The effect of stereotypes on black college test scores at a historically black university

Article *in* Journal of Economic Behavior & Organization · February 2022

DOI: 10.1016/j.jebo.2021.12.005

5 authors, including:

William Darity
Duke University

Catherine Eckel
Texas A&M University

Rhonda V. Sharpe
Women's Institute for Science, Equity and Race

Journal of Economic Behavior and Organization 194 (2022) 408–424

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

The effect of stereotypes on black college test scores at a historically black university

Mackenzie Alston[a], William A. Darity[b], Catherine C. Eckel[c], Lawrence McNeil[d], Rhonda Sharpe[e,*]

[a] Department of Economics, Florida State University, Tallahassee, FL 32306, USA
[b] Department of Economics, Texas A&M University, TAMU 4228, College Station, TX 77843-4228, USA
[c] Sanford School of Public Policy, Duke University, Durham, NC 27708, USA
[d] College of Business, Bowie State University, Bowie, MD 20715, USA
[e] Women's Institute for Science, Equity, and Race, USA

ARTICLE INFO

Article history:
Received 4 November 2020
Revised 23 November 2021
Accepted 5 December 2021

JEL Codes:
C91
J15

Keywords:
Stereotype threat
Lab experiment
Historically black university

ABSTRACT

We conducted lab experiments at a historically black university (HBCU), replicating the design and procedure, but not the results, of previous stereotype threat studies. The experimental design has two factors: stereotype salience (priming) and the identity of the experimenter (a less-threatening black woman vs. a more-threatening white man). Unlike previous studies, we found no effect of stereotype threat on student performance. We find little evidence that black students at the HBCU are affected by stereotype threat, regardless of the identity of the experimenter. We found no significant difference in the number of questions answered correctly by subjects in the control and treatment conditions in either the white male or the black female experimenter sessions. Finally, we found little evidence to support our prediction that subjects would respond differently to the identity of the experimenter. Having a black female experimenter, as opposed to a white male experimenter, had no effect on the number of questions answered correctly.

1. Introduction

According to a 2009 report from the National Center for Education Statistics, 75 percent of white first-time bachelor's students had a cumulative grade point average (GPA) of 3.0 or higher, whereas only 55.3 percent of black students had a similar GPA.[1] Researchers have tried to explain the black-white academic achievement gap through differences in quality of education, parenting, and wealth (e.g., Yeung and Conley, 2008; Hanushek and Rivkin, 2009). In addition, psychologists have explored another factor that could explain why black students are not performing as well as their white peers – stereotype threat. This phenomenon describes the feeling of being at risk of confirming a negative stereotype about one's social group. This feeling can potentially interfere with performance in the domain where the prejudicial belief applies, and it can hold

HBCU, Historically Black Colleges and Universities.

* Corresponding author. malston@fsu.edu
E-mail addresses: malston@fsu.edu (M. Alston), william.darity@duke.edu (W.A. Darity), ceckel@tamu.edu (C.C. Eckel), lmcneil@bowiestate.edu (L. McNeil), rhondavsharpe@wiserpolicy.org (R. Sharpe).

[1] https://nces.ed.gov/datalab/tableslibrary/viewtable.aspx?tableid=8836

https://doi.org/10.1016/j.jebo.2021.12.005
0167-2681/© 2021 Published by Elsevier B.V.

for females and racial groups like Asians (see Shih et al., 1999). Our study focuses on the effect that stereotype threat has on black college students.

A widely held prejudice has it that blacks are not as intelligent as whites, or, at minimum, have cognitive deficits relative to whites. This belief has been promoted in works like *The Bell Curve: Intelligence and Class Structure in American Life* (Herrnstein and Murray 1994), which insinuates that race and IQ are correlated, and papers like Rushton and Jensen (2005). According to the stereotype threat hypothesis, a black student who is primed or subtly reminded of their group membership and the prejudicial beliefs related to their group before completing a relevant task will suffer a decline in performance.

It has been proposed that those affected by stereotype threat either surrender to the stereotype or become so focused on the stereotype that they grow distracted from the task at hand. For example, Blascovich et al. (2001) showed that blood pressure rates of black subjects who had been primed to activate stereotype threat were significantly higher than the blood pressure rates of subjects who had not been primed. Regardless of whether individuals consciously change their effort or not, the main hypothesis involving stereotype threat is that those who are primed will perform worse on the task than those who are not. Consequently, stereotype threat could explain why two equally able individuals who are members of different racial groups may earn different grades.

Steele and Aronson (1995) conducted an experiment to test the impact of stereotype threat on the performance of black college students at Stanford University. They found that priming subjects to think about their race made them more aware of the stereotypes associated with a black identity. In addition, they found that subjects who were primed answered fewer questions correctly on a mock verbal Graduate Record Examination (GRE) than subjects who were not primed. These results were replicated in other studies like Brown and Day (2006) and Davis et al. (2006). Furthermore, these findings implied that test scores might misrepresent the true abilities of black students, especially when they have been exposed to a reminder of their race before completing a test.[2] As a result, white students could appear more intelligent even when compared to black students of the same ability.

The primary goal of our experiment is to determine whether black students from a historically black college respond negatively to stereotype threat on a mock verbal GRE. Almost all of the studies regarding black stereotype threat have been conducted at predominantly white institutions (PWIs) (e.g., Steele and Aronson, 1995; Cohen and Garcia, 2005). While PWIs are the most common type of institution in the United States, these studies fail to consider how stereotype threat may affect black students at institutions that are predominantly black – that is, historically black colleges and universities (HBCUs). Because HBCUs often have racially diverse faculty, it is possible that attending an HBCU and being surrounded by black professors and role models may ameliorate the negative effects of stereotype threat. This prediction is supported by previous studies like Marx and Roman (2002) and Marx and Goff (2005), who found that one way to protect against stereotype threat is to have role models or individuals of authority from one's social group present during the task.

While a handful of similar studies (e.g., McKay et al., 2002, 2003) have used black students at HBCUs as part of their sample, their focus has not been on any differences between students from these schools and PWIs.[3] Therefore, their analysis pools the black students from each type of institution together without treating them differently. In contrast, our study focuses exclusively on black students at an HBCU.

We conducted our experiment in classrooms at an HBCU in Texas. Subjects who participated in the treatment group were primed before completing a set of verbal GRE questions. They were asked to identify their race and ethnicity before the test and read test instructions that emphasized that the verbal GRE was a measure of their intelligence. Subjects in the control group were not asked about their race or ethnicity until after the exam, and their test instructions included no mention of the subjects' verbal ability or intelligence. By comparing the test performance of subjects in the treatment and control groups, we were able to assess the effect of stereotype threat on black students in this setting.

We also varied the level of the threat subjects experienced by using two types of lead experimenters: a white male and a black female (both graduate students from a different Texas university). Steele and Aronson (1995) used a white male experimenter in their studies. However, assuming we succeeded in replicating the original study, we also wanted to include a treatment with a lower threat level to test whether lower threat could mitigate the impact of stereotype threat. Like other studies that have used experimenters of different races (e.g., Marx and Goff, 2005), we predicted that subjects would feel most threatened in the presence of a white male experimenter because he would be seen as a critical outsider by black subjects at an HBCU given that the student body and faculty were predominantly black. Compared to white male experimenters, whose threat arises from both their race and gender, black female experimenters would be seen as least threatening. We anticipated that subjects in the treatment group would have a lower exam score when their experimenter was a white male compared to a black female because subjects with a white male experimenter would be exposed to compound threats (i.e., the presence of the white male experimenter plus the primes from the experiment). Therefore, we

[2] This is the phenomenon of "latent ability." See Walton and Spencer (2009).

