CollegeBoard
inspiring minds™

Research Report
No. 2008-2

Stereotype Threat Spillover and SAT® Scores

Michael E. Walker and Brent Bridgeman

www.collegeboard.com

College Board Research Report No. 2008-2
ETS RR-08-28

Stereotype Threat Spillover and SAT® Scores

Michael E. Walker and Brent Bridgeman

The College Board, New York, 2008

Acknowledgments

The authors wish to thank Dan Eignor, Jinghua Liu, Larry Stricker, and Cathy Wendler for reviewing earlier versions of this manuscript.

Michael E. Walker is a principal psychometrician and psychometric manager at Educational Testing Service (ETS).

Brent Bridgeman 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.

The College Board

The College Board is a not-for-profit membership association whose mission is to connect students to college success and opportunity. Founded in 1900, the association is composed of more than 5,600 schools, colleges, universities, and other educational organizations. Each year, the College Board serves seven million students and their parents, 23,000 high schools, and 3,800 colleges through major programs and services in college admissions, guidance, assessment, financial aid, enrollment, and teaching and learning. Among its best-known programs are the SAT®, the PSAT/NMSQT®, and the Advanced Placement Program® (AP®). The College Board is committed to the principles of excellence and equity, and that commitment is embodied in all of its programs, services, activities, and concerns.

For further information, visit www.collegeboard.com.

Additional copies of this report (item #080482549) may be obtained from College Board Publications, Box 886, New York, NY 10101-0886, 800 323-7155. The price is $15. Please include $4 for postage and handling.

© 2008 The College Board. All rights reserved. College Board, Advanced Placement Program, AP, SAT, and the acorn logo are registered trademarks of the College Board. inspiring minds and Computerized Placement Tests are trademarks owned by the College Board. PSAT/NMSQT is a registered trademark of the College Board and National Merit Scholarship Corporation. All other products and services may be trademarks of their respective owners. Visit the College Board on the Web: www.collegeboard.com.

Printed in the United States of America.

Contents

Abstract

A recent study by Beilock, Rydell, and McConnell (2007) suggested that stereotype threat experienced in one domain (e.g., math) triggered by knowledge of a negative stereotype about a social group in that particular domain can spill over into subsequent tasks in totally unrelated domains (e.g., reading). The authors suggested that these findings might have implications for how the ordering of sections on standardized tests such as the SAT® or GRE could affect examinee performance. To test the authors' assertions, this study used data from a recent SAT administration in which either a reading, a math, or a writing section preceded a reading section. Performance on the subsequent reading section of members of a stereotype threatened group (i.e., women) who took the math section first was compared to performance of those who took the reading or writing section first. Results were inconsistent with the stereotype threat spillover hypothesis and serve to justify the warning of Cullen, Hardison, and Sackett (2004) to exercise caution in generalizing lab findings on stereotype threat to operational testing situations.

Introduction

The current study investigates a recent assertion that stereotype threat can not only affect performance on tasks that are the target of the specific stereotype but also spill over into subsequent tasks unrelated to the stereotype (Beilock, Rydell, and McConnell, 2007). Stereotype threat refers to the fear that individuals feel of confirming negative stereotypes about a group to which they belong. This fear can lead to diminished performance on tasks associated with the negative stereotype. However, performance will be affected only when the domain to which the stereotype applies is of primary importance to the individual's self-definition (Steele, 1997). Furthermore, the effects of stereotype threat will be most evident if the threat is made salient in some manner, such as being told that the task measures intellectual ability, or being asked about race/ethnicity prior to performing the task (Steele and Aronson, 1995).

Stereotype threat is a rather robust phenomenon that has been demonstrated repeatedly in laboratory settings; it has been shown to apply to African Americans and Latinos with intellectual tasks and to females with quantitative tasks (See Wheeler and Petty, 2001, for a review). The argument has been made that stereotype threat could account at least in part for the persistent performance deficits for minorities and women on standardized tests of academic achievement and in school (Steele and Aronson, 1995). A related phenomenon has also been well documented in which the performance of a nonthreatened group actually improves in the presence of a condition that is threatening for a different group. In a meta-analysis of 43 studies, Walton and Cohen (2003) found substantial evidence for stereotype lift in the nonthreatened group ($d = 0.24$).

Although the effect of stereotype threat has been repeatedly observed in the laboratory, a demonstration of the theory in operational testing situations has proven more elusive. To test the generalizability of threat induction methods used in laboratory studies to operational testing situations, Stricker and Ward (2004) studied performance on the College Board's Advanced Placement Program® (AP®) Calculus AB Exam and Computerized Placement Tests™ (which include basic skills tests in reading, writing, and mathematics) under two conditions. In the first, a pretest questionnaire asked about examinees' gender and race/ethnicity. In the second condition, no such questionnaire appeared before the test. The researchers hypothesized that, when the questionnaire preceded the test, performance of women and African American students would suffer. The authors found no effects that met their standard for both statistical and practical significance, although the practical implications of their results remain controversial (Danaher and Crandall, in press; Stricker and Ward, in press).

