Stanford cepa

Center for Education
Policy Analysis

Did a “Traumatic” Test Question Create Racial Bias?

AUTHORS

Thomas S. Dee
Stanford University and NBER

Benjamin W. Domingue
Stanford University

ABSTRACT

On the second day of a 2019 high-stakes assessment, Massachusetts 10th graders faced a racially insensitive question, which was quickly invalidated. However, a remaining controversy is whether this exposure created a racial bias in performance on the remaining test items. We present results from a pre-registered analysis suggesting the controversial question reduced the comparative performance of Black students by a small amount (approximately 0.006σ). However, we also find a wide dispersion of such effects when examining similarly small sets of test items from prior state assessments that *lacked* a controversial question which suggests the 2019 assessment was not distinctive.

VERSION

August 2019

Suggested citation: Dee, T.S., & Domingue, B.W. (2019). Did a “Traumatic” Test Question Create Racial Bias? Retrieved from Stanford Center for Education Policy Analysis:
https://cepa.stanford.edu/content/did-traumatic-test-question-create-racial-bias

Last March, 10th graders in Massachusetts' public schools sat for the annual English Language Arts (ELA) exam that is part of the Massachusetts Comprehensive Assessment System (MCAS). A student's MCAS performance has high individual stakes.[1] The first day of the test (16 multiple-choice items and 2 essays) was without incident. The second day of the exam began with students reading a passage from the prize-winning 2016 novel *The Underground Railroad*. They then responded to 8 multiple-choice items before being asked to write a journal entry from the perspective of a White female character. The test then concluded with another short passage and 4 multiple-choice items.

Some students and organizations quickly criticized writing from the perspective of a character described in one press account as "openly racist" (Gerst 2019, State News House Service 2019) as inappropriate and "traumatic."[2] The book' author, Colson Whitehead, commented on the controversy stating "Whoever came up with the question has done a great disservice to these kids, and everyone who signed off on it should be ashamed." The Massachusetts Department of Elementary and Secondary Education (DESE) soon decided to remove the essay in question from the scored portion of the student response.

However, several concerned groups called for invalidating the results of the entire two-day exam. For example, the president of the Massachusetts Teachers Association noted "…all students need to be held harmless across the state and the test itself needs to be ruled invalid." A particular concern is that the question could introduce racial bias in student's performance. An extensive lab and field-experimental literature on "stereotype threat" has found that, in highly evaluative settings, priming awareness of a stereotyped identity can sometimes impede cognitive performance through mediators like anxiety.[3] The existence of stereotype threat could have lowered the comparative MCAS performance of Black students. Alternatively, the existence of "stereotype reactance" (i.e., motivational arousal from an unpleasant stimulus) could have increased their comparative performance. Given the

[1] For example, a student's performance must meet or exceed the "Proficiency" threshold (or an equivalent level on new next-generation grade 10 tests) to be eligible for graduation. Alternatively, a student can be eligible for graduation if their score instead exceeds the "Needs Improvement" threshold (or the next-generation equivalent) and they meet the requirements of an "Educational Proficiency Plan" (EPP). Furthermore, student MCAS performance above the "Proficiency" and "Advanced" thresholds (or the next-generation equivalents) are required for the John and Abigail Adams Scholarship which provides free tuition at state colleges and universities.

[2] The passage and the controversial question in question can be viewed online at http://www.doe.mass.edu/mcas/2019/release/g10ela-voidedessay.pdf.

[3] For overviews of this literature see Pennington et al. (2016) and Aronson and Dee (2012).

high stakes associated with the exam and the central role of fairness in educational measurement, the potential performance implications of the question are a serious concern.

In this study, we examine the evidence for such bias by analyzing student-level MCAS responses from both before and after the essay question. Critically, prior to receiving the data from the Massachusetts Department of Elementary and Secondary Education (DESE), we pre-registered our analytic strategy.[4] Our single, confirmatory hypothesis is to ask whether the test performance of Black students on the 4 post-question multiple-choice items differed significantly from that of White students conditional on their first-day performance. Our analysis is based on data from 49,034 students. We first estimated separate scores using items from the first-day and the post-question items. We used item-response models similar to those used by the state (MA DESE 2018; Thissen et al. 1995) and standardized these scores using the full sample. Focusing on just the Black and White students, we then regressed post-question scores on first-day scores and a binary indicator for Black students.[5] We report the key results from this regression in the first column of Table 1.