[3] Craemer and Orey (2017) conducted an experiment using only black students at HBCUs; however, they manipulated stereotype threat to see if there was any effect on students' political knowledge – not their general or verbal intelligence. They found that black students who were shown a picture of a white research team at the start of the experiment performed worse on the political test than black students who were shown a picture of a black research team.

tested two extremes: a high threat condition with the white male experimenter and a low threat condition with the black female experimenter.[4]

Our sample included over 170 black students at the HBCU. Our findings indicate that the performance of black students at the HBCU is not affected by stereotype threat regardless of the identity of the experimenter. There was some evidence that black students in the control group answered more questions than black students in the treatment group. However, we found no significant difference in the number of questions students answered correctly in either the white male or the black female experimenter sessions. Finally, we found little evidence to support our prediction that subjects responded differently to the type of experimenter.

Previous research showed that the test performance of black students at PWIs is depressed by reminders about the prejudicial belief that blacks are regarded as less intelligent or less achievement-oriented than whites. However, our findings provide preliminary evidence that this is not the case for black students at HBCUs. This is not to say that we conclude that every black high school student should choose to attend an HBCU.

In fact, there is debate over whether attending an HBCU is strictly positive. For example, Fryer and Greenstone (2010) find that blacks who attended HBCUs suffered from a "wage penalty" relative to blacks who graduated from a PWI, but Price et al. (2011) find that black students have higher permanent incomes if they attend an HBCU. We are unable to judge the relative net benefits of graduating from an HBCU. However, our experiment does suggest that black students at HBCUs may, at the very least, have the advantage of being less susceptible to stereotype threat. Thus, this may have potential implications for understanding more about the persistence of the black-white achievement gap. Further research on the benefits of attending an HBCU and on the types of students that choose to attend these institutions may shed light on ways in which the gap can be narrowed.

2. Previous research on historically black colleges and universities (HBCUs)

Most HBCUs were established with the primary mission of advancing the education of blacks under conditions of legal segregation in the United States. As of 2016, there were slightly more than 100 HBCUs in the United States, and these schools accounted for 14 percent of bachelor degree's earned by black undergraduates in 2015 – 2016. This is particularly significant given that HBCUs enrolled 9 percent of all black undergraduates in 2015 and 2016.[5] In addition, the National Science Foundation (2015) reports that about 18 percent of science and engineering degrees are awarded to black students by HBCUs. Therefore, HBCUs provide a unique, albeit understudied, environment to investigate stereotype threat.

Although there are mixed findings, many studies have shown that attending an HBCU affords a positive and beneficial experience for students who select such an institution. Nichols and Evans-Bell, 2017, find that HBCUs account for higher graduation rates for lower-income black students compared to PWIs. Price et al. (2011) utilized a potential outcomes approach to show that graduating from an HBCU has a positive and significant effect on permanent income. They also find that graduating from an HBCU has a positive effect on psychological outcomes like self-image. This confirms the findings of Gurin and Epps (1975) who determined that blacks at HBCUs had positive self-image and high aspirations.

Fries-Britt and Turner (2002) interviewed black students at PWIs and HBCUs to document differences in their experiences in the two environments. Black students at HBCUs reported feeling supported, being a part of the community, finding insulation from discrimination, and having a sense of confidence. On the other hand, black students at the PWIs felt excluded and weighed down by being the "token black" in their classrooms.

The positive impact of HBCUs on student attitudes and perceptions could enhance student performance; for example, studies have shown that self-affirmation improves the test performance of people facing stereotype threat (Martens et al., 2006). It is not unreasonable to then expect that black students attending HBCUs may be better insulated against stereotype threat than black students at PWIs.

If black students at HBCUs respond differently to stereotype threat than black students at PWIs, it could be the consequence of either selection or environment. Black students who choose to enroll in an HBCU could begin their studies with a lower vulnerability to stereotype threat vis-á-vis black students who enroll at PWIs. However, this is not supported by our other work, which is available upon request.[6] In addition, Sparks (2015) surveyed 48 black engineering students at various institutions and found minimal evidence of any differences in stereotype threat vulnerability between students attending HBCUs and PWIs.

If black students at both types of schools initially are equally susceptible to stereotype threat, it is possible that the environment of an HBCU encourages students to ignore or overcome the effects of stereotype threat. For instance, HBCUs typically have more racially diverse faculties than PWIs. In 2016, 61 percent of the faculty at our Texas HBCU were black,

[4] Conditional on finding a treatment effect when comparing the low threat treatment to the high threat treatment, we had intended to complete the full experimental design, measuring the effect of the race and gender of the experimenter by including sessions with white female experimenters and black male experimenters. However, because we failed to find strong evidence that subjects responded to the black female experimenter differently than the white male experimenter, we did not conduct any additional sessions with white female or black male experimenters.

[5] https://nces.ed.gov/fastfacts/display.asp?id=667

[6] We surveyed incoming black freshmen at two schools in Texas – an HBCU and a PWI – and found that they reported similar levels of stereotype threat vulnerability.

whereas only 3 percent of the faculty at our Texas PWI identified as black.[7] Black faculty may provide important same-race role models of someone highly educated, whose presence then offsets the stereotype that blacks are unintelligent or opposed to academic achievement.

The importance of role models has been supported by previous research showing that students receive higher grades when taught by people that physically resemble them (e.g., Dee, 2004; Egalite et al., 2015). There is also evidence in the literature that role models can be influential in overcoming the impact of stereotype threat. For instance, Marx and Roman (2002) showed that when women were primed to think of negative stereotypes, their math scores were higher when there was a competent female experimenter present in the room rather than a male experimenter. Marx and Goff (2005) find a similar result when they primed black subjects and used a black experimenter. Black subjects who had been primed performed just as well as on a verbal test as white subjects when there was a black experimenter. However, they performed worse than white subjects when the experimenter was white.

The content of instruction at HBCUs also might make a difference. The students interviewed in Fries-Britt and Turner (2002) reported appreciating the fact that class lectures highlighted the accomplishments of black scholars and authorities. Correspondingly, McIntyre et al. (2003) found that women who read about successful women performed better on a hard math test. And Aronson et al. (2002) showed that changing black students' mindset about intelligence to believe that intelligence was malleable helped reduce the negative effects of stereotype threat.

Our study answers the question of whether black students at an HBCU are vulnerable to stereotype threat. If any of the factors mentioned above affect black students' vulnerability to stereotype threat, then having more black students attend an HBCU could be an avenue to reduce the black-white achievement gap. These students would be provided with the tools and environment that could decrease the effect of stereotype threat on their performance. Additionally, if specific tools and strategies that successfully mitigate stereotype threat can be identified, it is possible that some of them could be used at PWIs to help black students there.

3. Experimental design

This experiment was conducted using a 2 × 2 design where we manipulated whether or not subjects were primed to think of their group membership and its associated stereotype (control or treatment) and the identity of the experimenter (white male or black female). In all four conditions, subjects received a test booklet at the start of the experiment. Each booklet contained five parts and was modeled after Steele and Aronson (1995) and a more recent set of instructions provided by Aronson.[8]

We chose to expose subjects in the treatment group to two primes in order to increase the intensity of stereotype threat manipulation: We asked them to identify their race prior to the test; and we varied the test instructions, as explained below. In Part One, students in the control group answered six personal questions regarding their age, year in school, major, number of siblings, and mother's and father's highest level of education. Students in the treatment group responded to the same set of questions; however, they also were asked if they identified as Hispanic/Latino, and to identify their race. Subjects in the control group answered these two questions after taking the test. This is similar to the procedure used in Study 4 of Steele and Aronson (1995). The intent was to prime students in the treatment group to think about their race *before* taking the test.

In Part Two, students received instructions that explained what to expect during the test. For those in the treatment group, these instructions acted as a second prime and are similar to what was done in Study 1 of Steele and Aronson (1995). Both Part One and Part Two for the control and treatment groups can be found in the Appendix with the accompanying script. Students in the control group read the following instructions:

> The problems you are about to solve are taken from the verbal portion of the GRE (Graduate Record Examination). You will be given twenty-five minutes to answer 18 questions. You will receive $12 in compensation for submitting your answers. Completing this test will allow you to familiarize yourself with the kinds of problems that appear on tests you may encounter in the future. Please try hard to correctly solve as many items as you can to help us in our analysis of the problem-solving process.

