Content relatability and test score disparities: Evidence from Texas[*]

Steven Lee[†] Matthew Schaelling[†]

February 27, 2024

(latest draft)

Abstract

One goal of standardized tests is to measure aptitude across heterogeneous students with minimal bias. However, students differ in their experiential knowledge, familiarity, or interest in topics that appear in written content. Could this contribute to bias and widen observed racial test gaps? We study this question empirically using item-level data from high-stakes reading comprehension exams in Texas. We use detailed time-use data and natural language processing techniques to define and build a race/ethnicity-specific measure of a student's "relatability" to passage content in the exams. Relying on quasi-random variation in topics across passages, we find that a one standard deviation increase in the presence of relatable text in a passage predicts a 1.7pp increase in performance on the passage. Extended to the test-level, a one standard deviation increase in test-level relatability leads to a 0.05 standard deviation change in student performance. Our results suggest that equalizing the relatability of passages in these standardized tests could reduce the Black-white and Hispanic-white test score gaps by 4% (0.5pp and 0.4pp, respectively). We then counterfactually estimate over 11,000 Black students and 15,000 Hispanic students during our sample period were designated to be at a lower reading comprehension level due to relatability.

[*] We appreciate valuable comments from Anna Aizer, Jesse Bruhn, John Friedman, Andrew Foster, Amy Handlan, Peter Hull, Stelios Michalopoulos, John Papay, Jonathan Roth, Neil Thakral, and Matt Turner, as well as feedback from participants at various seminars at Brown University and the Annenberg Institute. We thank Maxwell Macort and Elise Togneri for excellent research assistance. We are grateful to the Population Studies and Training Center at Brown University, which receives funding from the NIH, for training support (T32 HD007338) and for general support (P2C HD041020). We additionally acknowledge funding support from the Bravo Center for Economic Research at Brown University. The findings of this research do not necessarily reflect the opinion or official position of the Texas Education Research Center, the Texas Education Agency, the Texas Higher Education Coordinating Board, the Texas Workforce Commission, or the State of Texas.

[†] Department of Economics, Brown University; Population Studies and Training Center, Brown University

1 Introduction

Inequality along racial and economic dimensions is well-documented and widespread in educational contexts. Achievement gaps are observed among children as early as primary school and are especially notable in standardized testing (Fryer & Levitt, 2004; Fryer & Levitt, 2013; Bond & Lang 2013).[1] For example, in the state of Texas there is a 14 percentage point difference between white and Black students in the 3rd and 4th grades on end-of-year reading comprehension exams. In response, some observers and policymakers have called for a deeper understanding of this testing gap and the mechanisms that produce thes differences. While some have suggested abandoning standardized testing entirely, others insist that measurement is essential to accountability and progress, especially regarding racial and economic equality in education.[2]

If standardized tests fail to accurately measure achievement of students across all backgrounds, understanding the exact mechanisms by which it happens is crucial. One possible pathway might be the degree to which educational content is "relatable" to certain groups of students, whereby students learn or perform better when they encounter concepts or topics with which they are familiar or in which they hold interest. Several papers in educational psychology show that interest in a topic can impact performance on reading comprehension tests and that interests diverge by race and gender (Bray & Barron, 2004; Asher, 1979). Other existing research in this area focuses on the representation of different identity groups in educational materials, such as children's literature (Adukia et al., 2023). How-

[1] We recognize that the term "gap" is potentially problematic: some have argued that this term reinforces stereotypes about minority students more broadly (see "How We Talk About the Achievement Gap Could Worsen Public Racial Biases Against Black Students" in EducationWeek, https://www.edweek.org/leadership/how-we-talk-about-the-achievement-gap-could-worsen-public-racial-biases-against-black-students/2020/06). However, we utilize this terminology in a minimal way in accord to link us tightly to prior literature that has used this word. Further, our question is centered on concerns that the empirical measurement tools themselves may be capturing systemic differences.

[2] See, for example, discussions by the National Education Association ("The Racist Beginnings of Standardized Testing", https://www.nea.org/advocating-for-change/new-from-nea/racist-beginnings-standardized-testing) and The Atlantic ("Are Standardized Tests Racist, or Are They Anti-Racist", https://www.theatlantic.com/science/archive/2023/01/should-college-admissions-use-standardized-test-scores/672816/). There has also been an academic discussion of such topics, including by psychometricians (Boykin, 2023) and economists (Card & Giuliano, 2016).

ever, empirically analyzing the relationship between relatability and education performance is challenging. First, student familiarity with topics is inherently qualitative and not typically well-captured in existing datasets. Second, even with a quantitative representation of student background, there is the technical challenge of measuring the presence of topics within exams. Lastly, the topics students are familiar with may be correlated with student ability, which poses a threat to identification.

In this paper, we measure the impact of a student's relatability to a reading passage on standardized test performance. Complementing existing research on representation of identity groups in educational content, we consider the possibility that different racial/ethnic groups have divergent experiential knowledge or interest in topics which may appear in educational content. We develop a measure of "content relatability" for a racial/ethnic group to a piece of text through a novel measure constructed using administrative survey data on demographic differences from the American Time Use Survey (ATUS) and natural language processing (NLP) models. We apply this measure to reading comprehension passages in end-of-year standardized examinations for primary and secondary school students in Texas from 2013–2019, and we test whether content relatability improves student performance. We obtain causal estimates by treating demographic-level exposure to topics as endogenous "shares" and passage-level presence of topics as conditionally exogenous "shocks" in an adaptation of the shift-share instrument estimation framework. Afterwards, we explore whether the realized distribution of reading passages leads to differential test outcomes by race and ethnicity, focusing particularly on whether observed gaps are wider due to content relatability.

We find that the increased relatability of a reading comprehension passage causally raises student performance on questions connected to the passage. We estimate that a standard deviation increase in relatable topics in a passage leads to a 1.7pp increase in the probability of correct answers on that passage. This effect is equivalent to a 0.07 standard deviation increase in passage-level performance and a 0.05 standard deviation increase in test scores.

Since test makers ultimately select passages for inclusion in exams, we consider a variety of placebo outcomes with the same model specification, including prior student performance and demographic composition, and find no effect. Our estimate is also robust to alternative estimation strategies, such as changing the methodology for constructing the relatability measure. While our findings suggest that content relatability has small effects in absolute magnitude, they are moderately sized in comparison to the effects of other factors on test outcomes and test gaps. For instance, Chetty et al. (2014) finds that a standard deviation increase in teacher value-added raises English test scores by 0.1 standard deviations, suggesting that the effect of selecting a one standard deviation more relatable exam is only half that of selecting a one standard deviation higher value-add instructor.

We next investigate the extent to which content relatability contributes to racial disparities in test scores. We find that equalizing content relatability across groups would lead to a 4% smaller Black-white test gap and 4% smaller Hispanic-white test gap. We find that white students have higher average relatability than non-white students which is driven not only by differential exposure to topics but also by the selection of more relatable topics in passages. Instead of equalizing relatability, merely reducing the disproportionate selection of white-relatable topics could close the gap by 1% for both Black and Hispanic students. Results on average test scores differences mask the role of test score thresholds in exacerbating racial disparities. We counterfactually suppose how test scores would adjust if relatability had been set to a level which most closely equalizes relatability differences across race. We show that 1% of Black students in elementary school (grades 3 to 5) could have achieved a higher state-determined, reading comprehension standard if they took a test with more racially equal relatability. Overall, we counterfactually estimate over 11,000 Black students and 15,000 Hispanic students during our sample period were designated to be at a lower reading comprehension level due to relatability.

We contribute to existing work studying demographic representation in student learning materials and a how exam content can impact student performance. Recent work has explored the extent to which certain identities are underrepresented in textbooks using text analysis (Adukia et al., 2023; Lucy et al., 2020). Furthermore, researchers find that educational text can have an impact on students' outcomes and beliefs (Dee & Penner, 2017; Cantoni et al., 2017). Other papers such as Dobrescu et al. (2021) uses an experiment varying the cultural context in a standardized test in Australia, while Dee and Domingue (2021) test the theory of stereotype threat from Steele and Aronson (1995) in practice by looking at the impact of a culturally insensitive test question in Massachusetts. Cohen et al. (2023) finds that use of gender-neutral language in a high-stakes test in Israel improved women's performance on quantitative questions, but no changes in performances for men. In a closely related study to our own, Duquennois (2022) finds that monetarily themed math questions reduce standardized testing performance of low socioeconomic status students. In this context, Duquennois (2022) finds the effect of a 10% increase in financially salient math questions is about 6% of the total test gap between high and low income students, which is similarly sized to our decomposition results. This result is partially attributed to an attention capture mechanism, in which the monetary framing of a test item serves as a reminder of scarcity and leads to inattention. We add to this existing work in several ways. First, we obtain our estimates from a high-stakes exam. Second, we construct a rich measure of relatability that explores the demographic dimensions of race and ethnicity. Third, we propose an innovative identification strategy suited to large-scale exams administered uniformly on a population, whereas the approach of Duquennois (2022) relies on random assignment of digital homework questions or exam booklets.

Our paper relates to two additional strands of literature. First, the measurement and analysis of test score gaps, especially across race, has been a much-discussed topic among researchers and policymakers alike (see Fryer & Levitt, 2004; Fryer & Levitt, 2013; Bond & Lang 2013; Freedle, 2010; Harvard Educational Review, 2010). Recent work by Brown et al. (2022) considers the role of cognitive endurance in the socioeconomic test score gap. Other work considers potential sensitivity of group differences in test scores due to scaling decisions, with particular focus on Black-white test score differences (Bond & Lang, 2013; Nielsen, 2023). We contribute to this existing work by quantitatively examining whether the test content itself can lead to mismeasurement of racial test score gaps. Second, our work contributes to a growing literature which uses text analysis in causal social science research (Gentzkow & Shapiro, 2010; Loughran & McDonald, 2011; Hassan et al., 2017). Natural language processing methods are also increasingly used for analyzing educational content (Adukia, et al. 2023; Bruhn, et al. 2023). In our paper, we apply text analysis methods to high-stakes standardized tests, which serves as a large and important source of educational text data but may be relatively understudied.

We organize the rest of the paper as follows. Section 2 builds a simple conceptual framework which drives our estimation. Section 3 describes the student test data and administrative survey data. Section 4 explains our estimation strategy. Section 5 describes our main results. Section 6 discusses extensions and implications of our results. Finally, section 7 concludes.