We found the post-essay performance of Black students was 0.061σ lower than would be expected given their first-day performance (p-value < 0.001). We also found similar results across approaches we pre-registered as exploratory. For example, we estimated this specification using all non-White students as the focal group (i.e., a racial offset of -0.051σ; p-value < 0.001). Also, because students could navigate back to the 8 prior “second day” multiple-choice items after seeing the question, it may be that the outcome measure should include these additional items. In Table 1, we also report results based on this approach and find the estimated racial offset is larger (-0.128σ; p-value < 0.001).

Taken at face value, our results suggest the controversial question introduced statistically significant racial bias in post-question scores. However, the magnitude of this effect is extremely small. We offer two illustrative interpretations of this effect size. First, a 0.061σ reduction on 4 items representing 5 out of 51 available points roughly corresponds to an overall loss of 0.006σ (0.061*5/51). Under the state’s new standards, 4.2% of Black students fail to meet the ELA competency standard for graduation. Assuming a normal distribution for the population of 6,167 Black test-takers, such an

[4] Our registration is available here: https://sreereg.icpsr.umich.edu/framework/pdf/index.php?id=2313

[5] We also adjust for measurement error in the baseline test scores using the marginal reliability of the test and the “eivtools” package in R (Lockwood and affrey 2014). We also cluster standard errors at the school level.

impact would make 3 additional students ineligible for graduation. Second, the precise impact for a given student depends on their underlying performance but has an upper bound of 0.08 fewer points on these 4 items. Using the test response function for the entire test, this reduction in points translates to a 0.007σ decrease in performance on the whole test for a maximally affected student. Notably, both approaches similarly imply that the controversial question affected relatively few students.

Furthermore, the arbitrary partitioning of a small number of "post-question" test items could speciously create such racial differences in both positive and negative directions. For example, this variability could be due to differential item functioning (DIF) across small subsets of questions.[6] To assess the empirical relevance of this issue, we applied our confirmatory test to the student-level data from the 2017 and 2018 math and ELA MCAS tests given in grades 3-8 and 10. Specifically, for each of these 28 tests, we constructed comparisons corresponding to our focal case wherein students only took four items on the second day. We did this by constructing scores from four items among the second-day on each test. We took either all combinations of four items from tests with fewer than 9 second-day items and sampled 100 subsets of second-day items from tests with 9 or more such items, resulting in 2,645 comparison sets. We then estimated first- and second-day scores and applied our confirmatory test to each data set. Figure 1 shows the distribution of estimated performance offsets associated with Black students. This distribution (mean = -0.023, SD = 0.091) suggests that our approach may incorrectly suggest the presence of a racial bias. Seventy-six percent of these point estimates are statistically significant from zero though there were no known controversies involving particular questions on these tests. And 52% of these point estimates are larger in absolute value than our confirmatory estimate. Our overall conclusion is that, while our confirmatory test suggests that the controversial question creates a small racial bias in test performance, the broader context indicates that this is similar to effects, both positive and negative, observed in other MCAS settings.

[6] In supplemental analyses, we did in fact detect some potential DIF on one of the 4 post-question items.

REFERENCES

Aronson, J., & Dee, T. (2012). Stereotype threat in the real world. In Inzlicht, M., & Schmader, T. (Eds.). (2012). *Stereotype threat: Theory, process, and application*. New York, NY, US: Oxford University Press. 264-278.

Gerst, E. 2019. "MCAS Question about *The Underground Railroad* Thrown Out," Boston Magazine, https://www.bostonmagazine.com/education/2019/04/04/underground-railroad-mcas-question/, Accessed August 27, 2019.