Students in the treatment group read the following instructions:

> The test you are about to take, the verbal portion of the GRE (Graduate Record Examination), is in large part a measure of your verbal intelligence and verbal reasoning ability. You will be given twenty-five minutes to answer 18 questions. You will receive $12 in compensation for submitting your answers. Completing this test will allow you to familiarize yourself with some of your strengths and weaknesses.

Part Three consisted of 18 multiple-choice questions taken from *The* Educational Testing Service, 2014. These questions were identified as easy or medium in difficulty by the authors of the practice book and included a mixture of reading

[7] https://www.pvamu.edu/ir/faculty-data/ and https://accountability.tamu.edu/All-Metrics/Mixed-Metrics/Faculty-Demographics. Note that these statistics on black faculty members do not include international faculty.

[8] The authors would like to thank Joshua Aronson for emailing a copy of his recent instructions.

comprehension questions and vocabulary-based questions.[9] For the reading comprehension questions, students read a passage about a topic like the Antarctic and were asked several questions about what they read. Some of the answers could be found directly in the text while others required critical thinking. The vocabulary-based questions asked students to find one or two words that would fit the missing blank in the provided sentences. These questions were chosen to be challenging because Steele and Aronson (1995) argued that black students would find the stereotype more salient when the test was difficult.

In the final sections of the booklet, we asked subjects to answer several questions eliciting their demographic information and perceptions about the test. We also administered questions to measure their vulnerability to stereotype threat, evaluation of how important their verbal ability is to their identity or self, and the strength of their ethnic/racial identity. These questions were included as part of a pretest for another study. Details are available upon request.

In our white male experimenter sessions, there were two researchers: one white male who acted as the lead experimenter and one black female who acted as his assistant. The lead experimenter was the one who read the instructions, directed the experiment, and had a clear leadership role. On the other hand, the assistant handed out and collected materials, helped pay subjects, and maintained a subordinate role. In our black experimenter session, that same assistant took over as the lead experimenter and another black female acted as her assistant.[10]

Subjects in our experiment were recruited from two sections of an Agricultural Nutrition and Human Ecology course taught by the same professor in Fall 2016, a section of Principles of Microeconomics and a section of Economics and Human Resources taught by a second professor in Winter 2017, and two sections of Principles of Microeconomics taught by the same second professor in Fall 2018. At the end of the experiment, subjects were paid $12 in cash.

Because we conducted our experiment in classrooms during normal class times, we were not able to control the racial composition of the sessions. Consequently, not all of our subjects were black. In total, we had 214 subjects.[11] Of these, 193 identified themselves as black, including fifteen subjects who identified as multiracial.

In our analysis below, we only consider students that identified as black or African-American. If they identified as multiracial, they were included in our analysis so long as one of their self-identified races was black or African-American. Also, because of the nature of the experiment, some subjects came to class late and thus missed the introduction of the experiment. These subjects were always given control booklets but were excluded from the analysis because they did not have sufficient time to complete the booklet. In addition, careful notes were taken of students who were distracted (e.g., using their cell phones during the experiment) and students who skipped ahead. These students were also were excluded from the analysis. This made our final sample size 176: 104 were in the black female experimenter sessions, and 72 were in the white male experimenter sessions.[12] See Table 1 for the demographics of the students who agreed to participate in the experiment. Table 1A in the Appendix describes subject demographics by treatment.

4. Results

If black students at an HBCU respond the same way to stereotype threat as black students at a PWI, then our results should confirm those of Steele and Aronson (1995). Confirmation would mean that students in the control group correctly answered more questions than students in the treatment group.

We further hypothesized that students with the white male experimenter would experience an added level of threat compared to students with the black female experimenter. In other words, the difference in performance between the control and treatment groups would be larger in the white male experimenter sessions than the black female experimenter sessions, and the test scores for subjects in both groups would be lower in the white male experimenter sessions. If our hypotheses were correct, it would be consistent with findings from studies like Marx and Goff (2005), who found that stereotype threat was not as severe when there was a black rather than white experimenter.

To test our hypotheses, we ran three main sets of regressions for each dependent variable: number of correct answers and accuracy. For Tables 3 and 4, Column 1 displays the estimates from a regression where we measured the effect of the primes on all of the subjects in our final sample, and Column 2 adds control variables. Next, we added a dummy variable in Column 3 called *Black Female Experimenter*, which equaled one if the student had a black female experimenter. Column 4

[9] Steele and Aronson (1995) reported using 27 difficult verbal questions from GRE study guides. In a pilot study we conducted, we used 27 verbal questions from GRE practice exams. On average, the control group answered 17.2 questions but only 1.6 were correct (n = 5). The primed treatment group answered 24.5 questions. Of these, 3.67 questions were answered correctly (n = 6). We wanted subjects to correctly answer a higher proportion of questions, so we reduced the number of questions and the difficulty level of the questions used. Ultimately, our goal was to calibrate the question difficulty for subjects in our control group to correctly answer a similar percentage of the test questions as in Steele and Aronson (1995). Steele and Aronson (1995) wanted students to answer approximately 30 percent of the test correctly, and the average score for all students in the control group in our final sample was 28 percent.

[10] All experimenters and assistants were from a predominantly white institution. The students participating in the study were aware that this project was a joint project between the PWI and the HBCU. When the experimenters were introduced to the class, there was no emphasis made that they were from a PWI. More details on the procedure are in the Appendix.

[11] One subject participated in the study twice. Data from the second time he participated in the study were excluded from all analysis.

[12] Analysis with all black subjects who arrived on time (but may have skipped ahead or been distracted) can be found in the Appendix. The main results from this analysis are similar to the ones presented in the paper.

Table 1
Subject demographics.

	Full sample (1)	Blacks only (2)	Final sample (3)
Percent Blacks	90.19%	100%	100%
Percent Non-Hispanic Whites	1.40%	0%	0%
Percent Hispanic	8.88%	2.59%	2.84%
Percent Female	51.87%	50.78%	51.14%
Average Age	20.55	20.47	20.43
Average Year in School	2.59	2.61	2.60
Average Cumulative GPA	3.02	3.01	3.02
N	214	193	176

Notes: Anyone who indicated that they were black or African American, including those who identified as multiracial and also selected another race, were counted as black. Note that fifteen subjects identified as black or African American and at least one other race. The final sample includes blacks who came on time, did not skip ahead, and were not distracted. Year in School is 1 for freshmen, 2 for sophomores, 3 for juniors, and 4 for seniors. Four subjects never reported their race. Some other variables (e.g., age and cumulative GPA) were not reported for all subjects. For race, ethnicity, and gender, the percent was calculated as the number of observations fitting that description divided by N. Note that if a subject did not identify his/her race, ethnicity, and/or gender, he/she would appear in the denominator but not the numerator.

Table 2
Treatment effects for final sample.