2 Conceptual Framework

The primary objective of exams is to measure the ability and progress of students. However, this is complicated by the fact that the signal observed via standardized testing may be a function of learning *and* other factors that the testing administrator may not want to consider. For example, consider a reading passage excerpt from a biography about a sailor. Comparing two students who have identical reading comprehension ability but differ in exposure to boating and experience around the ocean, we may not be surprised to find that the student with greater exposure to the ocean performs better on these questions. The passage topic may help or hinder the ability of students to infer the meaning of vocabulary words or identify the main arguments of the passage. Further, if reading takes mental effort, perhaps the cognitive costs decrease in topical familiarity. If this difference is systematic across demographic groups, this may affect the signals test administrators receive.

To formalize this idea, consider a model of passage-level, student testing outcomes given by

$$y_{ip} = \theta_i + \phi_p + \rho_{ip} \tag{1}$$

where i indexes a student and p indexes passages.[3] Student performance is determined by three factors in this model: θ_i, individual student ability; ϕ_p, general passage difficulty; and ρ_{ip}, a passage-individual-specific term. Suppose we may parse ρ_{ip} into a portion that represents systematic variation, and another part that is idiosyncratic and uncorrelated with θ_i; that is,

$$\rho_{ip} = \underbrace{\vec{\varepsilon}_i' B \vec{\mu}_p}_{\textit{systematic}} + \underbrace{\nu_{ip}}_{\textit{idiosyncratic}},$$

if we model the systematic portion as linear interactions between some observed vector of student characteristics $\vec{\varepsilon}_i$ and an observed vector of passage characteristics $\vec{\mu}_p$. We assume that both $\vec{\varepsilon}_i, \vec{\mu}_p$ are of dimension $T \times 1$, and B is $T \times T$. In the pursuit of tractability, we impose some additional assumptions such that B is zero off the diagonal.

Whereas educators only observe y_{ip}, they will naturally draw inference about student learning from y_{ip}. For instance, educators will traditionally set $\tilde{\theta}_i \equiv \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}} y_{ip}$ as their estimate of student ability, where $\mathcal{P}$ is a set of passages (e.g., the entirety of a single reading comprehension exam). Educators are interested not only in using test results as a marker of individual-level learning but also in using these results to compare learning-level across groups, whether that be classrooms, schools, regions, or racial/ethnic groups. It follows that educators would set $\tilde{\theta}_{\mathcal{G}} \equiv \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}} \tilde{\theta}_i$ as the group-level performance indicator for group $\mathcal{G}$

[3] For ease of exposition, we omit from the model the possibility that each passage constitutes multiple questions, as is the case in most exams. Our framework would be largely unaffected accounting for this reality.

and $\tilde{\theta}_{\mathcal{G}} - \tilde{\theta}_{\mathcal{H}}$ as the difference, or gap, between groups $\mathcal{G}$ and $\mathcal{H}$.

However the testing outcomes model suggests these simple estimates of student- and group-level ability based only on y_{ip} may be biased due to ϕ_p and the systematic components of ρ_{ip}, $\vec{\varepsilon}_i$ and $\vec{\mu}_p$. Bias due to ϕ_p is typically not an issue: If test administrators give the same passages to all students to measure learning in a given grade and school year—as is often the case—there is no bias. Even if passages are differentially administered to students, we expect selection to not be correlated with θ_i. The potential correlation between $\varepsilon_{i,t}$ and θ_i, on the other hand, poses a threat to interpreting standardized test outcomes, $\tilde{\theta}$. Consider again the prior example of a reading passage with a sailor's biography. We can represent the presence of ocean and nautical themes in this passage as an element of $\vec{\mu}_p$, such that $\mu_{p,ocean} = 1$. Student exposure to the ocean and nautical activities may similarly be a component of $\vec{\varepsilon}_i$, thus student i has $\varepsilon_{i,ocean} = 1$ while student i' has $\varepsilon_{i',ocean} = 0$. Then, our model suggests a systematic wedge in observed performance between student i and i' of $\varepsilon_{i,ocean} B_{ocean,ocean} \mu_{p,ocean} - \varepsilon_{i',ocean} B_{ocean,ocean} \mu_{p,ocean} = B_{ocean,ocean}$. If we consider that attributes like exposure to topics are correlated with demographic groups, such as race or income, then group-level conclusions are also affected by this issue.[4]

This model framework and its implications drive our data collection and estimation strategy. First, we limit our focus to elements of $\vec{\varepsilon}_i$—and, by extension, $\vec{\mu}_p$—which are arguably outside the scope of evaluation for reading comprehension exam makers. In our setting, we consider topics which students are exposed to that are orthogonal to reading comprehension such as sports or arts and crafts. We use these topics to define the content relatability of passages to students. Second, while we acknowledge the potential for variation in the presence of these topics to cause differential performance at non-demographic group levels (e.g., baking-interested students compared to their counterparts), we focus on how topic interest and topic selection change test scores at the race/ethnicity-level which is of key interest to educators and researchers. Given our interest in race/ethnicity-level estimation,

[4] Educational psychology research suggests indeed that student interests do diverge by demographic factors and they can meaningfully affect student test performance (Bray & Barron, 2004; Asher, 1979).

one can see that aggregating ρ_{ip} (and implicitly other model objects) from many individuals to the group level, we can have similarly defined object $\rho_{dp} = \vec{\vec{\varepsilon}}_d B \vec{\mu}_p + \nu_{dp}$. This motivates our use of group-level data when constructing estimates of elements of $\vec{\vec{\varepsilon}}_d$. We are also interested in defining different notions of how to design "fairer" tests with respect to this model and our findings. We return to this question in section 6.

3 Data

3.1 Texas student and assessments data

Our primary data comes from the Texas Education Agency (TEA), which administers standardized tests to elementary, middle, and high school students in the state of Texas. We study the State of Texas Assessments of Academic Readiness (STAAR) which are end-of-year assessments taken by public school students in grades 3 through 8 and in high school. While tests are different across grade/course, most students receive the same test within a year.[5]

Within the set of STAAR assessments, we limit our attention to 2013–2019 reading comprehension exams from grades 3 to 8.[6] The reading exam format is standard across grades; students read a handful of text passages and answer multiple choice questions regarding each passage, including the content, vocabulary used, and grammar. We further restrict our analysis to students who take the non-Spanish language, non-alternatively designed STAAR test in a given grade and year.[7] We use the item-level student responses to calculate a binary outcome measure for whether or not a student answered a question correctly. We match each item to its corresponding reading comprehension passage.[8] Ultimately, we obtain 205

[5] Some students may be offered different test versions, such as tests written in Spanish or designed for students with certain cognitive limitations. Barring some minor exceptions, all public school students in Texas are required to take the STAAR test relevant for their grade.

[6] Students in grades 3 to 8 all take reading and math assessments. Students in grade 5 additionally take a science exam, while students in grade 8 also take science and social studies exams.

[7] Spanish langugage tests are only available from grades 3 to 5.

[8] In some cases, questions are associated with two different passages. In these cases, we consider the

unique passages from 42 different exams.[9]

We supplement the testing data with student demographic information including the race of a student and whether they are Hispanic. From the demographic information, we create four racial/ethnic categorizations, Asian, Black, Hispanic, and white. We first drop all students reporting as another race or multiple races. Then, all Hispanic-identifying students are categorized as Hispanic and the remainder of students are categorized according to their reported race.

Our sample consists of 13,180,138 student-exam observations (henceforth, just called "students").[10] Table 1 gives a summary of this student population. The plurality of students in the sample are Hispanic (46%), much higher than the U.S. average.[11] Among non-Hispanic students, the majority are white (34% of total) compared to Black (15%) or Asian (5%). We observe an average of 4.8 passages per exam, corresponding to just under 42 test items. Students answer roughly two-thirds of these questions correctly.

3.2 Time use data

We use additional data from the American Time Use Survey (ATUS) to support our analysis. ATUS, which is sponsored by the Bureau of Labor Statistics and conducted by the U.S. Census Bureau, is a nationally representative survey which provides estimates on how Americans, 15 years and older, spend their time. Randomly selected individuals from a subset of households within the Current Population Study sample report their prior day's activities to a phone interviewer. The interviewer assigns each reported activity to one of 442 six-digit classification codes. Each activity code encompasses 3-tiers of detail regarding

two passages to be one passage for the purpose of analysis. The released test materials are available here: https://tea.texas.gov/student-assessment/testing/staar/staar-released-test-questions

[9] Seven out of 212 reading passages are unavailable due to copyright restrictions. Responses associated with these test passages are removed from the analysis.

[10] Technically a given individual student will appear multiple times inasmuch as they take multiple exams throughout their schooling in Texas. However, since we do not make use of this fact in our paper, we will freely refer to student-exam observations as students.

[11] According to U.S. Census estimates for 2019, 20% of individuals aged 5 to 19 were of Hispanic origin.

Table 1: Summary statistics for student sample

	Mean	SD
Demographics		
Hispanic	0.46	
Non-Hispanic		
Asian	0.05	
Black	0.15	
White	0.34	
Test		
Passages	4.88	0.393
Test items	41.78	5.226
Correct	0.67	0.207
Observations	13,180,138	

Notes: The table displays sample means and standard deviations of characteristics in the student sample. The sample is all students grades 3 to 8 from 2013–2019 taking the standard STAAR reading comprehension test (excluding Spanish-language test takers and "alternative" format test takers).

the activity, with 17 major categories making up the highest tier.[12]

We reduce the dimentionality of this detailed ATUS data for our analysis. We first exclude activity codes that are either unrelated to leisure or activities with which the vast majority of individuals would be expected to be familiar. Appendix Table A1 provides a summary of excluded activity codes. We exclude almost two-thirds of all activity codes, but a substantial share of excluded respondent minutes are for common activities such as sleeping, working, and eating. Sixty-eight percent of excluded respondent minutes are associated with these activities for the average respondent. Next, we classify the remaining 140 activity codes into 25 "topics," including categories such as arts & crafts, animals, soccer, and winter sports (see Appendix Table A2 for examples). A detailed mapping of activity codes to topics is available upon request.