Lockwood, J. R., & McCaffrey, D. F. (2014). Correcting for test score measurement error in ANCOVA models for estimating treatment effects. *Journal of Educational and Behavioral Statistics*, *39*(1), 22-52.

Massachusetts Department of Elementary and Secondary Education. (2018). 2017 Next-Generation MCAS and MCAS-Alt Technical Report. http://www.mcasservicecenter.com/documents/MA/Technical%20Report/2017/NextGen/2017%20MCAS%20NextGen%20Technical%20Report.pdf, Accessed August 28, 2019.

Pennington, C. R., Heim, D., Levy, A. R., & Larkin, D. T. (2016). Twenty years of stereotype threat research: A review of psychological mediators. *PloS one*, *11*(1), e0146487.

State House News Service. 2019 "'Traumatic' Test Question Removed After Students Complained" https://www.wcvb.com/article/traumatic-mcas-question-removed-from-test-after-students-complain/27041456, Accessed August 27, 2019.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item Response Theory for Scores on Tests Including Polytomous Items with Ordered Responses. Applied Psychological Measurement, 19, 39-49.

Table 1 - Estimated Effects on Post-Question MCAS Performance

	Post-Question Score	
Independent variable	4 Items	12 Items
First-Day Score	0.693 (0.007)	0.781 (0.010)
Black	-0.061 (0.014)	-0.128 (0.013)

Notes: Standard errors reported in parentheses.

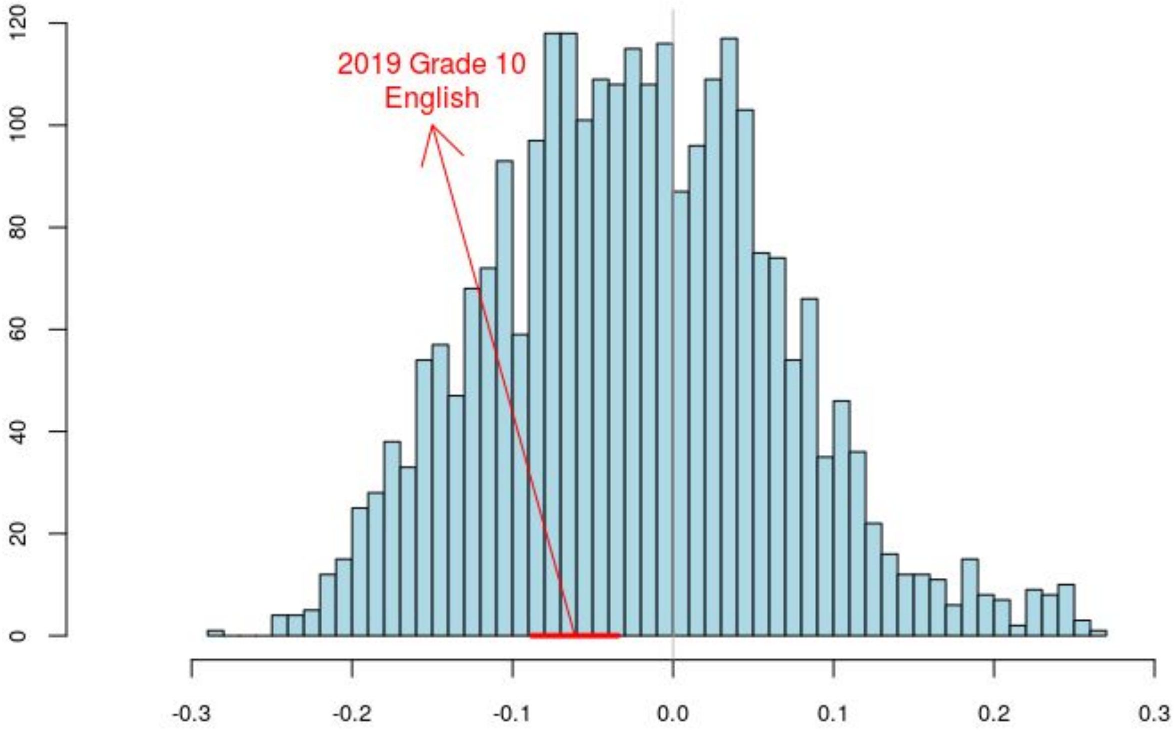
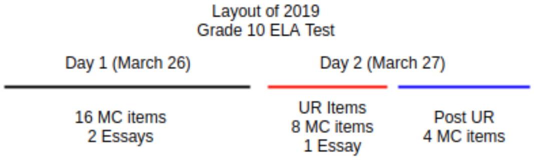