Panel A: Pooled Sessions				
	Control Group	Treatment Group	p-value for two-sample t-test	p-value for Mann-Whitney test
Average Questions Answered	17.82	17.49	0.03	0.05
	(0.59)	(1.26)		
Average Correct Answers	5.02	4.96	0.89	0.95
	(2.81)	(2.50)		
Average Accuracy (%)	28.27	28.42	0.95	0.75
	(15.96)	(13.96)		
N	91	85		
Panel B: Black Female Experimenter Sessions				
	Control Group	Treatment Group	p-value for two-sample t-test	p-value for Mann-Whitney test
Average Questions Answered	17.81	17.53	0.14	0.15
	(0.59)	(1.22)		
Average Correct Answers	4.75	4.80	0.92	0.90
	(2.81)	(2.46)		
Average Accuracy (%)	26.86	27.48	0.83	0.68
	(16.21)	(13.81)		
N	53	51		
Panel C: White Male Experimenter Sessions				
	Control Group	Treatment Group	*p*-value for two-sample t-test	*p*-value for Mann-Whitney test
Average Questions Answered	17.84	17.44	0.10	0.18
	(0.59)	(1.33)		
Average Correct Answers	5.39	5.21	0.77	0.87
	(2.80)	(2.59)		
Average Accuracy (%)	30.24	29.83	0.91	0.94
	(15.61)	(14.27)		
N	38	34		
Panel D: All Subjects				
	Black Female Session	White Male Session	*p*-value for two-sample *t*-test	*p*-value for Mann-Whitney test
Average Questions Answered	17.67	17.65	0.89	0.58
	(0.96)	(1.02)		
Average Correct Answers	4.78	5.31	0.20	0.17
	(2.63)	(2.69)		
Average Accuracy (%)	27.16	30.04	0.21	0.15
	(15.01)	(14.89)		
N	104	72		

Notes: Standard deviations reported in parenthesis. Questions answered and correct answers have a maximum value of 18. Accuracy has a maximum value of 100.

Table 3
OLS estimates for correct answers.

	(1)	(2)	(3)	(4)
Treatment	−0.06	−0.19	−0.19	−0.14
	(0.40)	(0.43)	(0.43)	(0.68)
Black Female Experimenter			0.27	0.31
			(0.85)	(0.95)
Treatment * Black Female Experimenter				−0.08
				(0.92)
Age		0.04	0.04	0.04
		(0.07)	(0.07)	(0.07)
Female		0.60	0.60	0.59
		(0.44)	(0.44)	(0.45)
Year in School		0.01	0.01	0.01
		(0.27)	(0.27)	(0.29)
GPA		0.84*	0.84*	0.84*
		(0.47)	(0.47)	(0.48)
Constant	5.02***	1.74	1.47	1.44
	(0.29)	(2.12)	(1.91)	(1.94)
N	176	163	163	163

Notes: Robust standard errors in parentheses. Specifications 2 through 4 include session effects. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 4
OLS estimates for accuracy (%).

	(1)	(2)	(3)	(4)
Treatment	0.15	−0.74	−0.74	−0.19
	(2.26)	(2.41)	(2.41)	(3.76)
Black Female Experimenter			−3.12	3.56
			(4.78)	(5.36)
Treatment * Black Female Experimenter				−0.96
				(5.09)
Age		0.15	0.15	0.15
		(0.39)	(0.39)	(0.40)
Female		3.60	3.60	3.50
		(2.45)	(2.45)	(2.49)
Year in School		0.25	0.25	0.31
		(1.53)	(1.53)	(1.63)
GPA		5.05*	5.05*	5.06*
		(2.64)	(2.64)	(2.65)
Constant	28.27***	11.22	8.10	7.79
	(1.67)	(12.04)	(10.74)	(10.90)
N	176	163	163	163

Notes: Robust standard errors in parentheses. Specifications 2 through 4 include session effects. $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

adds an interaction term called *Treatment*Black Female Experimenter*, which captures the differential effect of having a black female experimenter for those in the treatment group.

For the number of correct answers and accuracy, we expected the coefficient on *Black Female Experimenter* to be positive, which would suggest that students with a white male experimenter performed worse on the exam because they felt a higher level of threat in the presence of the white male experimenter. We also predicted that the interaction term would be positive. This would imply that stereotype threat was more detrimental for those who had a white male experimenter rather than a black female experimenter. The results of our analysis are presented below.

4.1. Results: effect of stereotype threat on correct answers

Before turning to an analysis of the treatment effect on correct answers, we first consider the impact on questions answered. It is possible that being exposed to the primes caused subjects in the treatment group to answer fewer questions. They may have found it pointless to disprove the stereotype and thus purposely chose to exert less effort, or they may have unconsciously been distracted by the stereotype and unable to answer as many questions. We found that 90 percent of subjects in the control group and 80 percent of subjects in the treatment group answered all eighteen questions. The distribution of questions answered for the treatment group is significantly different from that of the control group (Mann-Whitney, $p = 0.05$). Subjects in the control group answered an average of 17.82 questions, and this difference is statistically

significant from the treatment group's average of 17.49 (t-test $p = 0.03$). Therefore, there was a small, but statistically significant, treatment effect on questions attempted when studying the pooled sample.

However, when examining the black female and white male experimenter sessions separately, we do not find a significant difference between the number of questions attempted by subjects in the control group and the treatment group (t-test $p = 0.14$ and $p = 0.10$, respectively). These results are reported, along with other summary statistics of the treatment effects, in Table 2.[13] Table 3A in the Appendix reports the average marginal effects from a fractional logit for the proportion of questions answered. In the final specification, the effect of being in the treatment group loses significance.[14]

Our primary outcome of interest was the effect of stereotype threat on correct answers. On average, the control group correctly answered 5.02 questions, and the treatment group correctly answered 4.96 questions. This difference is not statistically significant (t-test $p = 0.89$); there is no evidence that priming had a negative effect on scores. This is true even when studying the black female experimenter and while male experimenter sessions separately. As seen in Table 2, the averages of the control groups and treatment groups are similar.

To estimate the effect of priming on students' tests scores, we ran an ordinary least squares (OLS) model using the number of correct answers as the dependent variable in Table 3. The dummy variable *Treatment* is small and insignificant, even after adding control variables. This means that there is no evidence of stereotype threat. In addition, the type of experimenter had no effect on test scores or how the treatment group responded to the prime. The only variable that predicted the number of correct answers was GPA. Those with a higher GPA answered more questions correctly.

4.2. Results: effect of stereotype threat on accuracy

Because the two groups responded to different numbers of questions, we compared each group's accuracy, which we calculated as the number of correct answers divided by the number of questions answered. The average control group's accuracy is 28.27%, which is not statistically distinguishable from the treatment group's average of 28.42% (t-test $p = 0.95$). The average accuracy is similar for both groups when only looking at the black female experimenter sessions (t-test $p = 0.83$) and when only looking at the white male experimenter sessions (t-test $p = 0.91$).

The results from Table 4 complement the results from Table 3. Mainly, we find that priming black students had no significant effect on their accuracy and neither did the type of experimenter. Furthermore, students in the treatment group were just as accurate, relative to the control group, when they had a white male experimenter as they were when they had a black female experimenter. The only significant predictor of accuracy is GPA. Therefore, these findings fail to support the idea that the test scores of black students at HBCUs are affected by stereotype threat. They also suggest that these students are impartial to the presence of white men when taking difficult tests under threat.[15]

5. Discussion

Based on our results, we fail to find strong evidence that black students at HBCUs react to stereotype threat. Apart from the HBCU environment, several factors may have contributed to our null finding. First, one may question whether our primes were effective; perhaps students did not respond to stereotype threat because the negative stereotype about blacks was never primed. Recall that we chose to expose subjects in the treatment group to two primes: We asked them to identify their race prior to the test; and we varied the test instructions. The primes that we used were taken from Steele and Aronson (1995) and similar studies, all of which found that the treatment group's average test score was lower than that of the control group.

The main prime (presented to the subjects as part of the instructions) was modeled after the first study reported in Steele and Aronson (1995). That study included three primes: diagnostic, nondiagnostic, and challenge. In the diagnostic treatment, subjects were told that researchers wanted to study "various personal factors involved in performance on problems requiring reading and verbal reasoning abilities." In addition, subjects were told that they would be given feedback that "may be helpful to [them] by familiarizing [themselves] with some of [their] strengths and weaknesses." Steele and Aronson (1995) chose this language to prime black subjects to think about the stereotype about blacks' intelligence and to potentially start to worry about the possibility of confirming this stereotype. For this reason, we used very similar language in the test instructions for subjects in the treatment group.

