Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Summer 2021, Vol. 40, No. 2, pp. 81–88

Assessing the Impact of a Test Question: Evidence from the "Underground Railroad" Controversy

Thomas S. Dee, *Graduate School of Education, Stanford University, NBER,* and Benjamin W. Domingue, *Graduate School of Education, Stanford University*

Abstract: *On the second day of a 2019 high-stakes English Language Arts assessment, Massachusetts 10th graders faced an essay question that was based on a passage from the novel "The Underground Railroad" and publicly characterized as racially insensitive. Though the state excluded the essay responses from student scores, an unresolved public controversy focused on whether this question created a racial bias in performance on the remaining test items. We present the results from an independent, preregistered study of this question. Our confirmatory results indicate that exposure to the controversial question is associated with a small reduction in the comparative performance of Black students on the overall test (approximately .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. Our approach offers a potential template that may be useful in other contexts where testing controversies occur and underscores the importance of carefully screening test items to avoid such occurrences.*

Keywords: fairness, high-stakes testing, MCAS, stereotype threat

Introduction

In March 2019, nearly 70,000 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 stakes, with respect to both graduating from high school and access to free tuition at state colleges and universities.[1] The first day of the test (i.e., 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 eight multiple-choice items before being asked to write a journal entry from the perspective of a White female character. The test then concluded with a final section containing four multiple-choice items.

Some students and organizations quickly criticized being asked to write a journal entry from the perspective of a character described in one press account as "openly racist" (Gerst, 2019; Lisinski, 2019), characterizing the question as inappropriate and "traumatic."[2] The book's 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." This question, like all others on the exam, had actually passed through multiple layers of vetting that included a committee of teachers and educators focused on age appropriateness and alignment with standards, a second "Bias and Sensitivity" committee, and two outside experts (Toness, 2019). In response to the controversy, the Massachusetts Department of Elementary and Secondary Education (DESE) quickly 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 2-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. For example, one student who took the test commented "While I was taking the test, I thought about other students in other towns taking the test and what

[Correction added on December 24, 2020 after first online publication: Author Tom Dee's correct affiliation has been updated]

[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.

they were writing and thinking about people like me. I imagined White students writing negative things about me" (Gerst, 2019). This type of subjective response among Black students has clear parallels in recent scholarship from the field of social psychology on student engagement and cognitive performance. In particular, an extensive lab and field-experimental literature on "stereotype threat" has found that, in highly evaluative settings (e.g., taking a high-stakes test like the MCAS), priming awareness of a stereotyped identity can sometimes impair cognitive performance. That is, when students become more aware that others may view them through the lens of a negative stereotype, test performance may suffer.

The available evidence suggest that these negative test-performance effects are due to mediators such as "anxiety, negative thinking, and mind-wandering," which "coopt working-memory resources" among threatened individuals (Pennington, Heim, Levy, & Larkin, 2016). An early meta-analysis (Nguyen & Ryan, 2008) found that the negative effects of race-based stereotype threats were larger (i.e., a standardized effect size of −.43) than those associated with gender (i.e., −.36). The evidence from experimental and quasi-experimental studies focused on real-world settings is broadly consistent with the leading laboratory studies (Aronson & Dee, 2012). However, a recent meta-analysis (Shewach, Sackett, & Quint, 2019) argues that the effects of stereotype threat are distinctly smaller (i.e., an effect size of −.14) in "operational test settings," which, like the MCAS, have high stakes.

This literature suggests that the controversial MCAS question could have lowered the comparative performance of Black students by threatening their social identity in the test-taking context. A closely related literature also suggests the controversial MCAS question may have been differentially harmful to the performance of Black students by reducing their subjective experience of belongingness in an academic setting. Field-experimental studies (e.g., Walton & Cohen, 2007, 2011) have found that interventions that promote social belongingness in school and that framed social adversity as a shared phenomenon unrelated to ability or race improved academic performance. To the extent that the MCAS question reduced students' sense of belongingness in the test setting, it may have reduced their performance. However, the existence of "stereotype reactance" (i.e., motivational arousal from an unpleasant stimulus) could have blunted the performance-dampening effects of such identity threats or even increased the comparative performance of Black students on the MCAS. Existing studies suggest that stereotype reactance can occur in evaluative settings that allow enough time to recover from the initial identity threat through effort (Jamieson & Harkins, 2007) and among individuals with a high degree of self-monitoring (i.e., a motivation to regulate one's actions in order to project a desired public image (Inzlicht, Aronson, Good, & McKay, 2006)).