Figure 1 - Distribution of Estimated Racial Offsets in Second-Day 2017 and 2018 MCAS Scores (n = 2645) Conditional on First-Day Scores

Online Supplement for "Did a 'Traumatic' Test Question Create Racial Bias?"

1. Methods

A schematic of the MCAS test in question is shown below. In our core analysis, we first estimate first-day scores based on the 18 items delivered on March 26. We then estimate second-day scores based on only the four items (shown in blue) not associated with the Underground Railroad passage. In a secondary analysis, we also consider a version of the second-day score using all items in blue and red except for the essay on the passage in question. Scores were generated using item response models along the lines of those used in the MCAS (1). Dichotomously coded item responses were modeled with the 3PL while polytomously coded responses were modeled with the graded response model. We then generated EAP estimates for each sum score (2). Item response models were estimated using mirt (3).



2. Exploratory Research Questions

We briefly note here analyses related to exploratory research questions posed in pre-registration.[1]

Question 1: Are the results of the confirmatory research question sensitive to the choice of functional form?

We considered the inclusion of a quadratic term for first-day test score alongside the linear term. Estimates were adjusted for measurement error in both the first-day test score, the quadratic term, and the covariance of those two. Results are as shown below and are qualitatively similar to those in main text (also shown here).

	Baseline		w/ quadratic term	
	Est	SE	Est	SE
(Intercept)	0.025	0.007	0.101	0.009
th1	0.693	0.007	0.673	0.005
group	-0.061	0.014	-0.041	0.014
th1squared			-0.086	0.006
N	49034		49034	

Question 2: Are the results of the confirmatory research question sensitive to designating as part of the post-treatment score the set of questions that followed the passage but preceded the essay prompt?

We discuss this in the main text (see column 2 of Table 1).

[1] Available here: https://sreereg.icpsr.umich.edu/framework/pdf/index.php?id=2313

Question 3: Are the results of the confirmatory research question sensitive to alternative definitions of which students are African American?

As noted in the main text, we find get similar results when all non-White students are considered the focal group possibly influenced by controversial question. We also find that, when Black students are the focal group and all non-Black students are in the comparison group, the estimated offset is -0.015 (p-value = 0.25).

Question 4: Are the results of the confirmatory research question sensitive to controlling for other student traits (e.g., SES, English Learner status, etc.)?

This was not conducted and is subject to data availability.

Question 5: What is the differential impact of the controversial essay question on the subsequent MCAS item missingness and response times for African American students relative to non-Hispanic White students?

Response Time

We focus here on the selected response items given that these are the focal items for the second-day analysis (and response times are much longer for essay items). Black students took less time on items after the essay passage than they did on first-day as compared to White students. In terms of total time spent on items, the mean Black student was at the 91th percentile of White test-takers for first-day items versus the 84th percentile for the second-day items following the essay passage.

Missingness

We again focus on the selected response items. Missingness (i.e., skipping items) is quite rare on the MCAS; e.g., 99.7% of the students had complete response strings on first-day. Missingness was similar in prevalence on second-day items that followed the essay passage amongst Black students. On first-day, the mean Black student skipped 4 times as many items as the mean White student. On second-day items following the essay prompt, the mean Black student skipped 3.8 times as many items as the mean White student.

Question 6: Assuming the responses to different test sections can be placed on comparable scales, what is the impact of being African American on performance in the post-essay sections conditional on student fixed effects and test-section fixed effects (i.e., a "difference in differences" or "DD" specification)?