In the nondiagnostic condition of Study 1, verbal abilities were not mentioned. The test was framed as a way for researchers to study the "psychological factors involved in solving verbal problems," and the feedback was framed as a way for subjects to become familiar "with the kinds of problems that appear on tests [they] may encounter in the future." Steele and Aronson (1995) decided on this wording because they believed that it would not trigger the negative stereotype

[13] Table 2A in the Appendix reports the same summary statistics but includes subjects who skipped ahead and who were distracted. The overall findings are consistent with those in Table 2.

[14] Table 4A repeats this analysis; however, it includes subjects who skipped ahead and who were distracted. The overall findings in the last specification are consistent with those in the last specification of Table 3A.

[15] OLS estimates for correct answers and accuracy for the sample that includes subjects who went ahead early and/or were distracted are reported in Tables 5A and 6A in the Appendix. The results are consistent with those in Tables 3 and 4.

about blacks' intelligence. Following this same logic, we used similar language when describing the test to subjects in the control group.

Steele and Aronson (1995) also conducted a study, Study 3, specifically to determine whether subjects in the treatment group had negative stereotypes on their mind. They measured stereotype activation by exposing a new group of subjects to either the prime used in the diagnostic condition or the ones used in the nondiagnostic condition and asked them to complete a fill-in-the-blank task. Subjects saw missing letters from a word and had to provide letters that would turn the fragment into a word. Black subjects in the diagnostic condition submitted more race-related words (e.g., minority, lazy) than black subjects in the nondiagnostic condition ($p < 0.01$). This showed that black subjects were more conscious of race and stereotypes when they were in the diagnostic condition and suggests that priming through test instructions is effective. This study supports our use of the diagnostic and non-diagnostic primes as our main manipulation.

To strengthen the treatment, we chose to expose subjects to an additional prime besides the main prime in Steele and Aronson (1995). This prime is taken from Study 4 of Steele and Aronson (1995). In this study, they tested whether stating one's race prior to completing the test would "prime the racial stereotype about ability for Black participants, and thus make them stereotype threatened." They found that black subjects in the treatment group performed significantly worse than black subjects in the control group ($p < 0.02$). Thus, adding the requirement that subjects first indicate their race before completing the test should reinforce the main manipulation.

Other studies (e.g., Davis et al., 2006; McKay et al., 2002) have used similar primes in their stereotype threat experiments and found a significant treatment effect with similar sample sizes. Therefore, we have no reason to believe that our null effects are the result of ineffective primes. To assess the effectiveness of the primes in our experiment, we tested whether the treatment led to significant differences in several of our post-experiment questions. First, we tested whether the prime affected strength of identity using the Multigroup Ethic Identity Measure (MEIM- see Phinney, 1992). This measure tests the strength of their racial or ethnic identity. We found no significant difference in MEIM scores between the treatment and control groups. However, we find some support for differences in response to the test. In particular, subjects in the treatment group perceived the test to be more biased than subjects in the control group ($p = 0.05$), which could suggest that the treatment group found connections between the test and negative stereotypes about blacks that the control group did not. The treatment group also scored marginally higher on the survey measure of stereotype threat vulnerability. The tests are reported in Appendix Table 7A.

Second, sample size could be an additional concern with our study. As noted, similar studies have found evidence of stereotype threat with similar sample sizes. Based on our best estimates of the mean and standard deviation for correct answers in Study 1 of Steele and Aronson (1995), we estimated that we would need approximately 20 black subjects in each of our treatment groups for a power of 0.80 to detect a treatment effect of the magnitude found in the original study (one standard deviation). Steele and Aronson (1995) had 114 black and white subjects in Study 1 who were randomized into one of three treatments. As such, approximately 38 subjects (black and white, combined) were in each treatment. Our study has a larger sample size, but of course an even larger sample would have allowed a more precise inference. Given the size of our treatment effect is close to zero, even a much larger sample size is very unlikely to lead to a significant finding.

Third, one could also imagine that the incentive structure could explain why we found null effects. In economics, experiments are traditionally designed so that subjects' payments depend on choices they make during the experiment. In our experiment, subjects were paid a flat rate of $12, irrespective of how many questions they answered or their overall score. We chose this incentive scheme to replicate Steele and Aronson (1995).[16] Subjects in their main study also earned a fixed rate of $10 in exchange for participating in the experiment. We first want to note that our subjects did not appear to shirk the task. When subjects handed in their booklets, the experimenters could not easily see how much of the test they completed. Therefore, a subject could choose to leave the test entirely blank and turn in her booklet without concern that the experimenter would immediately notice and/or confront her. We found that subjects attempted 17.66 out of 18 questions and reported guessing on only 4.70 questions, on average. Second, to the best of our ability, we tried to monitor and screen people who did not take the experiment seriously. If anyone appeared to be distracted or was not listening to the instructions and jumped ahead, they were excluded from our main analysis. Lastly, we note that Fryer et al. (2008) conducted a pilot study to see if subjects would respond differently if they were paid $2 per correct answer compared to if they were paid a flat fee of $20. The targeted group in their study was women, and they did not find any significant difference in performance between primed women in the piece rate treatment and women in the flat rate treatment.[17] For these reasons, we do not believe that the flat fee payment scheme explains why we failed to find evidence of stereotype threat.[18]

Another concern could be that stereotype threat is a phenomenon of the past. Steele and Aronson (1995) was published over two decades ago, and many things have changed since then. However, more recent studies (e.g., Marx and Goff, 2005,

[16] After replicating Steele and Aronson (1995), we intended to include new treatments where we incentivized subjects to maximize their number of correct answers. We hoped to study whether paying subjects per correct answer would exacerbate or reduce the effect of stereotype threat. However, because we failed to replicate Steele and Aronson (1995), we did not add these treatments.

[17] Indeed, they show that the un-threatened group, men, responded positively to both the prime and to incentives, producing the only significant difference in their analysis: an interaction effect between threat and incentives, which together increased men's performance without affecting women's.

[18] We were able to locate only one study examining stereotype threat with race, using an incentivized test. However, in this study there is no baseline, non-incentivized condition for comparison. The study finds the standard stereotype threat result with incentives (McFarland et al., 2003).

Brown and Day, 2006) have also found evidence of stereotype threat. As such, studies published in the twenty-first century have found that stereotype threat still harms people.

One remaining explanation for the null findings could be that black students at HBCUs respond differently to stereotype threat than black students at PWIs, who are the subjects most commonly used in stereotype threat studies. If the type of black student who chooses to attend an HBCU is different from the type who chooses to attend a PWI, then there is reason to believe that this difference in personal characteristics, backgrounds, and/or preferences could lead to differences in their natural resilience to stereotype threat. However, when we compared the level of stereotype threat vulnerability between incoming black freshmen at a HBCU to ones at a PWI in a separate survey we conducted, we did not find a significant difference.[19] That suggests that HBCU students' resilience is built up over time, and this could be because of the unique schooling environment HBCUs provide. For instance, HBCUs often have diverse faculty, and Marx and Roman (2002) have shown that competent role models that share your identity can reduce the effect of stereotype threat. As black students at HBCU spend more time around diverse instructors and curriculum, they could develop a resistance to the negative effects of stereotype threat. Therefore, our results could be the result of black students at HBCUs becoming more resilient to stereotype threat as they spend more years in school and gain exposure to black instructors, a supportive and inclusive environment, etc.

6. Conclusion

Stereotype threat is a term used to describe the feeling of being at risk of confirming a negative stereotype about one's social group. Steele and Aronson (1995) used black students at Stanford University to show that the performance of students on a mock verbal GRE fell when they were reminded, subtly, of the stereotype that blacks are not as smart as whites. Other research has produced similar findings (e.g., Brown and Day, 2006; Davis et al., 2006). Yet the majority of these studies focus on the effect of stereotype threat on black students attending predominantly white institutions.