Good, Aronson, and Inzlicht (2003) demonstrated that instructional interventions designed to reduce stereotype threat impacted performance on a state accountability test. It is uncertain whether this test should be considered as closer to a low-stakes laboratory experiment or a high-stakes test. Although the consequences may be high for the school, consequences for individual students are typically low. Even if the test is sometimes used to retain students in a grade, this is not a real threat for most students, especially for the high-ability students for whom stereotype threat is supposed to be greatest.

Cullen and his colleagues attempted to ascertain the role of self-identification in stereotype threat. In one study, Cullen, Hardison, and Sackett (2004) tested predictions from stereotype threat theory about the relationship between test scores and academic and job performance and the role of the individual's identification with the domain in this relationship. They reasoned (following Steele, 1997) that more able students would be more likely to identify with the domain in question. Stereotype threat theory maintains that such identification is necessary for threat-induced deficits to manifest themselves, so that test scores would increasingly underpredict actual performance at higher levels of test performance. The authors found no such evidence of stereotype threat.

In a subsequent study, Cullen, Waters, and Sackett (2006) measured identification more directly by using information from the Student Questionnaire

of the College Board's SAT, a college entrance examination taken by high school students. The questionnaire includes an item asking about intended college major and another asking about the student's certainty about this choice of major. The researchers classified as math-identified all those individuals who indicated a math-related major and who indicated certainty about the chosen major. The researchers classified as English-identified all those who indicated certainty about English as a major. Using these indicators, the authors examined the regression relationship between scores on the SAT mathematics section and grade point average (GPA) in English. They expected that the difference between regression lines for math-identified versus non-math-identified women would be greater than the difference for men. (The authors chose English GPA as the criterion because it should be unaffected by stereotype threat. Thus, any effects of stereotype threat would manifest themselves in altered regression relationships between the two variables.) Again, the researchers found no evidence of the operation of stereotype threat.

One possible reason for the differences in results between the laboratory and the real world could be motivation of the research participants. Cullen et al. (2006) argued that high motivation might help individuals to overcome the effects of stereotype threat. It is just as plausible that pressures on high-stakes tests are already so intense that the relatively minor stereotype threat manipulations result in no differences in performance across experimental conditions.

More recently, Beilock et al. (2007) investigated the mechanisms underlying stereotype threat. They cited Schmader and Johns (2003), who argued that stereotype threat serves to diminish working memory capacity needed to perform a task. Beilock et al. suggested that this may occur because worries about the perception of others and doubts about one's own ability tie up working memory so that it cannot be used to monitor the processes required for performance. They suggested a second possibility that these worries occupy the phonological aspects of working memory.

Beilock et al. (2007) devised a clever experiment in which they administered the same math items to two groups of individuals. The math items were written differently on the page, either vertically or horizontally. The horizontally oriented items require more phonological resources, according to the authors, "because the maintenance and rehearsal of intermediate steps are represented in different forms (p. 258)." By examining the differential effects of stereotype threat on the differently oriented items, the authors were able to test whether stereotype threat operated by targeting phonological working memory. Their results agreed with this basic supposition: The performance deficit in the presence of stereotype threat was greater for the horizontally oriented items than for the vertically oriented items.

In another study that provided the impetus for the current analysis, Beilock et al. (2007) tested an implication of their previous finding. In the fifth study in their series of related experiments, they presented participants with either a verbal or a spatial task after presenting the horizontal math problems under stereotype threat. The researchers found reduced performance in the stereotype threat condition as compared with the control condition for the verbal but not for the spatial tasks. The authors concluded that "in the current work we demonstrate that stereotype threat on a math task impacts performance on subsequent tasks unrelated to the stereotyped domain. As one might imagine, these findings have important implications for how overall performance may be affected by the ordering of sections on tests such as the SAT or the GRE" (p. 274).

We tested this hypothesis directly in an operational setting by taking advantage of the nonconstant ordering of subject matter on the SAT. The SAT is composed of critical reading, mathematics, and writing sections, each of which is subdivided into three separately timed sections. These sections are placed in different orders for each new SAT test form. By comparing a test form in which a reading section followed a mathematics section to a test form in which a reading section followed something other than a mathematics section (either critical reading or writing), this study examined whether negative spillover occurred. Following Cullen et al. (2006), a derived measure of math identification was included.

