Conducting full-dress experiments with operational tests is daunting, but so is the possibility that the usefulness of large-scale surveys that rely on such low-stakes tests is being sapped by stereotype threat. Establishing that stereotype threat is operating on these tests would signal the need to ameliorate this problem and modify the interpretation of the test results. Equally important, from a scientific perspective, experimental results with low-stakes operational tests would also contribute to our understanding of the boundary conditions for this phenomenon.

Acknowledgments

Thanks are due to Elise Christopher, Samantha Guff, and Catherine Millett for assistance in obtaining ELS databases; Min Hwei Wang for extracting the data and doing preliminary analyses; Michelle Najarian and Judith Pollack for providing background information about ELS; and Yigal Attali and Heather Buzick for reviewing a draft of this report.

Notes

1 All *N*s are rounded to the nearest 10, an Institute of Education Sciences requirement for access to the restricted-use ELS test data employed in this study (U.S. Department of Education, Institute of Education Sciences, 1996).

2 The ELS was designed to be representative of all high school sophomores in the United States in 2002, after weighting the total sample of 15,362 students for unequal probability of selection and nonresponse. The atypical nature of the present subsample (about 76.6% of the total sample of Black and White sophomores), as a result of excluding students who did not have usable test data or had not been given the standard questionnaire, precluded such weighting.

3 Race/ethnicity was imputed for 7 students and gender for 10 students in the entire base year sample of 15,362.

4 Self-efficacy items about tests were excluded: “I’m confident that I can do an excellent job on my math tests” and “I’m confident that I can do an excellent job on my English tests.”

5 The item scores appear in parentheses.

6 The mean number of test takers in the low-, middle-, and high-difficulty math groups was 4.69, 5.32, and 4.03 respectively. The corresponding means for the reading groups were 2.59, 6.01, and 5.59.

References

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 and Social Psychology*, *35*, 29–46.

Cadinu, M., Maass, A., Frigerio, S., Impagliazzo, L., & Latinotti, S. (2003). Stereotype threat: The effect of expectancy on performance. *European Journal of Social Psychology*, *33*, 267–285.

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.

Cullen, M. J., Hardison, C. M., & Sackett, P. R. (2004). Using SAT-grade and ability-job relationships to test predictions derived from stereotype threat theory. *Journal of Applied Psychology*, *84*, 220–230.

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

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). London, England: Chapman & Hall/CRC.

Gonzalez, P. M., Blanton, H., & Williams, K. J. (2002). The effects of stereotype threat and double-minority status on the test performance of Latino women. *Personality and Social Psychology Bulletin*, *28*, 659–670.

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

Ingels, S. J., Pratt, D. J., Rogers, J. E., Siegel, P. H., & Stutts, E. S. (2004). *Education longitudinal study of 2002: Base year data file user’s manual* (NCES 2004–405). Washington, DC: U.S. Department of Education.

Inzlicht, M., & Schmader, T. (Eds.). (2012). *Stereotype threat: Theory, process, and application*. New York, NY: Oxford University Press.

Mosier, C. I. (1943). On the reliability of a weighted composite. *Psychometrika*, *8*, 161–168.

Smith, J. L., & White, P. H. (2001). Development of the domain identification measure: A tool for investigating stereotype threat effects. *Educational and Psychological Measurement*, *2001*(61), 1040–1057.

Snedecor, G. W. (1946). Query. *Biometrics*, *2*, 56.

Spencer, B., & Cantano, E. (2007). Social class is dead. Long live social class! Stereotype threat among low socioeconomic status individuals. *Social Justice Research*, *20*, 418–432.

StataCorp. (2013). *Stata 13* [Computer software]. College Station, TX: Author

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

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

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

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

U.S. Department of Education, Institute of Education Sciences. (1996). *Restricted-use data procedures manual*. Washington, DC: Author.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, *16*, 219–242.

Walker, M. E., & Bridgeman, B. (2008). *Stereotype threat spillover and SAT scores* (Educational Testing Service Research Report No. 08-28; College Board Research Report No. 2008-02). Princeton, NJ: Educational Testing Service.

Walters, A. M., Lee, S., & Trapani, C. (2004). *Stereotype threat*, *the test-center environment*, *and performance on the GRE General Test* (GRE Board Research Report No. 01-03R). Princeton, NJ: Educational Testing Service.

Wei, T. E. (2012). Sticks, stones, words, and broken bones: New field and lab evidence on stereotype threat. *Educational Evaluation and Policy Analysis*, *34*, 465–488.

Williams, R. L. (2000). A note on robust variance estimation for cluster-correlated data. *Biometrics*, *56*, 645–646.

Suggested citation:

Stricker, L. J., Rock, D. A., & Bridgeman, B. (2014). *Stereotype threat*, *inquiring about test takers' race and gender*, *and performance on low-stakes tests in a large-scale assessment* (ETS Research Report No. RR-15-02). Princeton, NJ: Educational Testing Service. doi:10.1002/ets2.12046

Action Editor: Rebecca Zwick

Reviewers: Yigal Attali and Heather Buzick