In our experiment, we examine the role of stereotype threat on the verbal GRE test scores of black students attending a historically black university. There is evidence in economics and other fields that supports the idea that black students at HBCUs may benefit from this type of environment. For example, HBCUs are known for having a diverse faculty, and studies like Egalite et al. (2015) show that young black students have higher reading and math test scores when they have a black teacher. At an HBCU, black professors could serve as a valuable presence for students who are coping with the effects of negative stereotypes about their intelligence or academic motivation. Furthermore, Fries-Britt and Turner (2002) mention that students attending HBCUs are more likely to have the opportunity to learn about the scholarly achievements of other blacks. Research has shown that stereotype threat is reduced when people read about successful people from their identity group (McIntyre et al., 2003), so that is another way in which HBCUs may make a difference. Altogether, these factors could reduce the negative effects of stereotype threat for black students at HBCUs.

We conducted our experiment to test this idea with students at a historically black university in Texas. Students in the treatment group were primed to think about the negative stereotype about blacks' intelligence before answering questions on a mock verbal GRE. Students in the control group were not primed before taking the exam.

We also varied the level of threat by changing the identity of the lead experimenter. We chose to start by comparing two extremes. In some sessions, the lead experimenter was a black female (low threat); in others, the lead experimenter was a white male (high threat). According to standard hypotheses about the effects stereotype threat, students in the treatment group should have worse test scores than students in the control group. Additionally, we predicted that students with a white male experimenter would answer fewer questions correctly than students with a black female experimenter.

None of these predictions were confirmed in this study. In both the black female experimenter sessions and the white male experimenter sessions, we found little evidence that stereotype threat negatively affects students' test score. Students in the control group correctly answered a similar number of questions as students in the treatment group and had similar accuracy. We also failed to find strong evidence to support the idea that students' test scores were affected by the identity of the experimenter. The type of experimenter had no significant effect on the number of questions students answered correctly.

Thus, we find preliminary evidence that vulnerability to stereotype threat may be diminished for black students at HBCUs. If that is the case, then studying the recruitment strategies, teaching methods, on-site resources, and general environment of HBCUs could teach us how to prepare black students to deal with negative stereotypes about their intellect. In turn, when they are confronted with these biases, their performance may not be affected, and their grades may better reflect their latent ability. This may help reduce America's persistent black-white achievement gap. Therefore, we recommend that future research should investigate differences between HBCUs and PWIs and the students who choose to attend them and study how these differences affect the performance of black college students.

[19] Note that respondents in this survey were not intentionally primed; therefore, this measure of stereotype threat vulnerability should reflect their level of vulnerability when they are not exposed to any manipulations. Because of the timing of the survey, the measure should also show their level of vulnerability before they had the opportunity to be fully immersed in the culture and learning environment of their HBCU.

Acknowledgments

This research was supported by a collaborative grant from the National Science Foundation [SES-1530796 (Eckel, Sharpe, Alston), SES-1530730 (Darity), SES-1530746 (McNeil)]. We are grateful to Jeff Woolridge for econometric advice. We also thank Cheryl Mitchell for administrative support; Ryan Rholes and Joshua Witter, who acted as the "white male" experimenters; Gbenga Ojumu and Alfred Parks at Prairie View A&M University, for letting us use their classes to collect data; and Chandon Adger and Bethany Patterson for data assistance. We thank Heather Lench, Adrienne Carter Sowell, and Jane Sell at Texas A&M for comments and guidance. The paper benefitted from feedback of participants in seminars at the Department of Sociology at Texas A&M and presentations at the 2016 American Society of Hispanic Economists (ASHE) Conference, 2017 Economic Science Association (ESA) Conference, 2017 American Economic Association (AEA) Summer Mentoring Pipeline Conference, and 2018 Southern Economic Association (SEA) Conference. All remaining errors are our own.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jebo.2021.12.005.

Appendix: Experimental materials

SCRIPT

Thank you for participating in today's experiment!

Your seat is labeled with a number that corresponds with the number card you were given before. Please find your seat.

WAIT FOR THEM TO GET A SEAT

Thank you for participating in today's experiment!

Before we get started, please turn off your phones and put all of your belongings on the ground beside you. [In a moment, we will pass out booklets./As you may have noticed, there is a booklet on your desk.] It is really important for the experiment that you remain quiet and keep your eyes on your booklet for the duration of the experiment.

Your responses will have no effect on your course grade; however, you will have the opportunity to earn money during this experiment, which I will explain in a moment.

To start, please open your booklet.

WAIT FOR THEM TO OPEN THEIR BOOKLET

There is a blank line on the first page. Please write your subject ID number there. This is the same number that you were given earlier and that is on your desk.

WAIT FOR THEM TO WRITE THEIR NUMBERS

As with many assessments, your packet may differ from the packet that others receive. There is no need to be concerned about this. Simply follow the instructions written in your packet.

You may now turn the page.

WAIT FOR THEM TO TURN THE PAGE TO PART ONE: QUESTIONNAIRE

Okay, you will now be asked to fill out this questionnaire. Once you have finished the questionnaire, please wait until further instruction.

WAIT FOR THEM TO FILL OUT PART ONE: QUESTIONNAIRE

You may now turn the page to read the test instructions. We will give you three minutes to do so. Please read all of the instructions carefully, and then I will let you know when it is time to turn the page.

WAIT THREE MINUTES FOR THEM TO READ PART TWO: INSTRUCTIONS

You may now turn the page. You will have twenty-five minutes to answer the questions. We will keep track of the time remaining on the board.

START TIME, WRITE TIME REMAINING "25 MINUTES" ON BOARD, AND UPDATE TIME EVERY 5 MINUTES

Okay, time is up for the test. Please turn to Part Four. Once you have finished Part Four, please continue on to Part Five. You may begin to answer the questions now.

WAIT FOR THEM TO FILL OUT PART FOUR: QUESTIONNAIRE AND PART FIVE: OPINION SURVEY

All right. Thank you again for participating in today's experiment. We will call you up one by one to receive payment by your subject ID number. Please wait patiently at your desk until your number has been called. When we call your number, please bring your test booklet and subject ID card with you.

If people are still working on it and class time is ending soon:

Alright everyone! It is [time]. If you have finished answering the questions, feel free to come up to the front to turn in your booklet. Please bring your subject ID card with you, too. Otherwise, if you're still working on the questions, feel free to take your time.

If anyone asks about their score:

Hang around after we pay everyone, and we can discuss this more.

Hand graded test but don't let them keep it

Control part one

(1) What is your age? _____
(2) What year are you in school?
  a. Freshman
  b. Sophomore
  c. Junior
  d. Senior
  e. Graduate student
(3) What is your major? _____
(4) How many siblings do you have? _____
(5) What is your mother's highest level of education?
  a. No high school completed
  b. Some high school, no diploma
  c. High school graduate
  d. Some college, no degree
  e. Vocational degree
  f. Associate's degree
  g. Bachelor's degree
  h. Master's degree
  i. Doctoral degree
  j. I don't know.
(6) What is your father's highest level of education?
  a. No high school completed
  b. Some high school, no diploma
  c. High school graduate
  d. Some college, no degree
  e. Vocational degree
  f. Associate's degree
  g. Bachelor's degree
  h. Master's degree
  i. Doctoral degree
  j. I don't know.

Control part two

The problems you are about to solve are taken from the verbal portion of the GRE (Graduate Record Examination). You will be given twenty-five minutes to answer eighteen questions. You will receive $12 in compensation for submitting your answers. Completing this test will allow you to familiarize yourself with the kinds of problems that appear on tests you may encounter in the future.

Please try hard to correctly solve as many items as you can to help us in our analysis of the problem solving process.

Treatment part one

(1) What is your age? _____
(2) What year are you in school?
  a. Freshman
  b. Sophomore
  c. Junior
  d. Senior
  e. Graduate student
(3) What is your major? _____
(4) How many siblings do you have? _____
(5) What is your mother's highest level of education?
  a No high school completed
  b Some high school, no diploma
  c High school graduate
  d Some college, no degree
  e Vocational degree
  f Associate's degree
  g Bachelor's degree

h Master's degree
i Doctoral degree
j I don't know.