Given both the high stakes associated with the MCAS exam and its statewide scale, the potential performance implications of this testing controversy are a serious concern. In this study, we examine the evidence for potential performance implications 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 preregistered our analytic strategy.[3] Our single, confirmatory hypothesis is to ask whether the test performance of Black students on the four postquestion multiple-choice items differed significantly from that of White students conditional on their first-day performance. While this reduced-form impact is of central interest, we note that the available administrative data do not allow us to examine directly the theorized psychological mediators. However, alongside the central question we engage, we also conduct several ancillary analyses meant to both provide context for our finding and to probe its robustness. In particular, we assess the reliability of our confirmatory research design by applying it to the data from other recent MCAS tests, where there are no similarly controversial questions.

Our work provides a novel illustration of how researchers can engage questions of fairness in real-world testing environments. Such concerns about fairness abound in educational measurement (Camilli, 2006). A typical approach in post hoc analyses is to examine individual items for differential item functioning (DIF). In such conventional approaches, the key assumption is that nonfocal items contain minimal bias and performance on the focal item can be examined conditional on some unbiased index of ability constructed (largely) from the nonfocal items. However, in this context, the question of interest is about performance on a particular *subset* of items; specifically, those following the controversial essay prompt on the second day of the test. By examining students' performance on this final test section conditional on first-day performance, our core research design effectively leverages the position of the controversial essay prompt in the test's item sequence. We also complement this analysis with the results from traditional methods of detecting DIF. However, our core analytical approaches, which rely on differential prediction and have strong parallels with quasi-experimental impact evaluations, may be useful to researchers and practitioners interested in assessing testing controversies in other settings.

Methods

Data

Our analysis is based on data from 68,090 students taken from the 2019 Grade 10 ELA MCAS test. Specifically, our analysis focuses on the 49,034 students who took this test and identified as Black or White. As a way to complement our analysis of the controversial 2019 Grade 10 ELA administration and to understand the properties of our preregistered analytical design, we also constructed similar data sets from MCAS administrations conducted in the prior 2 years (i.e., 2017 and 2018). These 28 additional MCAS exams covered two subjects (i.e., mathematics and ELA) across seven different grades (i.e., grades 3 through 8 and grade 10) over these 2 years. Crucially, these tests had the same 2-day structure, a fact we use to enable parallel analysis of these data. Table 1 presents for each

[3]Our registration is available at the SREE registry, see ID 1759.1v1 at https://sreereg.icpsr.umich.edu/sreereg/. The preregistration of core hypotheses and research designs prior to conducting an analysis has become common in experimental studies as a way to attenuate the risk of flawed inferences resulting from searching over multiple outcome measures and analytical approaches. Similar risks exist in quasi-experimental studies like ours, though preregistration is not yet common in such applications. However, we believe that preregistration can be especially useful in contexts like ours where a high-profile controversy can more readily sow suspicion regarding researcher discretion.

Table 1. Descriptive Statistics for MCAS Datasets

Subject	Year	Grade	*N* Test Takers	% Black
English	2017	3	48,967	13.2%
English	2017	4	49,752	12.9%
English	2017	5	49,686	12.7%
English	2017	6	50,436	12.4%
English	2017	7	51,896	12.2%
English	2017	8	52,603	11.9%
English	2017	10	55,524	13.3%
Math	2017	3	48,967	13.2%
Math	2017	4	49,752	12.9%
Math	2017	5	49,686	12.7%
Math	2017	6	50,436	12.4%
Math	2017	7	51,896	12.2%
Math	2017	8	52,603	11.9%
Math	2017	10	55,524	13.3%
English	2018	3	47,182	13.9%
English	2018	4	49,050	13.5%
English	2018	5	49,915	13.3%
English	2018	6	49,715	13.1%
English	2018	7	50,417	12.8%
English	2018	8	51,898	12.4%
English	2018	10	54,686	13.9%
Math	2018	3	47,182	13.9%
Math	2018	4	49,050	13.5%
Math	2018	5	49,915	13.3%
Math	2018	6	49,715	13.1%
Math	2018	7	50,417	12.8%
Math	2018	8	51,898	12.4%
Math	2018	10	54,686	13.9%
English	2019	10	49,034	12.1%

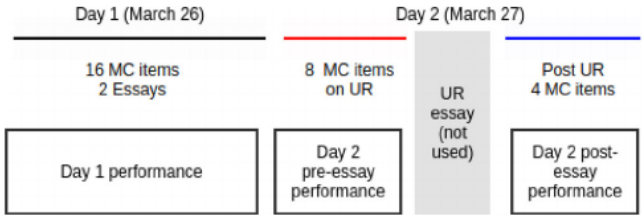

FIGURE 1. Layout of 2019 grade 10 English test emphasizing the multiday structure and the placement of the underground railroad (UR) essay. [Color figure can be viewed at wileyonlinelibrary.com]

of the tests the number of Black and White test-takers and the percentage who were Black.