We felt the challenges of creating a shared scale for the first-day and post-question items a DD design would require were prohibitive. Regardless, we also found that a DD analysis using person-by-item data resulted in a small and statistically insignificant estimated effect; a finding broadly consistent with the conclusion of our main analysis.

	Estimate	SE
Indicator (black student responding to item 28-31)	0.089	0.090
Person FE	X	
Item FE	X	
N	1078748	

Note: Standard errors clustered by person and item.

Question 7: If there is evidence that DD estimates are subject to a confound related to effects unique to African American students on the second day of a test, what is the estimated impact of the essay prompt in "difference in difference in differences" (DDD) specifications that include data from test(s) without a race-related essay prompt (i.e., tests based on other subjects or years)?

Given the measurement challenges noted in our main analysis (Figure 1) and the DD results and scaling issues noted above, we did not conduct a DDD analysis.

Question 8: Do we observe differential item functioning (DIF) for post-essay items as a function of race and ethnicity?

We conducted a variety of DIF analyses. One complicating factor is that the last item was polytomously scored (students could get 0/1/2 points on this item) while the others are dichotomously scored. All are based on the most stringent criteria available for Black versus White students (which we used in the main text).

We focus on the standardized DIF approach used by MCAS, computed via (5). Widely used standards (4) suggest that delta statistics greater than 1 in magnitude be viewed as evidence for moderate DIF'. By this standard, none of the items exhibit DIF.

We also consider a variety of alternative approaches; details are below. Collectively these suggest some potential DIF for item 29.

1. This is a test for uniform DIF based on a linear probability model and the first-day sum score. Estimates are the expected change in # of points on the item for a Black student. These estimates are biased due to the fact that they do not account for measurement error in the first-day test score.
2. Here we revise the approach above approach by correcting for measurement error on the first-day test as in the main text. Note that they cumulatively suggest an expected loss of about 0.07 scale points on these four items for Black students relative to expectation given first-day test scores. Items 29 and 31 are potentially exhibiting some degree of DIF.
3. The sibtest statistic computed by mirt (3), convenient given that it corrects for measurement error in the matching score and allows for different response formats.
4. The lordif statistic computed by package of same name (6).
5. We also use the lordif package to examine non-uniform DIF.

		Item 28		Item 29		Item 30		Item 31	
		Estimate	Delta	Estimate	Delta	Estimate	Delta	Estimate	Delta
	Standardized DIF	-0.016	-0.2193	-0.054	-0.5262	-0.008	-0.1166	0.006	0.1162
		Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
1	Linear	-0.018	<0.001	-0.053	<0.001	-0.013	0.005	-0.064	<0.001
2	EIV	-0.007	0.29	-0.032	<0.001	-0.004	0.5	-0.027	0.039
3	Sibtest	-0.011	0.015	0.014	0.038	-0.001	0.85	-0.029	0.003
4	LORDIF		0.513		<0.001		0.470		0.690
5	LORDIF (non-unif)		0.477		<0.001		<0.001		0.412

References

1. Massachusetts Department of Elementary and Secondary Education. 2017 Next-Generation MCAS and MCAS-Alt Technical Report [Internet]. [cited 2019 Aug 29]. Available from: http://www.mcasservicecenter.com/documents/MA/Technical%20Report/2017/NextGen/2017%20MCAS%20NextGen%20Technical%20Report.pdf

2. Thissen D, Pommerich M, Billeaud K, Williams VS. Item response theory for scores on tests including polytomous items with ordered responses. Applied Psychological Measurement. 1995;19(1):39–49.

3. Chalmers RP. mirt: A multidimensional item response theory package for the R environment. Journal of Statistical Software. 2012;48(6):1–29.

4. Holland P, Thayer D. An alternative definition of the ETS delta scale of item difliculty. ETS Research Report No; 1985.

5. Magis D, Béland S, Tuerlinckx F, De Boeck P. A general framework and an R package for the detection of dichotomous differential item functioning. Behavior research methods. 2010;42(3):847–862.

6. Choi SW, Gibbons LE, Crane PK. Lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. Journal of statistical software. 2011;39(8):1.