(6) What is your father's highest level of education?
a No high school completed
b Some high school, no diploma
c High school graduate
d Some college, no degree
e Vocational degree
f Associate's degree
g Bachelor's degree
h Master's degree
i Doctoral degree
j I don't know.

(7) Do you identify as Hispanic/Latino?
a Yes
b No

(8) How do you identify? **Circle all that apply.**
a White/Caucasian
b Black/African American
c East Asian
d Middle Eastern
e South Asian
f Native American
g Pacific Islander
h Other (Please specify.) _____

Treatment part two

The test you are about to take, the verbal portion of the GRE (Graduate Record Examination), is in large part a measure of your verbal intelligence and verbal reasoning ability. You will be given twenty-five minutes to answer eighteen questions. You will receive $12 in compensation for submitting your answers. Completing this test will allow you to familiarize yourself with some of your strengths and weaknesses.

It is absolutely vital that you try to do your best on this test. Please try hard to correctly solve as many items as you can to help us in our analysis of your verbal ability.

Appendix: Additional procedural information

Once the lead experimenter and assistant arrived at the classroom, the professor briefly introduced them before exiting the room. Because he left the room, students knew that their professor was not observing them. They also were explicitly told that their participation was not related to their class grades. After the professor left, the researchers asked the students to take their belongings and move into the hallway while one researcher remained in the room to set up the experiment.

The researcher inside the classroom shuffled the test booklets for the control and treatment groups together. Then the researcher randomly placed one numbered card and one test booklet on each of the student desks. In two of the sessions, there were enough desks so that booklets could be placed on every other desk, leaving an empty desk in between. When this was not possible, booklets were sometimes placed on two consecutive desks. At the same time, the researcher outside of the classroom made it clear that students' participation in the study was purely voluntary and passed out numbered cards to any student who chose to participate.

As soon as the room was ready, students were readmitted and asked to find the desk with the card number that matched the one that they had been given in the hallway. This seating arrangement was intended to reduce the probability that students were sitting next to close friends with whom they might exchange information during the experiment.

Appendix: Additional tables

Table 1A
Subject demographics by experimenter type (Final sample).

	Black Female Experimenter (1)	White Male Experimenter (2)	*t*-test for Difference (3)
Percent Hispanic	1.92%	4.17%	$p = 0.38$
Percent Female	53.85%	47.22%	$p = 0.39$
Average Age	20.04	21.00	$p = 0.04^{**}$
Average Year in School	2.38	2.93	$p < 0.01^{***}$
Average Cumulative GPA	3.14	2.86	$p < 0.01^{***}$
N	104	72	

Notes: p-values report the results of a two-sample *t*-test with equal variances. Age ($n = 70$), year in school ($n = 71$), and GPA ($n = 70$) were not reported for all subjects in the White Male sessions. GPA ($n = 95$) was not reported for all subjects in the Black Female sessions. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 2A
Treatment effects summary statistics for black sample who arrived on time.

	Control Group	Treatment Group	*p*-value for two-sample *t*-test	*p*-value for Mann-Whitney test
Panel A: Pooled Sessions				
Average Questions Answered	17.78	17.52	0.09	0.10
	(0.76)	(1.24)		
Average Correct Answers	5.02	4.94	0.84	0.90
	(2.78)	(2.47)		
Average Accuracy (%)	28.32	28.27	0.98	0.85
	(15.75)	(13.80)		
N	94	89		
Panel B: Black Female Experimenter Sessions				
	Control Group	Treatment Group	*p*-value for two-sample t-test	*p*-value for Mann-Whitney test
Average Questions Answered	17.73	17.56	0.37	0.28
	(0.86)	(1.19)		
Average Correct Answers	4.77	4.80	0.95	0.95
	(2.76)	(2.43)		
Average Accuracy (%)	27.02	27.39	0.90	0.80
	(15.85)	(13.65)		
N	56	54		
Panel C: White Male Experimenter Sessions				
	Control Group	Treatment Group	*p*-value for two-sample t-test	*p*-value for Mann-Whitney test
Average Questions Answered	17.84	17.46	0.11	0.20
	(0.59)	(1.31)		
Average Correct Answers	5.39	5.17	0.72	0.82
	(2.80)	(2.56)		
Average Accuracy (%)	30.24	29.61	0.86	0.99
	(15.61)	(14.11)		
N	38	35		

Note: The standard deviations in parenthesis. Questions answered and correct answers have a maximum value of 18. Accuracy has a maximum value of 100.

Table 3A
Average marginal effects from fractional logit for proportion of questions answered.

	(1)	(2)	(3)	(4)
Treatment	-0.02^{**}	-0.02^{*}	-0.02^{*}	−0.02
	(0.01)	(0.01)	(0.01)	(0.01)
Black Female Experimenter			−0.03	-0.04^{*}
			(0.02)	(0.02)
Treatment * Black Female Experimenter				0.01
				(0.02)
Age		0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)
Female		0.00	0.00	0.00
		(0.01)	(0.01)	(0.01)
Year in School		−0.01	−0.01	−0.01
		(0.01)	(0.01)	(0.01)
GPA		−0.01	−0.01	−0.01
		(0.01)	(0.01)	(0.01)
N	176	163	163	163

Notes: Average marginal effects of fractional logit regressions reported. Specifications 2 through 4 include session effects. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 4A
Average marginal effects from fractional logit for proportion of questions answered for black sample who arrived on time.

	(1)	(2)	(3)	(4)
Treatment	−0.01	−0.01	−0.01	−0.02
	(0.01)	(0.01)	(0.01)	(0.02)
Black Female Experimenter			−0.04*	−0.05**
			(0.02)	(0.02)
Treatment * Black Female Experimenter				0.02
				(0.02)
Age		0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)
Female		−0.01	−0.01	−0.01
		(0.01)	(0.01)	(0.01)
Year in School		0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)
GPA		0.00	0.00	0.00
		(0.01)	(0.01)	(0.01)
N	183	170	170	170

Notes: Average marginal effects of fractional logit regressions reported. Specifications 2 through 4 include session effects. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 5A
OLS estimates for correct answers for black sample who arrived on time.

	(1)	(2)	(3)	(4)
Treatment	−0.08	−0.18	−0.18	−0.15
	(0.39)	(0.42)	(0.42)	(0.67)
Black Female Experimenter			0.04	0.06
			(0.81)	(0.92)
Treatment * Black Female Experimenter				−0.04
				(0.89)
Age		0.03	0.03	0.03
		(0.07)	(0.07)	(0.07)
Female		0.56	0.56	0.55
		(0.42)	(0.42)	(0.42)
Year in School		0.03	0.03	0.03
		(0.27)	(0.27)	(0.29)
GPA		0.94**	0.94**	0.95**
		(0.46)	(0.46)	(0.47)
Constant	5.02***	1.42	1.38	1.37
	(0.29)	(2.09)	(1.87)	(1.91)
N	183	170	170	170

Notes: Robust standard errors in parentheses. Specifications 2 through 4 include session effects. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 6A
OLS estimates for accuracy (%) for black sample who arrived on time.

	(1)	(2)	(3)	(4)
Treatment	−0.06	−0.74	−0.74	−0.24
	(2.19)	(2.34)	(2.34)	(3.67)
Black Female Experimenter			2.01	2.41
			(4.54)	(5.16)
Treatment * Black Female Experimenter				−0.86
				(4.94)
Age		0.10	0.10	0.10
		(0.39)	(0.39)	(0.40)
Female		3.42	3.42	3.33
		(2.34)	(2.34)	(2.37)
Year in School		0.32	0.32	0.37
		(1.49)	(1.49)	(1.59)
GPA		5.57**	5.57**	5.58**
		(2.57)	(2.57)	(2.60)
Constant	28.32***	9.42	7.41	7.10
	(1.62)	(11.83)	(10.45)	(10.67)
N	183	170	170	170

Notes: Robust standard errors in parentheses. Specifications 2 through 4 include session effects. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 7A
Perceptions of test, multigroup ethnic identity measure, and stereotype threat vulnerability.