In Figure 1, we illustrate the key structural features of the 2019 Grade 10 ELA test. On day 1, students took 18 items in total (i.e., 16 multiple-choice questions and two essays). At the start of day-2, students read a passage from *The Underground Railroad* describing a teenage runaway slave, Cora, who is hidden by a man named Martin and his wife, Ethel, who treated Cora rudely while hiding her from "night riders or regulators who capture and return escaped slaves." Students then were presented with eight multiple-choice items about this passage. The next question (and the source of this controversy) asked students to imagine this story from Ethel's perspective and to write it as a journal entry. After concluding this section, students then proceeded to a final and unrelated section that consisted of four multiple-choice questions based on another short passage.

The primary, confirmatory outcome we preregistered reflects student performance on the four multiple-choice questions that followed the controversial essay. However, students had the capacity to navigate back to the eight multiple-choice questions that began the second day of the exam after viewing the essay prompt. This implies that student performance on these eight items could have also been influenced by the essay question. Therefore, we also examined an outcome measure based on the *twelve* second-day multiple-choice items. The key independent variables in the analytical plan we describe below are a measure of test performance on the first day and a binary indicator for whether the student is Black. It should be noted that the state quickly voided the controversial essay prompt and responses to it are not part of our analysis.

Rather than report sum scores that simply focus on the total number of points the student got on the test, the MCAS relies upon item response theory (IRT) models to map student performance onto a score scale. So as to mirror as closely as possible the approach used by the state, we used item–response models similar to those used by the state in their official score calculations (Massachusetts Department of Elementary and Secondary Education, 2018). Specifically, for different combinations of item responses, we needed to estimate student scores. We calibrated test items using the three parameter logistic (3PL) model for dichtomously scored responses (Birnbaum, 1968) and the graded response model (Samejima, 2016) for polyomously scored responses.[4] We then used item parameters to construct expected a posterior (EAP) sum scores (Thissen, Pommerich, Billeaud, & Williams, 1995). All IRT analyses were conducted using mirt (Chalmers, 2012), an R package for estimation of multidimensional item response theory models in R, and DIF analyses were computed via Magis, Béland, Tuerlinckx, and De Boeck (2010). We then standardized these scores using the full sample of test-takers; this standardization allows us to directly interpret estimates in terms of effect sizes (i.e., relative to the *SD* of the test-takers).

Analysis

Our main question is whether exposure to the essay prompt impacted student performance on the remaining items. If so, this may raise questions about the fairness of the MCAS which, given the central role of fairness in interpreting and using test scores (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014), would be a serious concern. Our preregistered, confirmatory research design involves an analysis of differential prediction of postessay scores as a function of student race while controlling for day-1 performance. That is, focusing on just the Black and White students, we regressed postquestion scores on first-day scores and a binary indicator for focal group membership (i.e., Black

[4]We briefly discuss the fit of the IRT models to this mixed format data. We obtain Root Mean Square Error of Approximation (RMSEA) values of .013 for the first day responses and .016 for the second day responses. Values below .06 are typically interpreted as indicating appropriate model fit (Hu & Bentler, 1999). These values are also comparable to those reported in the context of comparable state assessments; e.g., the Ohio (Table 1.5.1.1) and Maryland (Table 12) state assessments (American Institutes for Research, 2017; Pearson, 2016)

Table 2. Estimated Performance for Black Students on Items Following the Controversial Essay (After Controlling for Day-1 Performance)

Model	Outcome	Focal Group	Reference Group	Baseline Covariate	Estimate (in *SD* Units)	SE	*p* Value	*N*
1	Postessay items	Black students	White students	Linear	−.0612	.0138	<.001	49,034
2	Postessay items	Non-White students	White students	Linear	−.051	.0083	<.001	68,090
3	Postessay items	Black students	Non-Black students	Linear	−.015	.0132	.250	68,090
4	All day-2 items	Black students	White students	Linear	−.1278	.0133	<.001	49,034
5	Postessay items	Black students	White students	Linear and Quadratic	−.0406	.0136	.003	49,034

students). We accounted for two important issues in this regression model. First, we note that the performance measure (i.e., the EAP sum score) from the first-day items will contain measurement error. When this performance measure is used as a predictor, such measurement error is likely to lead to estimation bias; in particular, it may lead to an overestimate of the group-level coefficient. Thus, we also adjust for measurement error in the baseline test scores using the marginal reliability of the test via the error-in-variables linear regression framework implemented in Lockwood (2018). Second, because the clustering of students within schools may create a within-group dependence among the error terms, we cluster standard errors at the school level.