We hypothesized that the stereotype threat activated by taking a mathematics section would spill over onto a subsequent reading section resulting in reduced scores on the overall critical reading section relative to the reading scores of women who took a less threatening section (writing or critical reading) prior to the criterion critical reading section. Although this hypothesis refers specifically to effects on women, we also separately analyzed effects on men to evaluate the possibility of stereotype lift. The race/ethnicity-related stereotype threat to African American and Latino students should apply to both math and verbal sections, so we would not expect to be able to demonstrate any effects of purely race/ethnicity-related threats with this study. Nevertheless, there is value in determining whether the gender-related threat (or lift) applies equally across race-ethnic groups. Furthermore, it could be that the effects of stereotype threat are cumulative and do not produce a noticeable effect on test performance until a critical level is reached. White women who are threatened only by the math stereotype might fall below this critical level while African American women, who are threatened by both a general academic stereotype and by the gender-specific math stereotype, could be pushed beyond this critical level and show a performance decrement when this joint effect spills over onto the criterion critical reading subsection.

Method

Data

The data were taken from a recent Saturday administration of the SAT. These SAT forms have 10 timed sections: 3 for multiple-choice (MC) items in critical reading, 3 for mathematics MC items, 2 for writing MC items, and one for an essay. There is also a variable section that does not count toward the reported score; it is used for pretesting and other purposes, and can contain MC items in any of critical reading, mathematics, or writing. Test booklets containing the different variable sections are spiraled into the test packets sent to each testing center such that we would expect the groups of examinees receiving the different orders to be randomly equivalent. In every test form, the essay always appears in the first position in the test booklet. The particular form chosen for this study had the variable section in the second position, followed by a critical reading section.

Of the 20 distinct subforms (defined by the contents of the variable section), a mathematics section appeared twice in the second position, writing appeared 7 times, and critical reading appeared 11 times. This ordering of subject matter allows a direct test of the assertion of Beilock et al. (2007) that math stereotype threat can affect performance on subsequent nonmath tasks. Data for a total of 200,963 examinees from this administration were analyzed. Only juniors (187,402; 93 percent of the analyzed population) and seniors (13,561; 7 percent of the analyzed population) taking the test without special accommodations were retained for analysis.[1] The performance of this group was deemed to be fairly representative of the national average across all yearly administrations.

Measures

Criterion

The primary outcome measure in this study was the performance on the part of the critical reading section appearing in the third position. This section contained 24 critical reading items. The items were formula scored, as is the operational practice with the SAT: examinees received 1 point for a correct answer and 0 points for an omitted item. An incorrect response resulted in a deduction of 1/4 point. Thus, the scores on this critical reading section could range from –6 to 24.

Predictors

The main factors of interest in this study were gender, race/ethnicity, and the type of section (critical reading, mathematics, or writing) that preceded the criterion critical reading section. The race/ethnicity categories, taken from the SAT Questionnaire, included Asian, Asian American, or Pacific Islander; black or African American; Mexican or Mexican American; Puerto Rican; Other Hispanic, Latino, or Latin American; and white. For the analyses reported in this paper, categories for Mexican, Puerto Rican, and Other Hispanic were grouped together and labeled "Latino." High school grade point average (HSGPA) was examined as a check on the equivalence of groups.[2]

The population comprised 93,077 (46 percent) men and 107,886 (54 percent) women. Self-reported race/ethnicity of examinees included 1,156 (less than 1 percent) American Indian; 16,388 (8 percent) Asian American; 19,716 (10 percent) African American; 22,490 (11 percent) Latino; 135,147 (67 percent) white; and 6,066 (3 percent) other. A math section appeared in the second position for 20,342 (10 percent) examinees; a critical reading section for 110,836 (55 percent) examinees; and a writing section for 69,785 (35 percent) examinees.

According to stereotype threat theory, individuals who were more identified with a domain would be more likely to be affected by stereotype threat (Steele and Aronson, 1995). We created a measure for this study that used criteria of the same nature as those used by Cullen et al. (2006).[3] Students who indicated on the SAT Questionnaire an intention to major in a math-related field were identified. Included majors were all those in math-related business fields (e.g., accounting, actuarial science, economics, business statistics, finance, insurance, and management information systems and services), computer and information sciences and technology, mathematics education, engineering and engineering technologies, mathematics and statistics, mathematics and computer science, and physical sciences. Of students with these intended majors, those who indicated that they were "very certain" or "fairly certain" of their major were considered math-identified. This procedure resulted in 19,507 examinees (10 percent of the population) being designated math-identified.

[1] This group is considered the standard group for the purposes of data reporting and test score equating.

[2] Initially, HSGPA was used as a covariate in the analyses, to control for differences in ability across groups. However, one reviewer pointed out that HSGPA could itself have been affected by stereotype threat and could therefore contaminate the results. Thus, none of the analyses reported here use HSGPA as a covariate. The removal of HSGPA as a covariate did not change the basic findings.

[3] The number and titles of the college majors included in the SAT Questionnaire have changed since Cullen et al. (2006) conducted their study. Thus, the classifications used in this study could not be identical to that of the previous authors. However, we attempted as they did to label as math-identified all examinees stating a degree of certainty of majoring in any math-related field.