Our ATUS sample includes all respondents from 2013–2019, the same time period as our student data.[13] For each individual, we observe minutes spent on each ATUS activity, as well

[12] For example, "sewing" would be classified as code 020103, with 02 representing the major category of "Household Activities," 0201 representing an activity group of "Housework" within "Household Activities," and 020103 representing the activity of "Sewing, repairing, and maintaining textiles."

[13] We opt to use the full ATUS sample to reduce the potential for small cells to introduce noise in later

Table 2: Summary statistics for ATUS sample

	Mean	SD	5%	95%
Activities				
Number of activities	11.47	4.32	5	19
Number of min spent per done activity	145.28	65.17	73	288
Number of topics	2.49	1.32	1	5
Share of time spent in topics	0.22	0.15	0	1
Demographics				
Asian	0.04	0.20	0	0
Black	0.14	0.35	0	1
White	0.66	0.48	0	1
Hispanic	0.15	0.35	0	1
Observations	73626			

Notes: The table displays sample means, standard deviations, and the 5th/95th percentile value for each category. The sample includes all ATUS respondents from 2013–2019.

as their race, household size, household income range, and other demographic characteristics. We align the racial/ethnic categorization in this data to that of the student data, by excluding multi-race individuals and categorizing Hispanic respondents into their own category.

Table 2 summarizes the ATUS data. On average, a respondent does 11 activities per day, spending 145 minutes per activity.[14] Further, an average respondent does activities within 2 to 3 of our defined topics. This accounts on average to 22% of total time spent on activities within our set of topics. The ATUS sample is predominantly non-Hispanic white, making up 66% of the sample compared to just 34% for the student sample. Much of this difference is due to disproportionately higher Hispanic population in the student sample. The share of Asian and Black respondents is very similar to the shares in the student sample.

analyses. We are able to restrict the sample to groups which may be more "relevant" proxies for children, such as older children and adults (e.g., respondents aged 15 to 30) and adults with children. Our results remain relatively similar when we use these samples. More details on subsamples can be found in Appendix Section B and further discussion of results can be found in Appendix Section **??**.

[14] Dividing the number of minutes in a day by the average number of reported activities does not yield 145 minutes per activity. This occurs because the ratio of averages (average number of minutes divided by average number of activities) is not equivalent to the average of ratios (averaging minutes per activity across respondents).

4 Estimation strategy

4.1 Obtaining content relatability

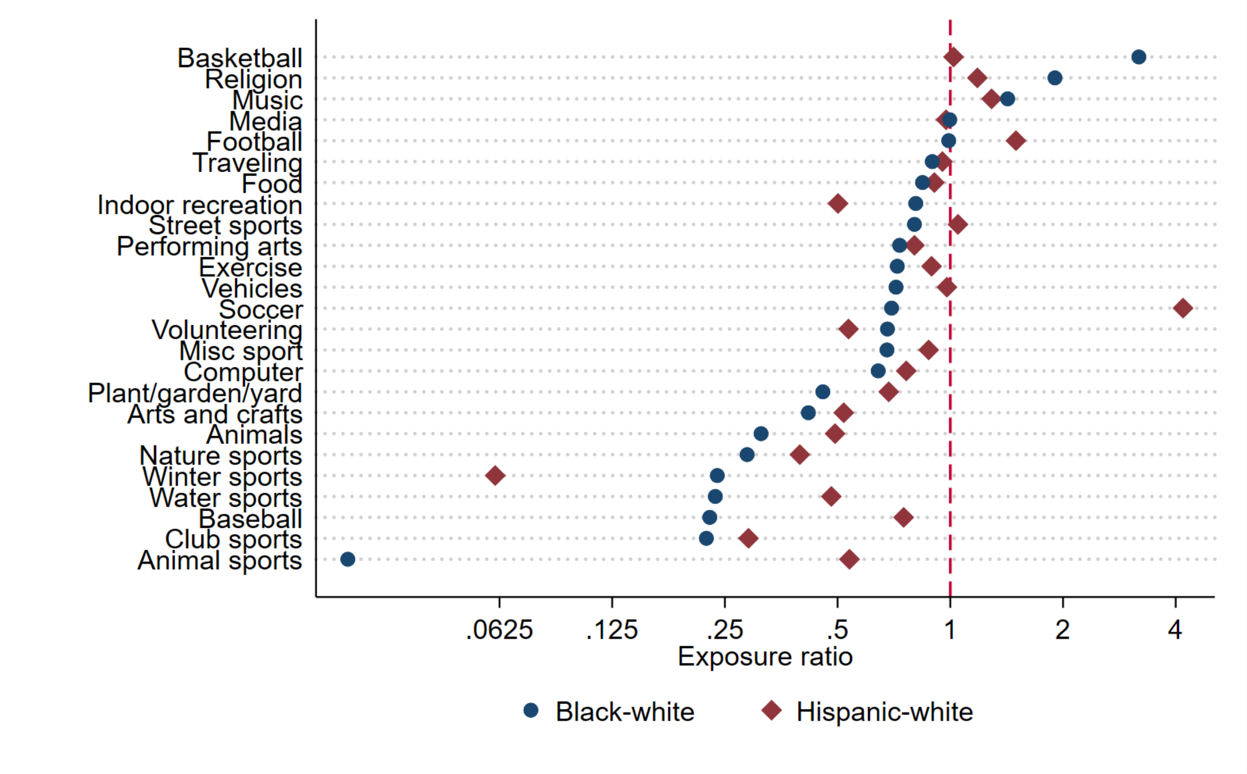
Following the model framework in section 2, our measure of content relatability requires three key components: (a) a vector of race/ethnicity-level elements indicating exposure to a set of topics, corresponding to $\vec{\varepsilon}_d$ in the conceptual framework; (b) a related vector of passage elements indicating salience to the set of topics, corresponding to $\vec{\mu}_p$; and (c) a method which relates these vectors into a single measure, which are restrictions on the B matrix. Using the set of topics described in section 3.2, we construct the first component by measuring *topic exposure* ($e_{d,t}$ for group d and topic t) using the ATUS. We construct the second component by measuring *topic salience* ($m_{t,p}$ for topic t and passage p) using natural language processing (NLP) methods on the text of passages. Finally, we combine these components together into a measure of content relatability, $r_{d,p}$.

4.1.1 Demographic topic exposure

We construct a measure for each demographic group of expected exposure to topics which may be present in student reading passages, corresponding to $\vec{\varepsilon}_d$ in the conceptual framework. Our immediate challenge is that we cannot observe the actual interests and hobbies of student testtakers in our sample. In order to proxy for this, we use reported time use across race from the ATUS data to create a measure of exposure to topics. We reason that even though the majority of ATUS respondents are not school-aged, children are exposed to activities of adults (e.g. parents, siblings) in their household.

Formally, we calculate $e_{d,t}$ simply as the fraction of respondents of demographic group $d \in \mathcal{D}$ who report any activity related to topic $t \in \mathcal{T}$. We also construct alternative measures of topic exposure by repeating this calculation using subsamples of respondents by age, Census region, day of week, and other characteristics. Further details on subsampling is discussed in Appendix Section B.

Figure 1: Racial/ethnicity differences in exposure to topics



Notes: Graph plots the ratio of topic exposures between Black and white respondents and Hispanic and white respondents. The vertical dotted line at 1 corresponds to the value at which both groups have identical exposure. The horizontal axis is presented in a logarithmic scale such that ratios on either side of the dotted line is symmetrical. Topic exposure is calculated separately for each racial/ethnicity group using the ATUS data. Topic exposure is the share of ATUS diaries for a demographic group reporting participating in any activity for a topic. The sample includes all ATUS respondents from 2013-2019.

We demonstrate differential exposure to activities across race and ethnicity in the ATUS data. In Figure 1, we present the Black-white respondent ratio of exposure and the Hispanic-white respondent ratio of exposure.[15] We observe wide divergence in exposure to activities by race and ethnicity. For example, Black individuals are significantly less likely than white individuals to have done a water sports-related activity, but much likelier to have participated in basketball. Hispanic adults are not as exposed to winter sports, but much more likely to be exposed to soccer. We also observe that the topics to which white respondents have relatively higher exposure differ across Black and Hispanic respondents.

[15] Table A3 directly presents the exposure ratios for all topics.

4.1.2 Topic salience in reading passages

We classify the content of each reading passage in the standardized tests. For each passage-topic pair, we use natural language processing (NLP) methods to create a measure representing how salient a topic is in a specific passage, corresponding to $\vec{\mu}_p$ in the conceptual framework. We follow a very intuitive algorithm, a type of dictionary method, which operates as follows. First, determine a set of words which indicate the presence of a topic: a set B_t for each topic $t \in \mathcal{T}$. Second, for each passage $p \in \mathcal{P}$, calculate the fraction of words in a passage that are in a given topic dictionary as our measure of topic salience.

We manually create a dictionary B_t for each topic. After grouping ATUS activities into the topics in $\mathcal{T}$, we separately list as many terms related to the activities within a topic for each topic $t \in \mathcal{T}$ (i.e. we created two dictionaries $B_{t,1}$ and $B_{t,2}$). We then construct $B_t = B_{t,1} \cup B_{t,2}$.[16] On average, each topic's dictionary has 29.7 words. To prepare the reading passages and dictionary terms for analysis, we follow standard steps for text data processing. This includes converting all words to lowercase, removing numbers, and removing stopwords (common words with no informational content in isolation).