We extended this main analysis and explored the robustness of our results in several ways. These included modifying the functional form of the model (i.e., adding a squared ability term), using different definitions of the focal group (e.g., all non-White students), and using all 12 day-2 multiple choice items (i.e., including in our outcome measure the eight items that students could have engaged after reading the essay prompt).[5] We also recognize that, because the outcome in our main design is based on only four items, our approach could generate a specious finding simply because Black and White students may tend to perform differently on these few items for reasons unrelated to the controversial essay question. We examine this important concern in two ways. One is to calculate directly the DIF on these items by race. Second, we also examined the empirical relevance of this concern by applying our basic research design to the data from 28 MCAS exams over the prior 2 years. Specifically, we constructed a distribution of effects that might occur with our research design due to chance. We did so by estimating the differential prediction on 4 day-2 items by race on these earlier exams where there was no controversy. Finally, we also extend our analysis of test performance by examining the impact of the controversial question on item response times and item missingness (e.g., due to skipping or items that were not reached).

Results

Confirmatory Findings

Our preregistered, confirmatory inference compares performance on the postessay items for Black students to the performance of White students after conditioning on day-1 scores and correcting for measurement error. We report the key results of this analysis in row 1 of Table 2. Note that this result, as well as the others in the table, correct for measurement error in the conditioning score (e.g., the day-1 score in this case). We found the postessay performance of Black students was significantly lower (i.e., by .061 of a population-level standard deviation) than would be expected given their first-day performance ($p < .001$). In isolation, this would support the argument that the essay prompt perhaps served to bias downward the postessay performance of Black students.[6] However, as we note in our discussion, the magnitude of this statistically significant estimate is rather small (i.e., a performance reduction of $.06\sigma$ on four of the test's questions).

We also explored the robustness of the finding based on this design in three ways. First, we examined the sensitivity of our findings to alternative definitions of the focal group of students that might be influenced by the controversial test question and of the corresponding group of students who serve as a reference point (models 2 and 3 in Table 2). We obtained similar results when all non-White students are considered the focal group that may be possibly influenced by the controversial question ($-.051\sigma, p < .001$). 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 $-.015\sigma$ ($p = .25$). These results suggest that our findings are qualitatively similar when we use reasonable alternatives for our definition of focal and reference groups.

Second, we considered the sensitivity of this finding to the choice of functional form. Specifically, we estimated a specification that included a quadratic term for the first-day test score alongside the linear term. Here, we adjusted for measurement error in the first-day test score and the quadratic term while accounting for the covariance of those two terms. The estimated postessay performance offset for Black students ($-.041\sigma, p = .003$; Table 2 Model 5) was qualitatively similar to the estimated performance based on our confirmatory design. Our results don't seem overly sensitive to the differences between these two functional forms.

Our third robustness check focused on an alternative definition of the dependent variable. After seeing the essay

[5]Our approach is closely related to a "difference in differences" approach. This widely used quasi-experimental design would rely on student-by-day performance measures and, conditional on student and day fixed effects, examine the day-2 impact unique to Black students. We chose not privilege this in our preregistration because of concerns over the comparability of day-1 and day-2 score scales.

[6]To probe heterogeneity as a function of day-1 score, we also considered analyses in data stratified into quintiles based on day-1 scores. Across all quintiles, Black students performed worse; effects ranged from −.056 to −.131. Given that the effects did not vary systematically (i.e., monotonically) across quintile and that this approach was not included in our preregistration, we focus on the main-effects analysis.