Analyses

Means and standard deviations were computed for the criterion measures, broken down by predictor categories. Effect sizes (standardized differences) and associated standard errors were computed. The reported standardized differences used the correction for bias listed in Hedges and Olkin (1985; pp. 79–81):

$$\tilde{d} = d\left(1 - \frac{3}{4\left(N_1 + N_2 - 2\right) - 1}\right),$$

where N_1 and N_2 are the sample sizes for the two groups. The standardized difference d was computed as

$$d = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{MSE}},$$

where the $\overline{X}_i$ are the sample means for the two groups; and MSE is the error term from the appropriate analysis of variance (ANOVA) including the predictor variables as factors. The estimated standard errors of the standardized differences were based on the formula found in Hedges and Olkin (p. 86):

$$\hat{\sigma}_d^2 = \frac{N_1 + N_2}{N_1 N_2} + \frac{d^2\%}{2\left(N_1 + N_2\right)}.$$

A 2 (gender) by 6 (race/ethnicity) by 3 (type of preceding section) ANOVA assessed group differences in the average score on the reading items in the third position. Because the design is unbalanced in terms of sample sizes, the general linear model approach to ANOVA was employed. Variance explained by each effect in the ANOVA design was estimated using the partial η^2 measure of association:

$$partial\ \eta^2 = \frac{SS_{effect}}{SS_{effect} + SS_{error}}.$$

This analysis was conducted on the entire population of examinees. Because deficits induced by stereotype threat should only manifest themselves in individuals who are identified with the threatened domain, the ANOVA was repeated for the subset of examinees classified as math-identified using the information from the SAT Student Questionnaire.

In an attempt to isolate gender-related stereotype spillover effects from race/ethnicity-related stereotype spillover, we conducted a single planned comparison of the differential performance (across preceding sections) of white females to white males:

$$L = \left[\left(\frac{R_f + W_f}{2}\right) - M_f\right] - \left[\left(\frac{R_m + W_m}{2}\right) - M_m\right],$$

where R_f, W_f, and M_f are the female participants' means for the critical reading section following critical reading, writing, and mathematics sections, respectively; and R_m, W_m, and M_m are the corresponding means for males. The associated variance for the contrast is

$$\sigma_L^2 = MSE \times \left(\frac{1}{N_{Mf}} + \frac{1}{4N_{Rf}} + \frac{1}{4N_{Wf}} + \frac{1}{N_{Mm}} + \frac{1}{4N_{Rm}} + \frac{1}{4N_{Wm}}\right),$$

where N_{Tg} is the sample size associated with preceding section T and gender g. The hypothesis of no gender differences was tested using $t = L/\sigma_L^2$, with degrees of freedom associated with the error term from the ANOVA. A 1-tailed test was used, as only a positive contrast should support the stereotype spillover hypothesis.

Results

As an initial step, we wanted to evaluate the equivalence of the academic ability of groups receiving each of the sections (mathematics, critical reading, and writing) in the second position. The self-reported HSGPA offered a convenient preexisting measure of ability. For each of the groups, the mean HSGPA was 3.04 with a standard deviation of 0.59. Group gender and racial/ethnic composition was also similar across the three preceding sections. These initial analyses offered strong evidence that random assignment resulted in equivalent groups. Thus, any performance differences among groups on the critical reading section in the third test position could more unambiguously be linked to which type of section appeared in the second position.

Figure 1 illustrates the relationship among the criterion critical reading score and the other measures (gender, preceding section, and HSGPA). The figure presents graphs of average performance on the critical reading items, conditioned on HSGPA. The conditional means were smoothed using loglinear models, preserving three univariate moments (Holland and Thayer, 2000). Separate graphs were produced for females and males by content of the section preceding the critical reading section (either critical reading, mathematics, or writing).

According to the stereotype threat spillover hypothesis, we would expect the graph for females taking the preceding mathematics section to be lower than that for the others, perhaps more so for those with higher HSGPA. We would not expect this pattern for males taking the preceding mathematics section. Instead, Figure 1 shows a similar pattern of results across all preceding sections. Males performed uniformly slightly better than females on the critical reading section. Females taking a mathematics section before the critical reading section did no worse than females with preceding sections in critical reading or writing.