To measure topic salience, denoted $m_{t,p}$, we count the number of times some word w is in passage p and denote it $count(w,p)$. Then, we calculate *term-frequency* as $tf(w,p) = \frac{count(w,p)}{|W_p|}$, where W_p is the set of words in passage p (including repeated words). As our main measure of topic salience, we define:

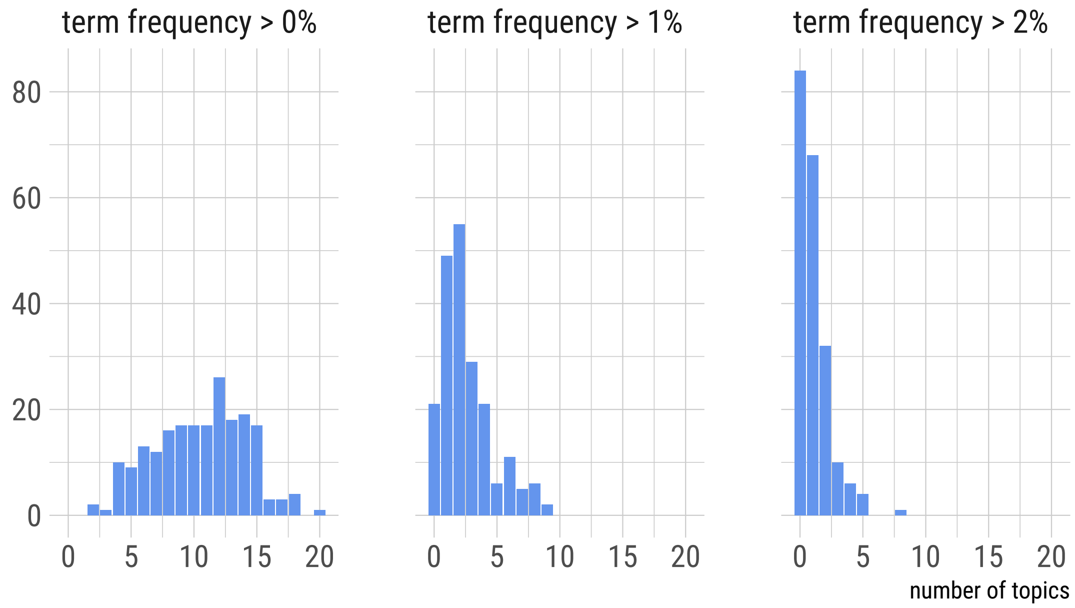
$$m_{t,p} \equiv \sum_{w \in B_t} \mathrm{tf}(w,p) \ .$$

In the distribution of this score displayed in Figure A1, the modal value is zero and is heavily right-skewed. We also consider other measures of topic salience and textual data cleaning, discussed further in Appendix Section B.2.

Considering a binary threshold of $m_{t,p} \geq 0.01$, we find the average passage contains

[16] Dictionaries can be provided upon request.

Figure 2: Number of topics appearing in each passage



Notes: The unit of observation in this figure is a passage. Reading passage measures are calculated for each passage-topic pair by the term-frequency metric discussed in Section 3.3. Using different thresholds, this histogram illustrates how many topics are detected in each passages. The sample of passages include grade 3 to 8 STAAR reading comprehension tests from 2013–2019.

2.54 topics; further, a threshold of 0.02 suggests the average passage contains 1.01 topics. A majority (60.9%) of topic-passages are zero-valued, with the average passage having 9.7 (out of 25) non-zero topics. Histograms illustrating how many topics appear in a passage using each of these thresholds are found in Figure 2. We also present the frequency which a topic appears in the top 3 matches for any passage using our dictionary-based method in Figure A2. This chart indicates that that many passages relate to nature sports (hiking, climbing, fishing, etc); arts and crafts; animals; plants, gardening, and yards; music; and water sports (swimming, boating, etc). Lastly, we validate our topic salience scores by hiring research assistants to manually label the reading passages, which suggest that our dictionary-based scores are capturing the intended variation (see Appendix Section C).

4.1.3 Relatability measure

Our final measure of relatability, $r_{d,p}$ is a straightforward combination of the demographic exposure measure and the passage salience measure. Specifically, given a vector of topic exposure and a vector of topic salience, we relate these vectors using the identity matrix. This choice is implied by the following conditions: (a) the exposure for an index topic does not affect relatability for non-index topics, (b) the effect of a relatable topic on outcomes is equivalent across topics, and (c) exposure and salience affect outcomes similarly for all demographic groups and for all passages. We can write the resulting relatability measure as

$$r_{d,p} \equiv \sum_{t \in \mathcal{T}} e_{d,t} m_{t,p}. \tag{2}$$

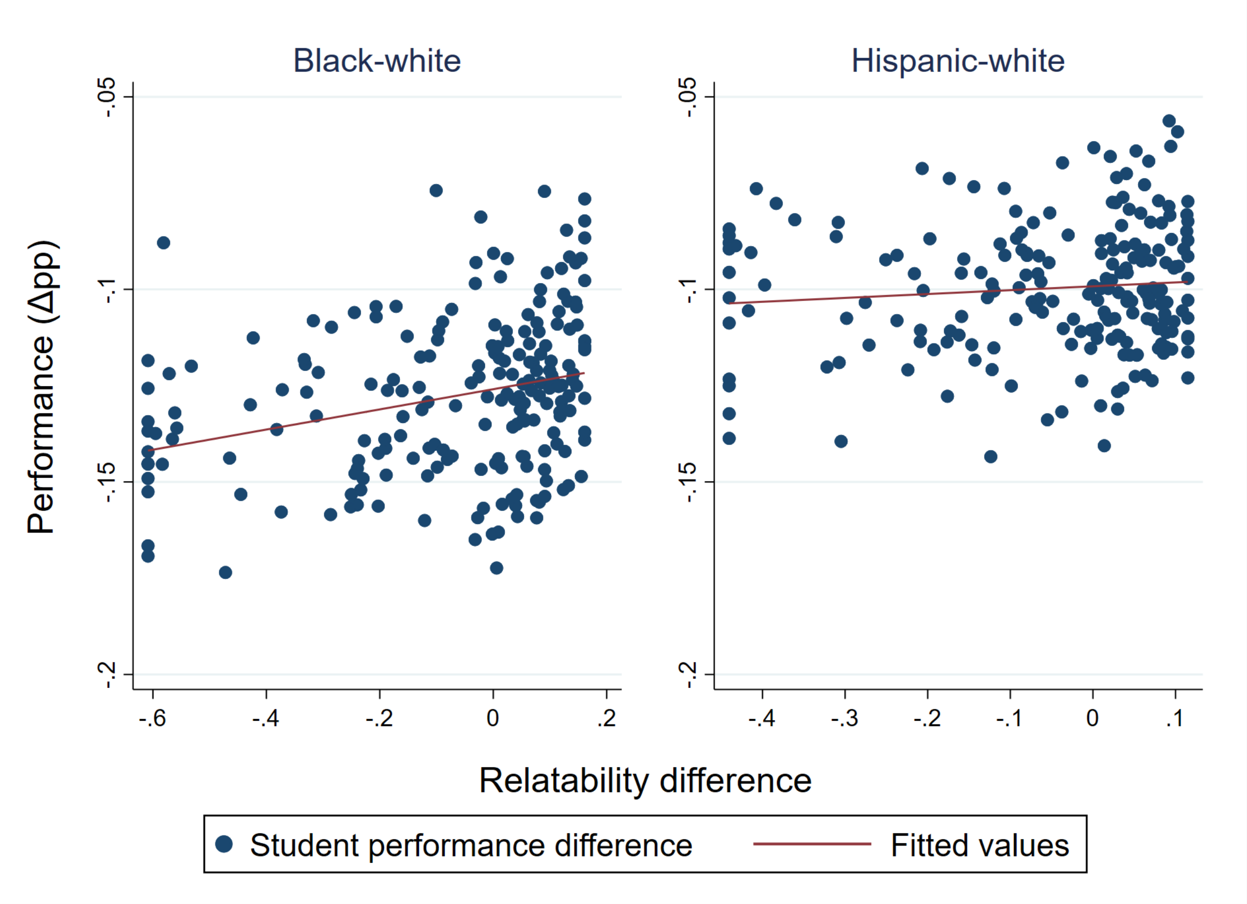
Intuitively, $r_{d,p}$ is a weighted average of topic salience for passage p weighted by the topic exposure for demographic group d.

To show preliminary evidence that our relatability measure is predictive of racial test differences, we standardize our relatability measure and plot the passage-level differential relatability and differential exam outcomes for Black vs. white students and Hispanic vs. white students in Figure 3. We observe a positive relationship between Black-white relatability differences and Black-white test outcome differences, which is suggestive of an impact of relatability on racial test gaps. However, we observe a less clear relationship between relatability and Hispanic-white test score gaps.

4.2 Estimating the causal effect of relatability

Given our relatability definition, consider the following data generating process and how it produces our identifying variation. Consider a test maker who is responsible for creating a 3rd grade exam each year which evaluates student competency on a fixed rubric. Due to state-mandated standards for third grade, she may need to include a poem, two fiction prose and two non-fiction prose passages each year with some fraction of vocabulary questions

Figure 3: Non-white vs. white students test outcome differences and relatability



Notes: Each graph plots average student test outcome differences between non-white and white students against relatability differences and the simple linear fit between these measures. Relatability differences at the passage-level are taken after the relatability measure is standardized. The left-hand side graph plots Black-white student differences while the right-hand side graph plots Hispanic-white student differences. Observations are at the passage-level. Differential relatability is winsorized at the 5th and 95th levels. The estimation sample is all students grades 3 to 8 from 2013–2019 taking the standard STAAR reading comprehension test.

and comprehension questions. While she assembles five passages designed to appeal to third graders, each year there is some amount of residual differences in the relatability of the topics to the student population, which may differ by racial background. Thus, while the topic exposure levels of students are non-random, the topic salience measures within a grade do have an element of randomness across exam years (2013 to 2019, within a grade) and even passages within an exam. We then isolate this residual variation in our regression specification by selecting fixed effects which remove the mean expected relatability score for each demographic group at each grade-level.

More formally, our estimation strategy relies on conditional randomness of topic salience to identify the causal effect of content relatability on student performance. To illustrate this, consider the race/ethnicity-aggregated version of the conceptual model described in equation (1)

$$Y_{dp} = \bar{\theta}_d + \phi_p + \rho_{dp}$$

where Y_{dp} is the share of questions associated with passage p answered correctly by students in group d. As before, we decompose ρ_{dp} into systematic and non-systematic portions, where the systematic portion is driven by content relatability. We therefore have $\rho_{dp} = \vec{\bar{\varepsilon}}_d B \vec{\mu}_p + \nu_{dp} = \beta r_{dp} + \nu_{dp}$ where $B = \beta I$ and I is the identity matrix. The resulting potential outcomes model is

$$Y_d(p) = \bar{\theta}_d + \phi_p + \beta r_{dp} + \nu_{dp}, \tag{3}$$

where β is the impact of content relatability on student performance.