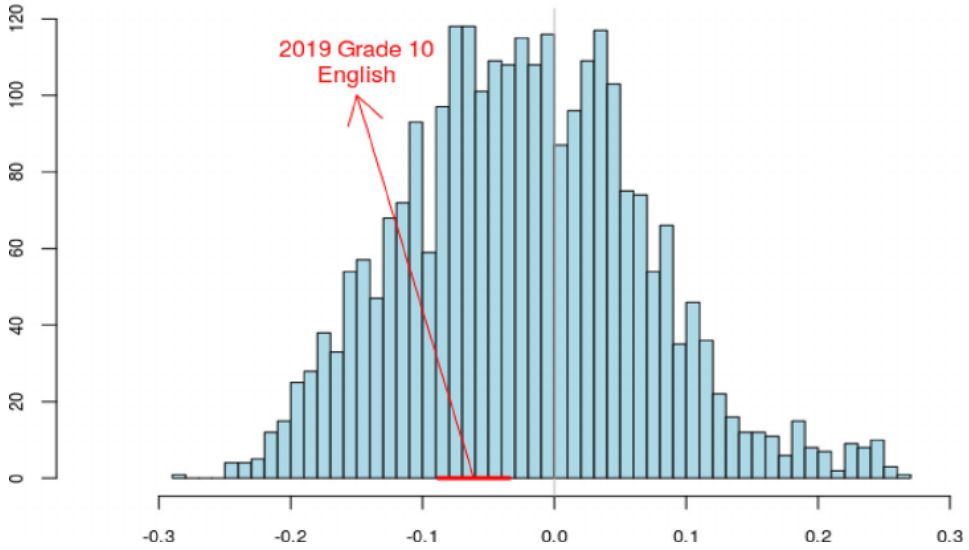

FIGURE 2. Distribution of estimated racial offsets in day-2 2017 and 2018 MCAS scores (n = 2,645) conditional on day-1 scores [Color figure can be viewed at wileyonlinelibrary.com]

prompt, students could navigate back to the eight multiple-choice items that began the second day of testing. Therefore, it is possible that the outcome measure should include these items. When doing so, we find that the estimated racial offset is larger in absolute value and statistically significant ($-.128\sigma$, $p < .001$; Table 2 Model 4). Given this finding, we also conduct additional analyses of the eight items that occur on day-2 prior to exposure to the essay (see below).

Out-of-Sample Benchmarking of the Confirmatory Design

Our confirmatory results suggest that the controversial essay question led to a modest but statistically significant reduction in the comparative performance of Black students on this high-stakes state test. However, another potential concern with our analysis is that examining a small subset of questions could lead to capricious results that do not necessarily reflect the theorized mediators (e.g., stereotype threat, social belongingness) suggested by critics of the exam. In particular, the arbitrary partitioning of a small number of "postquestion" test items could speciously create such racial differences in both positive and negative directions. For example, this variability could be due to a concentration of differential item functioning (DIF) unrelated to the essay passage across small subsets of questions. 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 (i.e., prior years that lacked the controversial question). 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 2 shows the distribution of estimated performance offsets (mean = $-.023$, $SD = .091$) associated with Black students along with the offset (plus confidence interval) from the test in question. The wide dispersion of these estimated offsets suggests that our preregistered design, which necessarily focuses on a small number of items, may incorrectly suggest the presence of a racial bias in either negative or positive directions. That is, the performance estimates for Black students when drawn from a similarly small set of day-2 items may differ from zero given day-1 performance. Seventy-six percent of these point estimates are statistically different from zero though there were no known controversies involving particular questions on these tests. And 52% of these point estimates are larger in magnitude than our confirmatory estimate. More formally, if we viewed Figure 2 as the distribution of estimated racial offsets under the null hypothesis of no effect, we would fail to reject that null hypothesis.

Differential Item Functioning on Postessay Questions

To provide complementary insight into the properties of the four items in the performance measure of interest, we also conducted a more conventional assessment of the racial patterns in answer to these questions. Specifically, we conducted a variety of analyses testing for differential item functioning (DIF) on the four postessay questions (Camilli, 2006; Clauser & Mazor, 1998). One complicating factor is that the last item was polytomously scored (i.e., students could get 0/1/2 points on item 31) while the others are dichotomously scored. We thus restrict analyses to DIF tests that can be deployed over both dichotomously and polytomously scored items. All DIF analyses are based on the comparison between Black versus White students with the day-1 score used as the matching variable.

We focus on the standardized DIF approach used by the Massachusetts Department of Elementary and Secondary Education as it is both widely used and can be used to analyze both the dichotomously and polytomously scored items. Widely used standards (Holland & Thayer, 1985) 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 (see Table 3); item 29 had the largest (in magnitude) delta, $-.53$. We also consider a variety of alternative approaches (rows 1–5 of Table 3). Collectively these

Table 3. Item-Level Analyses: All Analyses Involve Comparisons of Black Students (i.e., the Focal Group) to White Students

	Item 28		Item 29		Item 30		Item 31	
A. DIF								
	Estimate	Delta	Estimate	Delta	Estimate	Delta	Estimate	Delta
Standardized	−.016	−.2193	−.054	−.5262	−.008	−.1166	.006	.1162
	Estimate	*p*-Value	Estimate	*p*-Value	Estimate	*p*-Value	Estimate	*p*-Value
1 Linear[a]	−.018	<.001	−.053	<.001	−.013	.005	−.064	<.001
2 EIV[b]	−.007	.29	−.032	<.001	−.004	.5	−.027	.039
3 SIBTEST[c]	−.011	.015	.014	.038	−.001	.85	−.029	.003
4 LORDIF[d]		.513		<.001		.470		.690
5 LORDIF (nonunif)[e]		.477		<.001		<.001		.412
B. Response Time								
	Estimate	*p*-Value	Estimate	*p*-Value	Estimate	*p*-Value	Estimate	*p*-Value
6 *SD*[f]	.436	.663	.318	.750	.363	.716	.387	.699
7 Quantile[g]	.707		.657		.695		.659	
C. Missingness								
	% Missing		% Missing		% Missing		% Missing	
Black	.068%		.068%		.101%		.270%	
White	.021%		.021%		.019%		.072%	