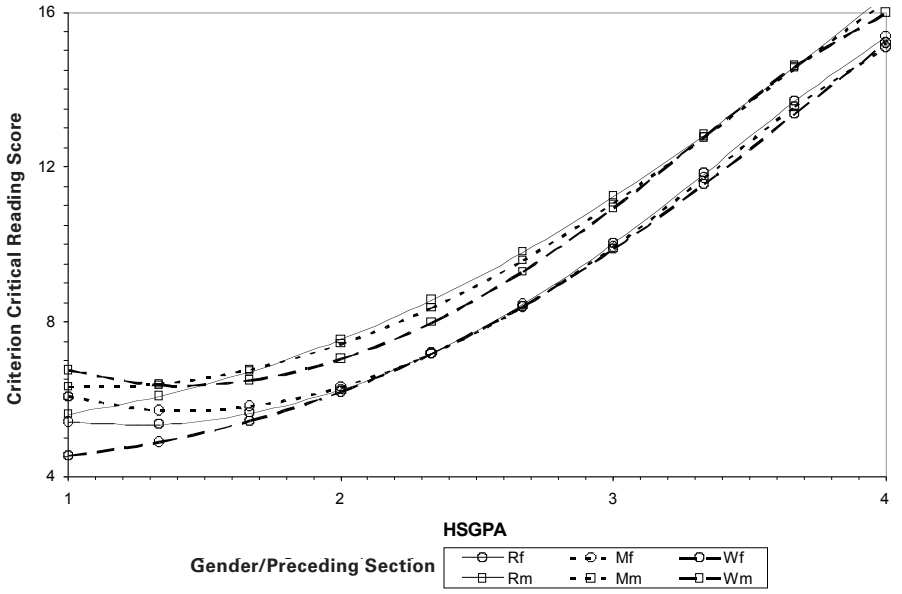

Figure 1. Smoothed critical reading scores conditioned on high school grade point average, by gender and preceding section (critical reading, mathematics, and writing).

Very few students report HSGPAs below 1.5, so the apparent divergence of the lines below this level can be safely ignored.

Table 1 shows average performance on the SAT critical reading section, broken down by gender, race/ethnicity, and the section that preceded the critical reading section on the test. The first thing to notice is for which and how many gender and race-ethnic groups' performances on the critical reading section were lowest when they followed a mathematics section. Of 12 gender and racial/ethnic groups, only two show the lowest mean performance when a mathematics section preceded critical reading. The two groups are Latino females (which might be predicted from the stereotype threat

Table 1

Means for the Critical Reading Score by Gender, Race/Ethnicity, and Preceding Section, with Standardized Differences Among Groups Receiving Different Preceding Sections

Gender	*Race/Ethnicity*	*Preceding Section*	*Sample Size*	*Reading Mean*	*Reading SD*	*Standardized Difference**	*Standard Error*
Female	American Indian	Mathematics	65	10.32	5.38	--	--
		Reading	346	9.68	5.83	–0.113	0.135
		Writing	211	10.51	6.44	0.032	0.142
	Asian American	Mathematics	802	10.16	5.89	--	--
		Reading	4,654	10.39	6.09	0.041	0.038
		Writing	2,923	10.10	6.08	–0.010	0.040
	African American	Mathematics	1,182	7.23	5.28	--	--
		Reading	6,124	6.90	5.24	–0.057	0.032
		Writing	3,878	6.67	5.27	–0.099	0.033
	Latino	Mathematics	1,315	8.26	5.37	--	--
		Reading	7,221	8.33	5.48	0.012	0.030
		Writing	4,438	8.28	5.45	0.006	0.031
	White	Mathematics	7,355	11.81	5.53	--	--
		Reading	39,371	11.97	5.54	0.027	0.013
		Writing	24,548	11.79	5.55	–0.006	0.013
	Other	Mathematics	321	10.83	6.16	--	--
		Reading	1,907	10.70	5.93	–0.022	0.060
		Writing	1,225	10.58	5.92	–0.058	0.063
	Total	Mathematics	11,040	10.74	5.78	--	--
		Reading	59,623	10.83	5.85	0.016	0.010
		Writing	37,223	10.66	5.85	–0.020	0.011

Table 1 (continued)

Means for the Critical Reading Score by Gender, Race/Ethnicity, and Preceding Section, with Standardized Differences Among Groups Receiving Different Preceding Sections

Gender	*Race/Ethnicity*	*Preceding Section*	*Sample Size*	*Reading Mean*	*Reading SD*	*Standardized Difference**	*Standard Error*
Male	American Indian	Mathematics	69	10.97	6.09	--	--
		Reading	286	9.94	5.92	–0.181	0.134
		Writing	179	10.13	5.98	–0.147	0.142
	Asian American	Mathematics	790	10.63	6.28	--	--
		Reading	4,389	11.04	6.35	0.072	0.039
		Writing	2,830	10.87	6.27	0.042	0.040
	African American	Mathematics	887	6.68	5.60	--	--
		Reading	4,574	6.66	5.44	–0.003	0.037
		Writing	3,071	6.49	5.40	–0.033	0.038
	Latino	Mathematics	914	8.75	5.88	--	--
		Reading	5,259	8.92	5.93	0.029	0.036
		Writing	3,343	8.61	5.76	–0.033	0.037
	White	Mathematics	6,352	12.29	5.74	--	--
		Reading	35,282	12.47	5.76	0.032	0.014
		Writing	22,239	12.14	5.80	–0.034	0.014
	Other	Mathematics	290	11.46	6.23	--	--
		Reading	1,423	10.81	6.27	–0.113	0.064
		Writing	900	10.95	6.13	–0.119	0.068
	Total	Mathematics	9,302	11.23	6.09	--	--
		Reading	51,213	11.41	6.11	0.031	0.011
		Writing	32,562	11.09	6.10	–0.032	0.012

* For the standardized differences, each mean is compared to the mean for the group receiving a mathematics section previously. These computations use the pooled variance estimate from the ANOVA shown in Table 2.

spillover hypothesis) and Asian American males (which would appear to contradict the stereotype threat spillover hypothesis).