Our goal is to identify β in equation (3) using the variation in topics across passages observed in our student testing data mediated by differential exposure to topics across racial groups. We face two main challenges with this approach. First, $\bar{\theta}_d$ and r_{dp} may be correlated. In particular, we expect r_{dp} to be directly and indirectly correalted with exposure levels e_{dt}. Perhaps students with higher $\bar{\theta}_d$ also have higher aggregate levels of exposure (direct correlation), or $\bar{\theta}_d$ is correlated with individual characteristics (e.g. race and grade-level) while simultaneously the salience (m_{tp}) for some particular topic is correlated with grade-level (indirect correlation). Second, ϕ_p may be correlated with r_{dp}. A passage characteristic such as passage length or passage genre may be associated with higher topic salience, but these characteristics also impact student performance directly.

To account for this potential endogeneity, we identify the impact of relatability on student performance by leveraging only the differences in topic salience (m_{tp}), after conditioning on: (a) grade-specific topic salience means for each topic and (b) passage-specific topic salience means across topics. That is, we account for the possibility that each topic is differentially salient across grades and that salience across topics is higher for certain passages. We assume that this variation is as-good-as-random with respect to other components in the model. Then, given the structure of r_{dp} as a linear interaction between "endogenous" shares (e_{dt}) and exogenous shocks (m_{tp}), we adapt the estimation framework laid out by Borusyak et al. (2022), which allows us to estimate a causal effect by controlling for conditional expected relatability.[17]

In our setting, conditional expected relatability is spanned by controlling for race-grade fixed effects and passage fixed effects interacted with $E_d \equiv \sum_t e_{dt}$. To see the former, consider first that topic exposure e_{dt} is passage-invariant. Thus, expected relatability within a topic-grade is simply given by interacting passage-invariant exposure e_{dt} with average topic salience $\bar{m}_{tg(p)}$, which only varies at the race-grade level. To see the latter, we note that expected relatability conditional on a given passage p is equivalent to the average topic salience for a passage interacted with exposure shares, $\sum_t e_{dt}\bar{m}_p = \bar{m}_p \sum_t e_{dt}$ This value varies at the passage-exposure sum level.

We formally estimate the following regression specification

$$Y_{dp} = \delta_{d,g(p)} + \pi_p E_d + \beta r_{dp} + \nu_{dp}, \tag{4}$$

where $g(p)$ indexes the grade level of passage p and β is the coefficient of interest. $\delta_{d,g(p)}$ are race-grade fixed effects, and π_p are passage-specific fixed effects of the sum of exposure

[17] While analyses that have units with a vector of differential exposure "shares" and a vector of common shocks (typically known as a "shift-share") typically feature the resulting exposure-weighted average of shocks serving as an instrument in a IV/2SLS analysis, Borusyak et al. (2022) explicitly note that identification assumptions follow through if such objects are used in "reduced form" analysis as is the case in our setting.

shares, E_d.[18] Since we have aggregated Y_{ip} to the race-passage level, we estimate the equation with weights representing the number of students and test items which make up (d, p).

We illustrate that race-grade fixed effects $\delta_{d,g(p)}$ and passage fixed effects interacted with exposure sums $\pi_p E_d$ are sufficient for identification using a simple example with two topics $t \in \{\text{soccer, animals}\}$, two groups $d \in \{A, B\}$, and two grades $g \in \{3, 4\}$. Suppose first that group A has higher θ than group B and these differences are greater in grade 3 than grade 4. If group A has higher exposure to soccer compared to group B and soccer is more likely to appear on a grade 3 exam, then θ predicts relatability, leading to bias in estimating β. However, $\delta_{d,g(p)}$ purges variation in relatability arising from differential topic salience across topics and grade, such that relatability is no longer predicted by θ. Now suppose that passage 1 has higher ϕ than passage 2 and also is more about soccer *and* animals than passage 2. Further suppose that group A has higher exposure to both soccer and animals compared to group B. As in the prior situation, ϕ predicts relatability and this biases the estimation of β. Nonetheless, we can directly account for the higher prevalence of topics in passage 1 by including $\pi_p E_d$. Crucially, if we simply included passage fixed effects, ϕ would still predict relatability through differential exposure sums between group A and group B.

Following Borusyak et al. (2022), we estimate standard errors for equation (4) using a topic-passage level regression, which accounts for the fact that groups face common topic salience shocks. Further, since our identifying variation comes from the presence of topics in a passage, we must allow clustering of standard errors within passage. In practice, we more flexibly allow clustering of standard errors within an exam.[19] This approach is motivated by two factors: (a) topic salience is both negatively and positively correlated within

[18] It is possible to estimate a version of equation (4) with passage fixed effects instead of passage interacted with exposure sum. This would yield an analysis featuring "unit" fixed effects at the race-grade level and "time" fixed effects at the passage-level in a canonical two-way fixed effects (TWFE) set-up. While there are appealing estimation properties of employing a commonly used "difference-in-differences" estimation strategy, we argue that the identification assumptions necessary for TWFE are not met in our setting. Crucially, treatment in our setting does not turn "on" and "off" consistently for any unit or time. As such, we also do not observe a unit which experiences a consistent level of relatability which might serve as part of an appropriate comparison group. Nevertheless, we estimate a TWFE version of equation (4) and obtain similar results under this specification.

[19] We obtain slightly smaller standard errors when clustering just within passage.

a passage (e.g., basketball and baseball may be present together) and (b) topic salience is likely negatively correlated within an exam (e.g., test makers may decide not include two passages regarding basketball in the same exam). Further details on the exact procedure used to obtain standard errors are available in Appendix section D.

4.3 Test score gaps

Our main regression in equation (4) returns a coefficient that represents the percentage improvement expected from a one standard deviation increase in content relatability. We investigate the extent to which relatability contribute to observed racial test score gaps.

We calculate the change in test scores if relatability were equalized for all students by predicting $\widehat{Y}_{dp}$ from equation (4) and predicting $\tilde{Y}_{dp}$ after setting $r_{dp} = 0$. These predictions represent predicted outcomes at average relatability and equalized relatability, respectively. We then generate conditional means of each predicted outcome by race: $\widehat{\mu}_d$ for $\widehat{Y}_{dp}$ and $\tilde{\mu}_d$ for $\tilde{Y}_{dp}$. For groups d and d' we can then compute the share of average test score gaps explained by relatability:

$$1 - \frac{\tilde{\mu}_d - \tilde{\mu}_{d'}}{\hat{\mu}_d - \hat{\mu}_{d'}} \tag{5}$$

We compute these measures for Black, Hispanic, and white students and calculate the difference in means.

5 Results

5.1 Impact of relatability on performance

Table 3 shows the results of estimating a regression of test performance on relatability. Column (1) shows results with race fixed effects and column (2) shows results with race-grade fixed effects corresponding to equation (4). We find in both specifications that relatability has

Table 3: Impact of relatability on test performance

	(1)	(2)
Relatability (r_{dp})	0.0174***	0.0166***
	(0.00484)	(0.00510)
Race FEs	X	
Race-by-Grade FEs		X
Outcome mean	.674	.674
No. of student-items	550,633,893	550,633,893
No. of topic-passages	5,125	5,125

Notes: $^{*}p < .10, ^{**}p < .05, ^{***}p < .01$. Each specification is an OLS regression of the share of items answered correctly on a passage on content relatability. Unreported controls include differential slopes on exposure sums by passage. Standard errors reported in parentheses below coefficiene sestimates are (a) obtained using a shock-level regression following Borusyak et al. (2022) and (b) clustered by exam. Observations are at the race-passage level. The estimation sample is all students grades 3 to 8 from 2013–2019 taking the standard STAAR reading comprehension test.

a positive impact on test performance. In our preferred specification, a standard deviation increase in relatability for a passage leads to a 1.66 percentage point increase in the share of items answered correctly for that passage, an effect that is significant at the 99%-level. As is standard in the education literature, we rescale our effect size in terms of standard deviation (SD) changes in student performance and find that our main effect is equivalent to a 0.07 SD increase in passage-level performance. Next, we bring our results to the exam-level, taking into consideration that relatability variation is larger across passages than across exams. We find that a SD increase in average relatability at the exam-level predicts a 0.05 SD unit increase in exam-level student performance.

As described in section 4.2, our specification relies on an orthogonality assumption between relatability and the unobserved error term after including race- and passage-based controls. We identify three ways such an assumption may be violated. First, test makers may incorporate differential performance across groups when building a test for a given year. If the resulting adjustment jointly changes attributes of passages and the topics within passages, then our main estimates may be biased. Second, test makers may adjust the distribution of topics in response to changing underlying demographics of test makers.[20] For instance, an increase in the number of Hispanic students in Texas may lead to more Hispanic-relatable passages in tests and improved in-classroom instruction for Hispanic students. Third, topic selection may be correlated with observable and unobservable passage characteristics which also affect outcomes.

We test for violations of our orthogonality assumption by regressing potential confounders on our relatability measure. We define these confounders consistent with the three violation examples discussed above. To test for test maker responsiveness to prior year performance, we compute one-year lagged average performance for each race-passage dyad rp. Lagged performance is calculated either fixing grade or fixing cohort.[21] This allows for flexibility in how test makers may respond to observing differential performance on a particlarly exam: They may seek to correct (or enhance) these differences in the same-grade exam in the next year, or they may correct the exam for the affected cohort of students in the next year. We then either aggregate lagged performance at the exam level or assign lagged passage performance to rp through matching by passage position. Next, we consider as a confounder the share of testtakers for passage p that identify as race r. Finally, we consider passage characteristics as confounders: calendar year, passage position, word count, and testing category. For the latter measure, we use official reporting categories used by the state of Texas to categorize reading comprehension items and assign passages based on whether it is associated with more items with the "literary texts" category or more test items with the "informational texts" category.[22]

Table A4 shows the results of our balance tests using these confounders. If after account-

[20] Since exams are held at the end of the school year and almost all students are required to take them, it is possible for test makers to either directly observe demographic breakdown of students at the beginning of the school year or project out the breakdown for future years.

[21] Consider, for instance, Black students taking the 6th grade exam in 2017. Lagged performance *within grade* would correspond to performance for Black students taking the 6th grade exam in 2016. Lagged performance *within cohort* would correspond to performance for Black students taking the 5th grade in 2016.