[a] This is a test for uniform DIF based on a linear model and the first day sum score. Estimates are the expected change in number 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.
[b] Here we revise the approach above 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 .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.
[c] The SIBTEST statistic computed by mirt (Chalmers, 2012), convenient given that it corrects for measurement error in the matching score and allows for different response formats.
[d] The LORDIF statistic computed by package of same name (Choi, Gibbons, & Crane, 2011).
[e] We also use the LORDIF package to examine nonuniform DIF.
[f] Standardized difference (based on overall *SD*) between logged response time of Black students versus White students.
[g] Quantile of median response time for Black student in the distribution of response times from White students.

suggest some potential DIF for item 29. We view evidence from our DIF analyses as being consistent with evidence from the confirmatory approach as they both suggest the potential for some small degree of overall bias against the Black students on the four postessay items.

Effects on Other Item–Response Behaviors

Another possibility is that the controversial test question differentially influenced the test engagement of Black students in a manner not fully captured by their performance measure. To assess this possibility, we focus on two such behaviors: how long respondents take to complete items and missingness. For response time, we focus on the multiple-choice items (i.e., we remove the essay items from day 1) given that the focal items for the second-day analysis are multiple choice and response times are much longer for essay items. Although Black students had longer response times in general, Black students took less time on items after the essay passage than they did on the first day as compared to White students. In terms of total time spent on items, the median total time for a Black student was at the 83th percentile of White test-takers for first-day items. The median response time for Black students was between the 66th and 71st percentiles across the four focal second day items (see Table 3). However, the estimated offsets unique to Black students' response times following the essay question were not statistically significant (Table 3).

For missingness, we again focus on the multiple-choice items from both days. Missingness (e.g., skipping items and failing to reach an item) is quite rare on the MCAS. Of the MCAS respondents, 99.7% of the students had complete response strings on first-day items. Comparing the first and second day of testing, rates of missingness were similar. On first-day items, the mean Black student skipped four 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 (see Table 3 for missingness rates of individual postessay items).

Analysis of Day-2 Preessay Questions

Based on the results from Model 4 of Table 2, which suggested a larger performance decline on the entire set of second-day items as compared to only those that just followed the essay, we also analyzed item responses and associated behaviors (e.g., missingness and response time) for these items. Results are shown in Table 4. In terms of differential item functioning, six of the eight items showed delta statistics suggesting reduced performance of Black students although none of them exceeded the threshold of 1 in magnitude. Amongst Black students, responses were somewhat slower for these items than those shown in Table 3 (e.g., responses were around the 75th percentile of responses from white students as compared to around the 70th percentile for the items following the essay) but again not statistically significant. Missingness rates were again higher for Black students on these items. We view these analyses as suggestive that behavior by Black students on the day-2 items prior to the essay prompt is similar in many respects to behavior following the essay prompt. We cannot rule out that this is due to the prompt (i.e., students read the prompt and then navigated backward to, for example, change their responses).

Table 4. Item-Level Analysis for First 8 Items of Day-2 Testing

	Differential Item Functioning		Response Time			% Missing	
	Estimate	Delta	*SD*	*p*-Value	Quantile	Black	White
Item 19	−.017	−.287	.621	.535	.74	.02%	.00%
Item 20	−.044	−.562	.704	.482	.756	.02%	.00%
Item 21	−.006	−.128	.728	.466	.78	.02%	.01%
Item 22	−.072	−.799	.638	.523	.761	.02%	.01%
Item 23	−.054	−.85	.788	.431	.792	.00%	.01%
Item 24	−.032	−.56	.658	.511	.741	.02%	.01%
Item 25	.008	.258	.655	.513	.734	.02%	.01%
Item 26	.041	.444	.678	.497	.753	.61%	.08%

Discussion

After public concern was raised about the content of a particular essay prompt students were asked to respond to on the 2019 Grade 10 ELA MCAS, the Massachusetts DESE contacted us to independently evaluate whether there was reason to invalidate responses to items following the essay. State officials had already made the decision to invalidate the responses to the controversial essay. We preregistered an analytic approach prior to receiving data that focused on analysis of performance on postessay items relative to day-1 performance. Our core analysis suggests that relative to day-1 performance, Black students performed .061σ worse on the small subset of postessay items than expected.