Table 1 also shows standardized differences. These are computed within each gender and racial/ethnic group, and compare critical reading performance for each preceding section (either critical reading or writing) with performance for the group receiving the mathematics section previously. The standardized differences illustrate very small effects. For the critical comparison when mathematics versus critical reading is the preceding section for female students, the largest standardized difference is only 0.11 (for American Indian females), and this difference is in the opposite direction of the prediction from the stereotype threat spillover hypothesis.

We focused specifically on white examinees, under the supposition that race/ethnicity-related stereotype threat may affect all subject areas and not just math, thereby masking any math-specific stereotype spillover. When we did so, we found that for white females, the observed critical reading mean was indeed lower when the reading section was preceded by a mathematics section than when it was preceded by another critical reading section. However, the difference between the means was trivial, as evidenced by the standardized difference estimates. We also examined the results for white males for any evidence of stereotype lift. This hypothesis might predict superior performance on a section that followed a mathematics section. Here, average critical reading performance was highest when the section was preceded by critical reading and lowest when it was preceded by writing, with average performance after a mathematics section falling squarely in the middle. Thus, there would appear to be no evidence for the operation of any stereotype lift.

For the total sample in Table 1, Table 2 gives the results of the gender by race/ethnicity by preceding task ANOVA.

Table 2

Results of the Gender by Race/Ethnicity by Preceding Section Analysis of Variance with Critical Reading Score As the Dependent Variable, Total Sample

Source	*DF*	*Mean Square*	*F Value*	*Pr > F*	*Partial* η^2
Gender	1	427.6942	13.25	.0003	0.000
Race/Ethnicity	5	80,080.3846	2480.87	<.0001	0.058
Gender x Race/Ethnicity	5	355.7197	11.02	<.0001	0.000
Preceding Section	2	45.2824	1.4	.2459	0.000
Gender x Section	2	10.7548	0.33	.7166	0.000
Racial/Ethnicity x Section	10	60.0145	1.86	.0458	0.000
Gender x Race/ Ethnicity x Task	10	28.2689	0.88	.5553	0.000
Error	200,927	32.279			

The effects of interest involve interactions of gender and race/ethnicity with preceding section, and none of these is statistically significant even in this very large sample. The direct contrast of female and male differences was statistically significant: $L = 0.05$, $t(200927) = 4.61$, $p < .01$. However, the contrast represents an average difference of differences of 1/20 of a score point. Further, the η^2 values show that these gender by section interaction effects account for less than 1/10 of 1 percent of the variation in critical reading scores. These results summarize and reinforce the generally negative trends noted in Table 1.

Although results for the full sample are certainly of interest, effects of stereotype threat spillover should be most evident in the subsample of students who are math-identified. Following Cullen et al. (2006), we classified students who indicated that they intended to major in math-related fields, and who were very or fairly certain of their choices, as math-identified. Table 3 shows the relevant numbers. A total of 19,507 students (10 percent of the sample) stated that they were very certain or fairly certain that they would major in math-related disciplines and were classified as math-identified.

We repeated the analyses illustrated in Tables 1 and 2 using only math-identified participants. Results are shown in Tables 4 and 5. Limiting the analysis to math-identified students does not change the story. Table 4 shows that for more than half of the gender and racial/ethnic groups, the critical reading mean was lowest when the critical reading section followed a mathematics section. Overall, females' observed critical reading mean was lowest with a preceding mathematics section. The same was true for males. However, these differences were small and not practically meaningful. There is no evidence that any gender and racial/ethnic group does significantly worse if the preceding section is mathematics.

Table 5 gives the results of the gender by race/ethnicity by preceding section ANOVA, for math-identified participants only. The effects of interest involve interactions of gender and race/ethnicity with preceding section. As with the total sample, none of these is

Table 3

Numbers of Students Reporting the Intention to Major in Math-Related Subjects and the Students' Degree of Certainty About the Intention

	Certainty			
Intended Major	***Very***	***Fairly***	***Not***	***Total ****
Business Math	994	1,937	788	3,719
Computer and Information Sciences	995	2,189	863	4,047
Mathematics	259	754	473	1,486
Math Education	180	230	64	474
Engineering	3,130	7,304	3,155	13,589
Math and Computer Science	5	5	8	18
Physical Science	492	1,033	637	2,162
Total Math-Related	6,055	13,452	5,988	25,495
Not Math-Related	38,576	56,978	25,237	120,791

* 54,677 examinees did not respond to the question.