[22] In our study period, items on the reading test are classified in one of three categories: "understanding and analysis across genres," "understanding and analysis of literary texts," "understanding and analysis of informational texts."

ing for race-by-grade fixed effects and period indicators interacted with exposure sums, our relatability measure is quasi-random, then it should not predict these confounders. Indeed, we fail to reject the null hypothesis that relatability predicts 10 of 11 confounders at the 99% significance level. Considering the lagged performance confounders, we find that the coefficient estimates are substantially lower in magnitude than that of our main regression and we fail to reject the null hypothesis that the coefficient on relatability is equal to zero.[23] We similarly find that student population characteristics and most passage characteristics are not predicted by relatability. However, the last row shows that relatability is predictive of being a literary or informational passage. We show in a subsequent robustness specification that our main results remain unchanged while controlling for passage category.[24]

Thus far, our falsification approach is unable to address concerns that there are unobservable confounders which co-vary with our relatability measure. One particular problem may be a passage characteristic which differentially affects performance across racial groups and is also correlated with a certain topic or set of topics. Unless exposure to these topics are equivalent across groups, this issue would bias our coefficient, incorrectly attributing other factors contributing to test performance to content relatability.

We make progress on these issues by repeating our estimation with *non*-racial demographic characteristics. To the extent that unobservable characteristics at the race-passage-level are driving our main result, we can mitigate their influence by removing race from the calculation of relatability. We focus on three characteristics present in both the ATUS and the student data: economic disadvantage, urbanicity, and sex.[25] We limit our analysis to white students, which allows us to remove variation in relatability driven by our race/ethnicity groupings.[26] Separately for each of these characteristics, we recalculate topic

[23] Given how these confounders are defined, these results can be seen as roughly analogous to a "pre-trend" test in a difference-in-differences estimation setting.

[24] We also conduct a regression interacting race-grade fixed effects with passage category and find similar results to our main estimate.

[25] Neither the ATUS nor the student data include markers for gender during the study period.

[26] Using the full sample of students but only utilizing variation in non-race characteristics can still lead to the same bias concerns as previously outlined due to correlation between race/ethnicity, economic disadvantage, and urbanicity.

exposure along the relevant demographic dimension, compute a new relatability measure, and re-estimate equation (4) appropriately replacing all race/ethnicity-based covariates. Further details on this process can be found in Appendix Section B.3.

Table 4 displays the results of this exercise, where each column represents a regression of student performance on relatability for a different demographic grouping method of white students. We find that for estimates relying on economic disadvantage and urbanicity variation, relatability has a statistically significant effect on test performance at the 99% level and both point estimates are similar in magnitude to that of relatability at the race/ethnicity-level in our main specification. We also find smaller and statistically insignificant effect sizes on student performance when relatability is defined at the sex-level. We posit that this may be due to differences in the underlying assumption which drives the construction of the exposure measures across demographic characteristics For race, economic disadvantage, and urbanicity, we allow adult's exposure to proxy for children's exposure under the assumption that because most households are homogeneous in those demographic characteristics, adults simply exert intra-household influence on children's interests. However, since most households are not homogeneous in sex, we must make an added assumption that male adults are only influencing male children and female adults are only influencing female children. Taken together, these results provide suggestive evidence that our main finding that relatability predicts test outcomes is not due solely to unobservable confounders.

Finally, we test if our results are robust to alternative specifications and alternative formulations of our main relatability measure. First, we re-run our main specification replacing race-grade fixed effects with alternative fixed effects. Second, we construct various respondent subsamples of the ATUS data to calculate topic exposure e_{dt} and estimate our main specification with the resulting relatability measure r_{dp}. Third, we employ different NLP methods and measures to calculate topic salience m_{tp} and estiamte our main specification with the resulting relatability measure. Finally, we drop each topic when calculating the relatability measure to test sensitivity to the construction of our topic set. We find our main

Table 4: Effect of non-racial relatability on student performance for white students

	(1)	(2)	(3)	(4)
Relatability	0.0159*** (0.00528)	0.0161*** (0.00485)	0.0150*** (0.00501)	0.004 (0.00350)
Relatability variation	Economic disadvantage	Urbanicity	ED X Urban	Sex
No. of students	190,552,960	190,549,007	190,549,007	190,552,960
No. of topic-passages	5,125	5,125	5,125	5,125

Notes: $^{*}p < .10,^{**} p < .05,^{***} p < .01$. Each specification is an OLS regression of the share of items answered correctly on a passage on content relatability. Content relatability used in each column is calculated based on the groups listed in "Relatability variation". Unreported controls include group-by-grade fixed effects differential slopes on group-level exposure sums by passage. Standard errors reported in parentheses below coefficient estimates are (a) obtained using a shock-level regression following Borusyak et al. (2022) and (b) clustered by exam. Observations are at the race-passage level. The estimation sample is white, non-Hispanic students grades 3 to 8 from 2013–2019 taking the standard STAAR reading comprehension test.

result is robust throughout this battery of exercises. Details on the methodology and results of the robustness analysis is in Appendix Section E.

5.2 Estimating racial test gaps

We observe large average test score differences between non-white and white students on reading comprehension tests. Table A5 displays the Black-white and Hispanic-white test gap by grade, calculated as the percentage point difference in the percentage of items answered correctly. We find that Black students in the 3rd grade have a 13.8 percentage point lower probability of answering a question correctly than white students, a difference which shrinks to 11.9 percentage points by grade 8. The difference for Hispanic students is slightly smaller: between 9 and 11 percentage points across grades.[27] These gaps are economically meaningful, equating to 16-18% and 13-15% of the white student mean for Black students and Hispanic students, respectively.

If content relatability has an impact on student performance and relatability can differ

[27] We observe largely stable test gaps for Hispanic students between 3rd and 5th grade, and then a relatively large jump between 5th and 6th grade. We believe this increase in the test gap is because Spanish-language reading comprehension tests are no longer offered after elementary school.

across students, a natural question is whether differences in relatability contribute to these observed test gaps. Using expression (5) we estimate what our regression results would imply about test score gaps if topic salience, and ultimately relatability, were set to zero. We find that given lower average relatability of Black and Hispanic students compared to white students, racial test gaps are potentially smaller than can be detected using raw student performance. Relatability accounts for 4% of the test score gaps between Black and white students and between Hispanic and white students. Further, relatability is a larger contributor to test score gaps in early grades. In 3rd grade the Black-white and Hispanic-white test gaps are 5% and 6%, respectively, while in 8th grade the gaps are 2%.

As detailed in section 4.3, the implications of these findings could differ based on the source of variation in relatability. If the primary driver of relatability differences is due to differences across groups in levels of exposure, then test makers would have to modify the overall levels of topic salience to mitigate the impact of relatability in exams. Further, this would imply that a planner seeking to purge relatability heterogeneity from test score gaps would need to change student exposure levels directly. However, if a portion of relatability differences cannot be attributed to exposure levels, this implies that a part of the issue is being exacerbated by the selection of topics more relatable to one group over the other.

6 Extensions

6.1 Reconsidering student test standards

Our analyses thus far show that relatability is predictive of performance on reading comprehension tests, leading to inflated calculations of the Black-white and Hispanic-white test gap. We now consider whether passage relatability can disadvantage students directly, through the classification of Black and Hispanic students to different performance standards. Meeting performance standards can have important impacts on future student learning. Failure to reach a certain category could lead to differential classroom assignment or changes to the intensity of resourcing for a student which will affect learning and performance on future tests. In some cases, a student's standards classification could prevent grade promotion, which would substantially affect the learning trajectory of the student.

The TEA annually sets STAAR performance standards which are meant to link raw STAAR test scores to performance categories based on state-mandated curriculum standards. We collect, for every year and grade in our analysis sample, the conversion of raw reading comprehension scores to performance standards categories. Our analysis period spans a period in which performance standards were continually changing, so we adjust the original categories to enable us to consistently aggregate results across years. Further institutional details and process for category adjustment can be found in Appendix Section F.

We test what share of students would have been categorized for a higher standard category if topics—and, therefore, relatability—for a given test had been different. Given that our estimation approach assumes that topic salience may be determined jointly within a given exam, we adjust tests for passage relatability that hews to this assumption. First, we calculate the average relatability for each test by race. Next, we identify a test within each grade with (a) the lowest Black-white average relatability difference and (b) the lowest Hispanic-white average relatability difference. For each non-"benchmark" test, we calculate the difference between relatability for that test compared to this "benchmark" test and adjust overall scores based on the predicted relationship between relatability and item correctness from estimating equation (4). Formally, we calculate

$$score'_i = score_i + N_{g(i)m(i)}\hat{\beta}\left(\tilde{r}_{g(i),d(i)} - \bar{r}_i\right), \tag{6}$$

where i indexes student-exams, $g(i)$ indicates the grade level, $d(i)$ is the student's race, and $m(i)$ indicates the calendar year of the exam. $score_i$ is the score for the student on the exam, $N_{g(i)m(i)}$ is the number of questions on the exam, and $\hat{\beta}$ is the estimated coefficient from equation 4. $\tilde{r}_{g(i),d(i)}$ is the average relatability for $d(i)$ on the benchmark exam for grade

$g(i)$ and $\overline{r}_i$ is the average relatability for the student's exam.

We assign students to performance standard categories based on this new, relatability-adjusted score, $score'_i$. Since raw scores can only be a whole number, students with non-integer $score'_i$ are partially attributed to the two nearest integers. Specifically, we assign $(\lceil score'_i \rceil - score'_i)$% of a student to score $\lfloor score'_i \rfloor$ and the rest to $\lceil score'_i \rceil$. For example, if 100 students have a computed score of 30.4, our method essentially counts 40 students having an adjusted score of 30 and 60 students having an adjusted score of 31. After these adjustments, we assign students to one of four performance standards categories..

Tables 5 and 6 summarizes the results of applying this exercise to a 1% sample of Black students and Hispanic students, respectively.[28] Each row corresponds to a different performance standards category and the columns are organized by grade. Each cell shows the mean difference in the share of students labelled in the performance category between the adjusted and unadjusted exams. We find that the benchmark test would have led to fewer Black and Hispanic students being classified as not meeting each of the three standards. Across all standards, on average 1.1% more Black students in Grade 3 would have been placed in a higher performance category with relatability adjustments. While effects are larger for lower grades, relatability still has large impacts for some exams in higher grades; for example, in one 7th grade exam, we find 1.5% more Black students would have been placed in a higher performance category with relatability adjustments. Adjustments have a smaller impact for Hispanic students, with only an average of 0.1% to 0.6% of Hispanic students experiencing upwards adjustments in performance categories with relatability changes. Extrapolating our results to the full sample, we estimate that over 11,000 Black students and over 15,000 Hispanic students may have achieved a higher reading comprehension standard if relatability had been more equalized across groups.