Taken at face value, our results suggest the controversial question introduced statistically significant racial bias in postquestion scores. We also found that this basic finding was robust to alternative functional forms and definitions of the student groups and the outcome variable. However, the magnitude of this impact is quite small. We offer two illustrative interpretations of this effect size. First, a .061σ reduction on 4 items representing 5 out of 51 available points roughly corresponds to an overall loss of .006σ (i.e., .061×5/51) on the overall test score. Under the state's new standards, 4.2% of Black students fail to meet the ELA competency standard for graduation. Assuming a standard normal distribution for the population of 6,167 Black test-takers, such an impact would make three additional students ineligible for graduation.[7] Second, we can alternatively use the test–response function (for an illustration, see figure 7 in Partchev, 2004) to reason about the potential impact on the points a student receives. The precise impact for a given student depends on their underlying performance but has an upper bound of .08 fewer expected points on these four items (i.e., using the test–response function for just the postessay items). Using now the test–response function for the entire test, we assume that students at every level of performance felt this impact. In that case, the .08 fewer points translates to a .007σ decrease in performance on the whole test for a maximally affected student. Notably, both approaches similarly imply that the controversial question had quite small effects, which would have been consequential for very few students. However, we also note that, given the well-established economic benefits of graduating from high school, the long-run consequences for these few students may have been dire.

Our analysis of out-of-sample MCAS tests (i.e., exams from the prior 2 years when there was no such controversy) complicates that inference by raising serious questions about whether the modest racial offset we found can be reliably determined as an impact. Specifically, we found that, when similarly modeling racial offsets on a subset of test items on other tests, a broad number of both positive and negative estimated effects are quite common. This finding suggests that the modest racial offset we found for the controversial 2019 grade 10 ELA exam is a specious reflection of the item-level variation associated with focusing on a subset of test items. As a general matter, we also note that using out-of-sample data in this way can be a useful and important component of assessing testing controversies in other contexts.

This combination of findings, along with our other exploratory results, implies that the essay did not clearly create racial bias on the state test or, at most, an effect of .006σ. With regard to the theorized psychological mediators, these findings are consistent with the hypothesis that the essay did not constitute a substantial identity threat or that the impact of such threats are attenuated in operational test settings (Shewach et al., 2019). However, this incident and our analysis of it still have several constructive implications. First and foremost, this controversy has compellingly underscored the need for a careful and attentive screening of possible test items. Like Massachusetts, many states and test-development consortia have "Bias and Sensitivity" reviews. However, it may be possible to adapt the design features of these reviews to reduce the chance that objectionable test items are fielded at scale. For example, if the review procedures for a given item rely exclusively on the input of a "large" group of assessors, a "diffusion of responsibility" could increase the chance that an inappropriate item passes review. A long-standing literature in economics and political science (e.g., Olson, 1965) argues that larger groups increase the probability that individuals will "free ride" on the contributions of others. In contrast, a review process that instead assigned the responsibility for specific items to smaller subsets of assessors with overlapping reviews could increase the quality of the overall scrutiny. We note that the Massachusetts DESE has revised and expanded its committee training in response to this incident.

The approach adopted in our analysis may also provide a useful roadmap for analyzing future testing controversies. A careful analysis of testing irregularities such as this one—as well as careful communication of findings back to the specific stakeholders and broader public—are important if high-stakes standardized tests are to retain public faith. Such analyses will be especially important in an era wherein there

[7]Under a standard normal distribution, 4.2% of students not meeting the ELA standard implies a critical z-value of −1.728. An effect size of .006 would shift this critical value to −1.722 and the resulting probability mass to 4.25%. With 6,167 test-takers, this implies 3 additional students not meeting the standard (i.e., .0005 × 6,167).

is an elevated possibility of items that engage sensitive issues quickly becoming highly public and controversial. Established professional standards also underscore the importance of engaging the issues such controversies raise. For example, the standards articulated by AERA, APA, and NCME (2014) stress the fundamental role of fairness in testing. Relevant to the context we study, these standards specifically underscore the importance of valid score interpretations for relevant subgroups (Standard 3.1) and the responsibility to develop tests that minimize the potential influence of "construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics" (p. 63).