	Control Group	Treatment Group	*p*-value for two-sample *t*-test	*p*-value for Mann-Whitney test
Perceived Test Difficulty	4.47	4.56	0.61	0.64
	(1.14)	(1.11)		
Perceived Test Bias	3.07	3.64	0.05	0.06
	(1.86)	(1.97)		
Multigroup Ethnic Identity Measure	3.36	3.34	0.82	0.59
	(0.43)	(0.37)		
Stereotype Threat Vulnerability	28.00	30.04	0.08	0.06
	(7.47)	(7.72)		

Note: Standard deviations are in parentheses. Perceived test difficulty and perceived test bias were rated on a scale of one (not at all) to seven (extremely). Ninety subjects in the control group and eighty-five subjects in the treatment group rated the test difficulty. Eighty-nine subjects in the control group and eighty-three subjects in the treatment group stated how biased they thought the test was. Subjects were asked to rate how strongly they agreed with twenty statements on a scale of one (strongly disagree) to four (strongly agree). The multigroup ethnic identity measure depends on the average of their responses to a subset of these questions. A higher number indicates a higher racial/ethnic identity measure. Eighty-one subjects in the control group and seventy-three subjects in the treatment group completed the measure. Subjects were asked to rate how strongly they agreed with seven statements on a scale of one (strongly disagree) to seven (strongly agree) to measure their stereotype threat vulnerability. A higher number indicates a higher vulnerability to stereotype threat. Ninety subjects in the control group and eighty-one subjects in the treatment group completed this measure.

References

Aronson, J., Fried, C.B., Good, C., 2002. Reducing the effects of stereotype threat on African American college students by shaping theories of intelligence. J. Exp. Soc. Psychol. 38 (2), 113–125.

Blascovich, J., Spencer, S.J., Quinn, D., Steele, C., 2001. African Americans and high blood pressure: the role of stereotype threat. Psychol. Sci. 12 (3), 225–229.

Brown, R.P., Day, E.A., 2006. The difference isn't black and white: stereotype threat and the race gap on raven's advanced progressive matrices. J. Appl. Psychol. 91 (4), 979–985.

Cohen, G.L., Garcia, J., 2005. "I am us": negative stereotypes as collective threats. J. Pers. Soc. Psychol. 89 (4), 566–582. doi:10.1037/0022-3514.89.4.566.

Craemer, T., Orey, D.A., 2017. Implicit black identification and stereotype threat among African American students. Soc. Sci. Res. 65, 163–180.

Davis III, C., Aronson, J., Salinas, M, 2006. Shades of threat: racial identity as a moderator of stereotype threat. Journal of Black Psychology 32 (4), 399–417.

Dee, T.S., 2004. Teachers, race, and student achievement in a randomized experiment. Rev. Econ. Stat. 86 (1), 195–210.

Educational Testing Service, 2014. *Official GRE verbal reasoning practice questions. with practice for the analytical writing measure.*, 1. McGraw-Hill, New York.

Egalite, A.J., Kisida, B., Winters, M.A., 2015. Representation in the classroom: the effect of own-race teachers on student achievement. Econ. Educ. Rev. 45, 44–52.

Fries-Britt, S., Turner, B., 2002. Uneven stories: successful black collegians at a black and a white campus. Rev. High Ed. 25 (3), 315–330.

Jr Fryer Jr., R.G., Greenstone, M, 2010. The changing consequences of attending historically black colleges and universities. Am. Econ. J. Appl. Econ. 2 (1), 116–148.

Jr Fryer Jr., R.G., Levitt, S.D., List, J.A, 2008. Exploring the impact of financial incentives on stereotype threat: evidence from a pilot study. Am. Econ. Rev. 98 (2), 370–375.

Gurin, P., Epps, E., 1975. Black consciousness, identity, and achievement: a study of students in historically black colleges.. John Wiley and Sons, New York.

Hanushek, E.A., Rivkin, S.G., 2009. Harming the best: how schools affect the black-white achievement gap. Journal of Policy Analysis and Management 28 (3), 366–393.

Herrnstein, R.J., Murray, C., 1994. The Bell curve: The reshaping of American life By Differences in Intelligence. The Free Press, New York.

Martens, A., Johns, M., Greenberg, J., Schimel, J., 2006. Combating stereotype threat: the effect of self-affirmation on women's intellectual performance. J. Exp. Soc. Psychol. 42 (2), 236–243.

Marx, D.M., Goff, P.A., 2005. Clearing the air: the effect of experimenter race on target's test performance and subjective experience. Br. J. Soc. Psychol. 44 (4), 645–657.

Marx, D.M., Roman, J.S., 2002. Female role models: protecting women's math test performance. Pers. Soc. Psychol. Bull. 28 (9), 1183–1193.

McFarland, L.A., Lev-Arey, D.M., Ziegert, J.C., 2003. An examination of stereotype threat in a motivational context. Hum. Perform. 16 (3), 181–205.

McIntyre, R.B., Paulson, R.M., Lord, C.G., 2003. Alleviating women's mathematics stereotype threat through salience of group achievements. J. Exp. Soc. Psychol. 39 (1), 83–90.

McKay, P.F., Doverspike, D., Bowen-Hilton, D., Martin, Q.D., 2002. Stereotype threat effects on the raven advanced progressive matrices scores of African Americans 1. J. Appl. Soc. Psychol. 32 (4), 767–787.

McKay, P.F., Doverspike, D., Bowen-Hilton, D., McKay, Q.D., 2003. The effects of demographic variables and stereotype threat on Black/White differences in cognitive ability test performance. J. Bus. Psychol. 18 (1), 1–14.

National Science Foundation. 2015. *Report to congress on advancing historically black colleges and universities.* Education and Human Resources Directorate, US National Science Foundation. Retrieved from https://www.nsf.gov/ehr/Materials/HBCU_Report_Oct_2015.pdf

Nichols, A.H., Evans-Bell, D., 2017. A look at Black student success: identifying top- and bottom-performing institutions. The Education Trust, Washington, DC http://edtru.st/2mD7pWL.

Phinney, J.S., 1992. The multigroup ethnic identity measure: a new scale for use with diverse groups. J. Adolesc. Res. 7 (2), 156–176.

Price, G.N., Spriggs, W., Swinton, O., 2011. The relative returns to graduating from a historically Black college/university: Propensity score matching estimates from the national survey of Black Americans. The Review of Black Political Economy 38 (2), 103–130. doi:10.1007/s12114-011-9088-0.

Rushton, J.P., Jensen, A.R., 2005. Thirty years of research on race differences in cognitive ability. Psychol. Public Policy Law 11 (2), 235–294. doi:10.1037/1076-8971.11.2.235.

Shih, M., Pittinsky, T.L., Ambady, N., 1999. Stereotype susceptibility: identity salience and shifts in quantitative performance. Psychol Sci 10 (1), 80–83.

Sparks, D., 2015. An exploration of the connections between institution type and perceived levels of stereotype threat in African American Engineering Students. J. Afr. Am. Males Educ. 6 (1), 42–58.

Steele, C.M., Aronson, J., 1995. Stereotype threat and the intellectual test performance of African Americans. J. Pers. Soc. Psychol. 69 (5), 797–811. doi:10.1037/0022-3514.69.5.797.
Walton, G.M., Spencer, S.J., 2009. Latent ability: grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students. Psychol Sci 20 (9), 1132–1139.
Yeung, W.J., Conley, D., 2008. Black-white achievement gap and family wealth. Child Dev. 79 (2), 303–324.