The results demonstrate that even if relatability has modest effects on average scores in the aggregate, the effect is strong enough to reclassify students as meeting or not meeting

[28] While we have access to the full sample of student results, we are only able to use a 1% sample for individual student analyses due to external data constraints.

Table 5: Effect of balancing relatability across tests on share of Black students meeting performance standards

	Grade 3	4	5	6	7	8
Below Approaches	-0.003	-0.003	-0.003	-0.002	-0.003	-0.001
Below Meets	-0.004	-0.003	-0.003	-0.002	-0.003	-0.001
Below Masters	-0.004	-0.002	-0.002	-0.001	-0.002	-0.001

Notes: Numbers show the change in the share of students in a given performance standards category by rescaling relatability to a benchmark test within the grade. Observations are at the test-level. The sample starts with a 1% draw of all students grades 3 to 8 from 2013–2019 taking the standard STAAR reading comprehension test. Students taking the benchmark test within each grade is excluded.

Table 6: Effect of balancing relatability across tests on share of Hispanic students meeting performance standards

	Grade 3	4	5	6	7	8
Below Approaches	-0.002	-0.002	-0.001	-0.000	-0.001	-0.000
Below Meets	-0.003	-0.002	-0.001	-0.001	-0.001	-0.000
Below Masters	-0.002	-0.001	-0.000	-0.000	-0.001	-0.000

Notes: Numbers show the change in the share of students in a given performance standards category by rescaling relatability to a benchmark test within the grade. Observations are at the test-level. The sample starts with a 1% draw of all students grades 3 to 8 from 2013-2019 taking the standard STAAR reading comprehension test. Students taking the benchmark test within each grade is excluded.

performance standards. Further, even though our results suggest smaller relatability-based distortions as students progress in grade-level, our single-year analysis masks potential cascading effects due to misclassification in earlier years. Improper classification in earlier grades could meaningfully change the learning trajectory of students.

6.2 Counterfactual policies

We consider a few policies that the test maker may implement in the context of our model and estimates. In light of our discussion thus far, the most obvious goal a policymaker may want to consider is to reduce the influence of relatability on test scores, thereby more precisely measuring student ability, θ. Given racial differences in relatabilty, these changes to exams would ensure that estimates of θ are not systematically divergent by race. A policymaker may first consider simply equalizing relatability across different demographic groups. This can be expressed formally, using our model's notation from section 2, as choosing some $\vec{\mu}^*_{\text{rel}}$, representing the expected value of the distribution of topics being drawn from, such that

$$\vec{\mu}^*_{\text{rel}} = \arg\min_{\vec{\mu}} \left| \sum_t \left(\varepsilon_{dt} - \varepsilon_{d't} \right) \mu_t \right|$$

for group d and group d'. In theory, if the set of feasible μ_t values is unconstrained, the effect of implementing $\vec{\mu}^*_{\text{rel}}$ is equivalent to the results obtained in section 5.2 regarding the contribution of relatability to racial test gaps. We find that if relatability is theoretically equalized in this way, policymakers could observe 4% smaller test score gaps between Black and white students and between Hispanic and white students. Test makers can achieve this either by selecting passages free of our studied topics or selecting passages that on average generate relatability for all demographic group at some level r^*.

In reality, however, test makers may face constraints on the possible values of μ_t. For instance, our analysis thus far does not take into account the "supply" of passages from which test makers can choose when constructing exams or the presence of additional political or educational considerations which influence topic selection in exams.[29] To account for this possibility, we repeat a version of the exercise in section 6.1. First, we search for the exam within each grade which minimizes the nonwhite-white gap in average relatability and designate the race-level relatability on these tests to be the feasible relatability for their respective grades. Second, for all other exams in a grade, we calculate counterfactual test scores under the scenario where the relatability measure was changed to the feasible relatability measure. We estimate 2% smaller Black-white and Hispanic-white test gaps through this procedure. Since the relatability adjustments in this exercise reflect actual

[29] Authors of children's stories or books may themselves be from a selected population. Their exposure to topics as a child or an adult may influence the topics they write in the work, which may then mean the corpus of selectable passages is already biased prior to selection by test makers. This may further be exacerbated (mediated) if there are considerations which prioritize Texan writers or stories that are based primarily in Texas.

relatability measures we observe in our sample, we consider the results of this exercise to demonstrate what is feasible for policymakers even in the presence of existing constraints on topic seleciton in passages.

Alternatively, a test maker could object to the previous goal of equalizing relatability as it may "disadvantage" students with higher aggregate topic exposure. While a policymaker who prefers the earlier approach may claim that different levels of aggregate exposure are due to differential constraints across groups (e.g. differences in financial resources), this test maker may want to be agnostic about that heterogeneity. Thus, they may prefer to focus on adjusting the topic distribution such that a student from either demographic group is equally likely to see a topic with which they are either very familiar or very unfamiliar. Formally, this is optimizing for some $\vec{\mu}^*_{\text{topic}}$ such that

$$\vec{\mu}^*_{\text{topic}} = \arg\min_{\vec{\mu}} \left| \sum_t \left(\frac{\varepsilon_{dt}}{\sum_t \varepsilon_{dt}} - \frac{\varepsilon_{d't}}{\sum_t \varepsilon_{d't}} \right) \mu_t \right| .$$

Intuitively, this will result in selecting a topic distribution that equalizes the probability of choosing a topic that is very (un)popular for a student of demographic group d or d'.

To illustrate this precisely, we decompose the empirical relatability differences across groups in our sample into a portion that is due to the topic distribution as opposed to differences in topic exposure. For this exercise, we use the objects m, e (see section 4) which are the empirical analogues to μ, ε respectively. We decompose cross-group differences in average relatability as

$$\bar{r}_d - \bar{r}_{d'} = (E_d - E_{d'})\bar{m} + \frac{\sum_p \sum_t (e_{dt} - e_{d't})\tilde{m}_{tp}}{|\mathcal{P}|} , \tag{7}$$

where $\bar{m}$ is average topic salience across all topics and passages, $\tilde{m}_{tp}$ is the residual of m_{tp} from $\bar{m}$, $|\mathcal{P}|$ is the number of passages, and, as before, E_d and $E_{d'}$ are the sums of exposures across topics. The first term of the right-hand side of equation (7) represents differences in relatability due to differences in overall levels of exposure. The second term of the question represents differences in relatability due to selection of differentially favorable topics in passages.[30] Intuitively, we can think of the first term as reflecting the fact that if overall exposure is higher for one group than another, then any randomly selected topic will lead to some baseline difference in relatability on average. However, if the second term is non-zero, any differences in relatability over and beyond this baseline difference must be due to topic selection being skewed toward one group over the other.

Applying equation (7) to our data, we find that the second term contributes to just under one-third of Black-white (32%) and Hispanic-white (33%) differences in average relatability. Said differently, if test makers were to optimize for $\vec{\mu}^*_{\text{topic}}$, relatability differences would be almost a third smaller than currently observed relatability differences by race and ethnicity. Combined with our original estimate of β, this implies that test makers could theoretically close 1% of both nonwhite-white test score differences by optimizing for $\vec{\mu}^*_{\text{topic}}$.

7 Conclusion

We study the extent to which the sociocultural content in standardized tests impact performance using end-of-year reading comprehension exams for 3rd to 8th grade students in the state of Texas. We develop a measure of "relatability" between a reading comprehension passage and a demographic group using natural language processing methods and time use data from the American Time Use Survey. We find that our measure is predictive of students' standardized test performance—a standard deviation higher race-based relatability for a passage leads to a 1.7pp increase in probability of answering questions for that passage correctly. Given differences in average content relatability across race and ethnicity, we find that relatability accounts for 4% of the Black-white and Hispanic-white test gaps.

Our results have implications both for test writers and education policymakers. First, it highlights that in order to write balanced assessments, test makers should take into account

[30] We note since $\tilde{m}_{tp}$ is a residual, it is possible for this term to be negative even if $e_{dt} - e_{d't} > 0$, $\forall t$.

not only the identities of characters, but also the general content of the passage or question itself as we show that this may influence performance. Second, when policymakers consider outcome differences along demographic dimensions, one additional component to examine might be the standardized tests used to calculate those differences. However, we also note that the contribution of test construction to the gaps we find are both non-negligible and modest; that is, they cannot explain a substantial portion of why Black and Hispanic students on average have lower performance on tests than white students.

Finally, an alternative interpretation of our results may be that adaptability to different environments and new concepts is an important part of student learning. To test this ability in standardized tests, students should read passages on topics with which they are unfamiliar, captured in our case by "relatability." We contend our findings still have meaningful policy-relevant implications in this scenario. Recast in this light, our main result demonstrates that students on average are not fully "adapatable" given that on average we are able to predict they will perform worse on topics with which they are less familiar. This suggests that education curriculum should put more emphasis on teaching students skills to be "adapatable." Further, our racial gaps results demonstrate that if test makers seek to test the ability of students to be interested in concepts and settings unfamiliar to them, they must incorporate the fact that familiarity differs by demograhpic group; an unfamiliar topic to one group may be a familiar topic to another. Ultimately, our preferred interpretation of our results stems from the fact that the stated goals of most reading comprehension exams do not include testing for breadth of topic knowledge or whether students are adaptable to unknown topics. Insofar as stated testing standards reflect the knowledge and skills educators truly expect students to have, we take these standards seriously in our conclusions and set aside any ancillary skills that educators would like to test.

References

Adukia, A., Eble, A., Harrison, E., Runesha, H. B., & Szasz, T. (2023). What We Teach About Race and Gender: Representation in Images and Text of Children's Books. *The Quarterly Journal of Economics, 138*(4), 2225–2285.

Asher, S. R.Influence of topic interest on Black children's and White children's reading comprehension. *Child Development, 50*(3), 686-690.

Bond, T. N., & Lang, K. (2013). The Evolution of the Black-White Test Score Gap in Grades K–3: The Fragility of Results. *The Review of Economics and Statistics, 95*(5), 1468–1479. doi: https://doi.org/10.1162/REST_a_00370

Borusyak, K., Hull, P., & Jaravel, X. (2021). Quasi-Experimental Shift-Share Research Designs. *The Review of Economic Studies, 89*(1).