Table 4

Means for the Critical Reading Score by Gender, Race/Ethnicity, and Preceding Section, with Standardized Differences Among Groups Receiving Different Preceding Sections, Math-Identified Examinees Only

Gender	***Race/Ethnicity***	***Preceding Section***	***Sample Size***	***Reading Mean***	***Reading SD***	***Standardized Difference****	***Standard Error***
Female	American Indian**	Mathematics	4	15.50	5.45	--	--
		Reading	19	11.53	6.62	--	--
		Writing	8	6.25	7.03	--	--
	Asian American	Mathematics	28	9.79	5.38	--	--
		Reading	230	9.23	6.62	–0.097	0.200
		Writing	159	9.25	6.37	–0.094	0.205
	African American	Mathematics	85	8.15	5.16	--	--
		Reading	352	7.19	5.23	–0.169	0.121
		Writing	202	7.50	5.67	–0.116	0.129
	Latino	Mathematics	78	8.22	5.56	--	--
		Reading	399	8.30	5.39	0.015	0.124
		Writing	199	8.37	5.45	0.027	0.134
	White	Mathematics	269	12.06	5.47	--	--
		Reading	1,396	12.51	5.56	0.080	0.067
		Writing	855	12.33	5.43	0.049	0.070
	Other	Mathematics	8	9.63	8.43	--	--
		Reading	78	11.65	6.15	0.355	0.372
		Writing	48	11.06	6.62	0.253	0.383
	Total	Mathematics	472	10.57	5.76	--	--
		Reading	2,474	10.73	6.02	0.029	0.050
		Writing	1,471	10.73	5.96	0.036	0.053

Table 4 (continued)

Means for the Critical Reading Score by Gender, Race/Ethnicity, and Preceding Section, with Standardized Differences Among Groups Receiving Different Preceding Sections, Math-Identified Examinees Only

Gender	*Ethnicity*	*Preceding Section*	*Sample Size*	*Reading Mean*	*Reading SD*	*Standardized Difference**	*Standard Error*
Male	American Indian**	Mathematics	8	10.50	6.35	--	--
		Reading	47	10.38	5.36	–0.020	0.382
		Writing	32	8.97	6.18	–0.268	0.396
	Asian American	Mathematics	157	10.62	6.33	--	--
		Reading	835	10.45	6.34	–0.030	0.087
		Writing	572	9.95	6.10	–0.118	0.090
	African American	Mathematics	164	6.44	5.44	--	--
		Reading	826	6.68	5.23	0.042	0.085
		Writing	557	6.50	5.24	0.012	0.089
	Latino	Mathematics	155	8.32	5.79	--	--
		Reading	1,003	9.01	5.67	0.121	0.086
		Writing	598	8.58	5.66	0.045	0.090
	White	Mathematics	941	12.47	5.42	--	--
		Reading	5,379	12.75	5.66	0.050	0.035
		Writing	3,409	12.65	5.65	0.032	0.037
	Other	Mathematics	48	9.58	6.45	--	--
		Reading	216	10.42	5.86	0.147	0.160
		Writing	143	10.26	6.26	0.119	0.167
	Total	Mathematics	1,473	11.06	5.99	--	--
		Reading	8,306	11.39	6.05	0.059	0.028
		Writing	5,311	11.17	6.08	0.025	0.029

* For the standardized differences, each mean is compared to the mean for the group receiving a mathematics section previously. These computations use the pooled variance estimate from the ANOVA shown in Table 5.

** Because of the small sample sizes, standardized differences and associated standard errors were not computed.

statistically significant. The η^2 values indicate very small effects.

We also examined the results for white examinees only. Table 4 shows that white males as well as white females performed least well on average on the critical reading section when it followed a mathematics section. The effects were not very large, but they were slightly larger for females than for males. The direct contrast of female and male differences was not statistically significant: $L = 0.13$, $t(19{,}471) = 0.77$, $p = .22$.

Table 5

Results of the Gender by Race/Ethnicity by Preceding Section Analysis of Variance with Critical Reading Score As the Dependent Variable, Math-Identified Examinees Only

Source	*DF*	*Mean Square*	*F Value*	*Pr > F*	*Partial* η^2
Gender	1	17.27007	0.54	.4632	0.000
Race/Ethnicity	5	7,706.89317	240.13	<.0001	0.058
Gender x Race/ Ethnicity	5	134.43743	4.19	.0008	0.001
Preceding Section	2	125.96996	3.92	.0198	0.000
Gender x Section	2	36.73408	1.14	.3184	0.000
Race/Ethnicity x Section	10	44.08822	1.37	.1854	0.001
Gender x Race/Ethnicity x Section	10	25.02238	0.78	.6487	0.000
Error	19,471	32.095			

Discussion

If Beilock et al. (2007) were correct, then performance on critical reading should have suffered when it followed a mathematics section for individuals susceptible to math stereotype threat. This hypothesis would suggest interactions of gender with preceding task, especially for students who were math-identified. These interactions were not statistically significant, and even the nonsignificant differences in the means were frequently in the opposite direction from what would be predicted from stereotype threat theory. There also was no evidence of stereotype lift for male students or for effects that were unique to a particular racial/ethnic group.