In this regard, we believe several features of our analysis are worth underscoring and, possibly, emulating. First, the detailed preregistration of an analytic plan is a uniquely good practice in such settings where public concerns about researcher bias are likely to be paramount. Second, our preregistration substantially benefited from consulting with other researchers and measurement specialists who provided insights on balancing the demands of a sensible research design and of test measurement, which are sometimes in tension. Measurement expertise can also be particularly relevant to examining testing controversies because of the need to focus on performance on context-dependent subsets of the overall items. Third, in an effort to provide an appropriately comprehensive analysis, we believe it is also important to complement preregistered confirmatory hypotheses with sensible exploratory analyses. We note that, in our context, this included an exploratory analysis that interrogated the properties of our core confirmatory analysis, which was based on a small set of test items. Finally, we also note that studies like ours rely critically on the shared purpose and values of the relevant practitioners and public leaders. In our case, this analysis was only possible because the leadership of Massachusetts' Department of Elementary and Secondary committed both publicly and in deed to supporting our entirely independent analysis and to facilitating the relevant data access.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *The Standards for Educational and Psychological Testing*. Washington, D.C.: American Educational Research Association.

American Institutes for Research. (2017). Annual Technical Report: Ohio's State Tests in English Language Arts, Mathematics, Science, and Social Studies. Retrieved from https://oh.portal.cambiumast.com/core/fileparse.php/3094/urlt/OST_Annual_Technical_Report_Spring2016.pdf.

Aronson, J., & Dee, T. (2012). Stereotype threat in the real world. Inzlicht, M & Schmader, T, *Stereotype Threat: Theory, Process, and Application*. New York: Oxford University Press, 264–278.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* pp. 374–472. Charlotte, NC: Addison-Wesley Publishing.

Camilli, G. (2006). Test fairness. *Educational Measurement*, *4*, 221–256.

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

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). 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*, *39*(8), 1.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*(1), 31–44.

Gerst, E. (2019). *MCAS question about the underground railroad thrown out*. Boston Magazine. Retrieved from https://www.bostonmagazine.com/education/2019/04/04/underground-railroad-mcas-question/

Holland, P., & Thayer, D. (1985). An alternative definition of the ETS delta scale of item difficulty. *ETS Research Report*.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

Inzlicht, M., Aronson, J., Good, C., & McKay, L. (2006). A particular resiliency to threatening environments. *Journal of Experimental Social Psychology*, *42*(3), 323–336.

Jamieson, J. P., & Harkins, S. G. (2007). Mere effort and stereotype threat performance effects. *Journal of Personality and Social Psychology*, *93*(4), 544.

Lisinski, C. (2019, April 4). *"Traumatic" MCAS question removed from exam after students complain*. WBUR. Retrieved from https://www.wbur.org/edify/2019/04/04/underground-railroad-mcas-question

Lockwood, J. R. (2018). *eivtools: Measurement error modeling tools*. Retrieved from https://CRAN.R-project.org/package=eivtools

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*(3), 847–862.

Massachusetts Department of Elementary and Secondary Education. (2018). 2017 Next-generation MCAS and MCAS-Alt. Technical Report. Retrieved from http://www.mcasservicecenter.com/documents/MA/Technical%20Report/2017/nextgen/2017%20MCAS%20nextgen%20Technical%20Report.pdf

Nguyen, H.-H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology*, *93*(6), 1314.

Olson, M. (1965). The logic of collective action: Public goods and the theory of groups. *Second Printing with a New Preface and Appendix*. Cambridge MA: Harvard University Press.

Partchev, I. (2004). A visual guide to item response theory. Retrieved from https://www.metheval.uni-jena.de/irt/VisualIRT.pdf

Pearson. (2016). Maryland School Assessment (MSA): Science Grades 5 and 8 Technical Report 2016 Operational Test. Retrieved from http://marylandpublicschools.org/about/Documents/DAAIT/Assessment/MISA/msa_sci_5_8_tech_report_2016.pdf

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), 1–25.

Samejima, F. (2016). Graded response models. van der Linden, W, In *Handbook of item response theory* (vol. 1, pp. 123–136). Boca Raton, FL: Chapman and Hall/CRC.

Shewach, O. R., Sackett, P. R., & Quint, S. (2019). Stereotype threat effects in settings with features likely versus unlikely in operational test settings: A meta-analysis. *Journal of Applied Psychology*, *104*(12), 1514.

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

Toness, B. (2019). *The life cycle of a controversial MCAS Question*. WGBH. Retrieved from https://www.wgbh.org/news/education/2019/08/02/the-life-cycle-of-a-controversial-mcas-question

Walton, G. M., & Cohen, G. L. (2007). A question of belonging: Race, social fit, and achievement. *Journal of Personality and Social Psychology*, *92*(1), 82.

Walton, G. M., & Cohen, G. L. (2011). A brief social-belonging intervention improves academic and health outcomes of minority students. *Science*, *331*(6023), 1447–1451.