When white participants alone were examined in an attempt to isolate gender-specific stereotype threat spillover effects, there was no evidence of stereotype lift for males. Males as well as females performed slightly worse on average on the critical reading section when it followed a mathematics section. This difference was larger for females than for males in the total sample, and the difference was statistically significant. However, the effect was too small to be viewed in any way as meaningful. When only math-identified participants were included in the analyses, the effects were still in the same direction but were no longer statistically significant.

Previous theory and research have indicated that certain conditions must be met in order for stereotype threat to affect performance. First, a negative stereotype must exist. Second, the individual must identify with the threatened domain. Third, the effects of stereotype threat may be heightened by making the threat more salient, either by asking participants about group membership before the test administrator or by mentioning that the test measures cognitive ability. We believe that all of the conditions were met in the current study to activate any stereotype threat for which there was a predisposition. Previous research has supported the existence of negative stereotypes for women with respect to math ability. As is routine with the SAT, all examinees were asked to specify gender just prior to the test administration. This study focused on the participants who indicated an intention to major in math-related fields. Finally, the SAT is widely known to be a high-stakes test of cognitive ability. Thus, although this study did not test for the direct effects of stereotype threat, we would expect such effects to manifest themselves under these conditions. And if there were any negative effects of stereotype threat, according to Beilock et al. (2007), there should also be negative spillover into subsequent sections. And yet no such effects were evident.

In their study, Beilock et al. (2007) used an explicit stereotype threat, reminding female participants of gender differences in math performance. Such a manipulation was of course not used in the current study. We cannot completely rule out the possibility that the difference in findings in the two studies is related to this difference in manipulations. However, we would argue that if such explicit triggers are necessary to observe stereotype threat and subsequent spillover effects, then the authors' assertion that their findings have implications for section ordering on standardized tests is without merit; such manipulations would never be included in a standardized testing situation.

We cannot contest the assertion from Beilock et al. (2007) that performance in a testing situation might be harmed by worry over test performance, using phonological working memory resources that could otherwise be productively used by the test being administered. Indeed, it seems likely that such effects occur, but in a high-stakes testing environment, the worry level might already be so high that stereotype threat plays little or no role. In the laboratory, on the other hand, where there is little intrinsic threat in the task itself, experimentally induced stereotype threat could have a substantial impact on worry level with the concomitant reduction in working memory resources that could be applied to the laboratory task.

Whatever the reason for the generally null results in this real-life test administration, it was not statistical power, as this particular study could detect effect sizes of less than 0.05 standard deviations. Thus, like other studies using data from high-stakes operational settings (e.g., Cullen et al., 2004, 2006; Stricker and Ward, 2004), this study showed little evidence consistent with a stereotype threat hypothesis. This does not mean that such effects do not exist in operational settings. It does mean that effects that appear to be ubiquitous in the laboratory may be much more difficult to demonstrate with real-world test administrations. As Cullen et al. (2004) stated clearly, laboratory results should be generalized beyond the laboratory with caution.

References

Beilock, S. L., Rydell, R. J., & McConnell, A. R. (2007). Stereotype threat and working memory: Mechanisms, alleviation, and spillover. *Journal of Experimental Psychology: General, 136* (2), 256–76.

Cullen, M. J., Hardison, C. M., & Sackett, P. R. (2004). Using SAT-grade and ability-job performance relationships to test predictions derived from stereotype threat theory. *Journal of Applied Psychology, 89*, 220–30.

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* (4), 421–40.

Danaher, K., & Crandall, C. S. (in press). Stereotype threat in applied settings re-examined. *Journal of Applied Social Psychology*.

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,* 654–62.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25* (2), 133–83.

Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology, 85*, 440–52.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52*, 613–29.

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., & Ward, W. C. (2004). Stereotype threat, inquiring about test takers' ethnicity and gender, and standardized test performance. *Journal of Applied Social Psychology, 34* (4), 665–93.

Stricker, L. J., & Ward, W. C. (in press). Stereotype threat in applied settings re-examined: A reply. *Journal of Applied Social Psychology*.

Walton, G. M., & Cohen, G. L. (2003). Stereotype lift. *Journal of Experimental Social Psychology, 39*, 456–67.

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


CollegeBoard
inspiring minds™