Boykin, C. M. (2023). Constructs, Tape Measures, and Mercury. *Perspectives on Psychological Science, 18*(1), 39–47. https://doi.org/10.1177/17456916221098078

Bray, B. G. & Barron, S. (2004). Assessing reading comprehension: The Effects of Text-based Interest, Gender, and Ability. *Educational Assessment, 9*(3-4), 107-128.

Brown, C. L., Kaur, S., Kingdon, G., & Schofield, H. (2022). Cognitive Endurance as Human Capital (Working Paper No. 30133; Workign Paper Series). National Bureau of Economic Research. http://www.nber.org/papers/w30133

Bruhn, J. M., Gilraine, M., Ludwig, J., & Mullainathan, S. (2023). What's in a Question? Harnessing the Information Value of Item Response Data to Better Capture and Represent Learning. *Manuscript in preparation.*

Cantoni, D., Chen, Y., Yang, D. Y., Yuchtman, N., & Zhang, Y. J. (2017). Curriculum and Ideology. *Journal of Political Economy, 125*(2), 338–392. https://doi.org/10.1086/690951.

Card, D., & Giuliano, L. (2016). Universal Screening Increases the Representation of Low-Income and Minority Students in Gifted Education. *Proceedings of the National Academy of Sciences, 113*(48), 13678–13683. https://doi.org/10.1073/pnas.1605043113

Chetty, R., Friedman, J. N., & Rockoff J.E. (2014). "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review, 104*(9), 2593-2632.

Cohen, A., Karelitz, T., Kricheli-Katz, T., Pumpian, S., & Regev, T. (2023). Gender-Neutral Language and Gender Disparities (Working Paper No. 31400; Working Paper Series). National Bureau of Economic Research. https://doi.org/10.3386/w31400.

Dee, T. S., & Domingue, B. W. (2021). Assessing the Impact of a Test Question: Evidence from the "Underground Railroad" Controversy. *Educational Measurement: Issues and Practice, 40*(2), 81–88. https://doi.org/10.1111/emip.12411.

Dee, T. S., & Penner, E. K. (2017). The Causal Effects of Cultural Relevance: Evidence From an Ethnic Studies Curriculum. *American Educational Research Journal, 54*(1), 127–166. https://doi.org/10.3102/0002831216677002.

Dobrescu, L. I., Holden, R., Motta, A., Piccoli, A., Roberts, P., & Walker, S. (2021). Cultural Context in Standardized Tests. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.3983663.

Duquennois, C. (2022). Fictional Money, Real Costs: Impacts of Financial Salience on Disadvantaged Students. *American Economic Review, 112*(3), 798–826. https://doi.org/10.1257/aer.20201661.

Freedle, R. (2010). On Replicating Ethnic Test Bias Effects: The Santelices and Wilson Study. *Harvard Educational Review, 80*(3), 394–404. https://doi.org/10.17763/haer.80.3.l050025058204016

Fryer, R., & Levitt, S. (2004). Understanding the Black-White Test Score Gap in the First Two Years of School. *The Review of Economics and Statistics, 86*(2), 447–464. https://doi.org/10.1162/003465304323031049

Fryer, R. & Levitt, S. (2013). Testing for Racial Differences in the Mental Ability of Young Children. *American Economic Review, 103*(2), 981-1005.

Gentzkow, M., & Shapiro, J. M. (2010). What Drives Media Slant? Evidence from U.S. Daily Newspapers. *Econometrica, 78*(1), 35–71.

Hassan, T. A., Hollander, S., van Lent, L., & Tahoun, A. (2017). Firm-Level Political Risk: Measurement and Effects (Working Paper No. 24029; Working Paper Series). National Bureau of Economic Research. https://doi.org/10.3386/w24029.

Loughran, T., & Mcdonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance, 66*(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x.

Lucy, L., Demszky, D., Bromley, P., & Jurafsky, D. (2020). Content Analysis of Textbooks via Natural Language Processing: Findings on Gender, Race, and Ethnicity in Texas U.S. History Textbooks. *AERA Open, 6*(3). https://doi.org/10.1177/2332858420940312.

Nielsen, E. (2023). How Sensitive are Standard Statistics to the Choice of Scale?. *Working paper.* https://drive.google.com/file/d/13QDzIHbpm1T5QE7Np4_4oRxvtn3GNz14/view

Steele, C., & Aronson, J. (1995). Stereotype Threat and The Intellectual Test-Performance of African-Americans. *Journal of Personality and Social Psychology, 69*, 797–811. https://doi.org/10.1037/00 3514.69.5.797.

SYMPOSIUM: Bias in the SAT? Continuing the Debate. (2010). *Harvard Educational Review, 80*(3), 391–394. https://doi.org/10.17763/haer.80.3.l1678014u04m3504.

Willett, P. (2006). The Porter stemming algorithm: Then and now. *Program electronic library and information systems, 40*, 10.1108/00330330610681295.

A Appendix tables and figures

Table A1: Summary of ATUS activity codes not selected for analysis

	(1) Mean Min./Respondent	(2) # of activity codes
All activities	1427	442
Excluded activities	1109	302
Sleeping	534	2
Personal care	46	4
Child care	39	66
Work	156	22
Education	14	22
Shopping	24	10
Eating	66	4
Telephone	7	11
Traveling (non-leisure)	58	51
Other	165	110
Included activities	319	140
Observations	73626	73626

Sample includes all ATUS respondents from 2013–2019. Each row represents the statistics for each group of activity codes. Column 1 represents the average reported minutes per respondent for the group of activities. Column 2 represents the total number of activity codes for the group of activities.

Table A2: Examples of ATUS activities in each topic

Topic	Example activities	# of six-digit activity codes
Animal sports	(*equestrian, rodeo*)	4
Animals	(*caring for pets, going to the vet*)	9
Arts and crafts	(*sewing, decorating*)	5
Baseball		4
Basketball		2
Computer		1
Exercise	(*running, lifting*)	10
Food	(*baking, cooking*)	4
Football		2
Indoor recreation	(*billiards, bowling*)	4
Media	(*movies, TV*)	2
Misc. sports		21
Music		2
Nature sports	(*kayaking, fishing, climbing*)	6
Performing arts	(*musicals, dancing*)	4
Plant/garden/yard	(*gardening*)	3
Religion	(*attending church*)	7
Club sports	(*golf, tennis*)	6
Soccer		2
Street sports	(*skateboarding, scootering*)	6
Traveling		2
Vehicles	(*fixing car*)	3
Volunteering		25
Water sports	(*swimming, water polo*)	5
Winter sports	(*skiing, ice skating*)	4

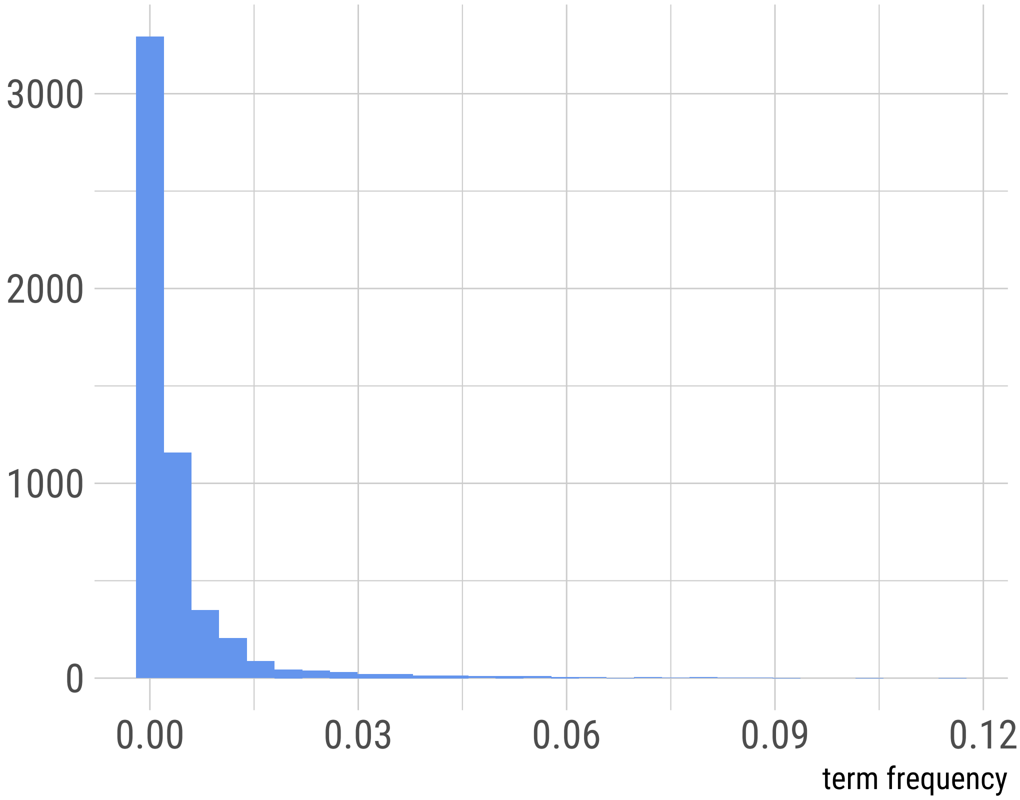
Each topic is associated with a set of ATUS activities/activity codes. A full mapping of activities to topics are available upon request.

Table A3: Relative exposure of groups to topics

	Asian-white	Black-white	Hispanic-white
Animal sports	0.000	0.025	0.538
Animals	0.328	0.312	0.492
Arts and crafts	0.573	0.418	0.519
Baseball	0.310	0.228	0.751
Basketball	0.991	3.189	1.020
Club sports	0.851	0.223	0.289
Computer	1.491	0.642	0.763
Exercise	1.381	0.722	0.891
Food	1.023	0.843	0.906
Football	0.305	0.990	1.497
Indoor recreation	0.558	0.809	0.501
Media	0.883	0.997	0.975
Misc sport	1.092	0.677	0.875
Music	1.147	1.423	1.289
Nature sports	0.261	0.287	0.396
Performing arts	0.796	0.732	0.802
Plant/garden/yard	0.675	0.457	0.685
Religion	1.218	1.903	1.181
Soccer	1.371	0.697	4.184
Street sports	1.639	0.802	1.048
Traveling	0.858	0.895	0.952
Vehicles	0.598	0.716	0.979
Volunteering	0.507	0.679	0.535
Water sports	0.854	0.236	0.481
Winter sports	0.704	0.239	0.061

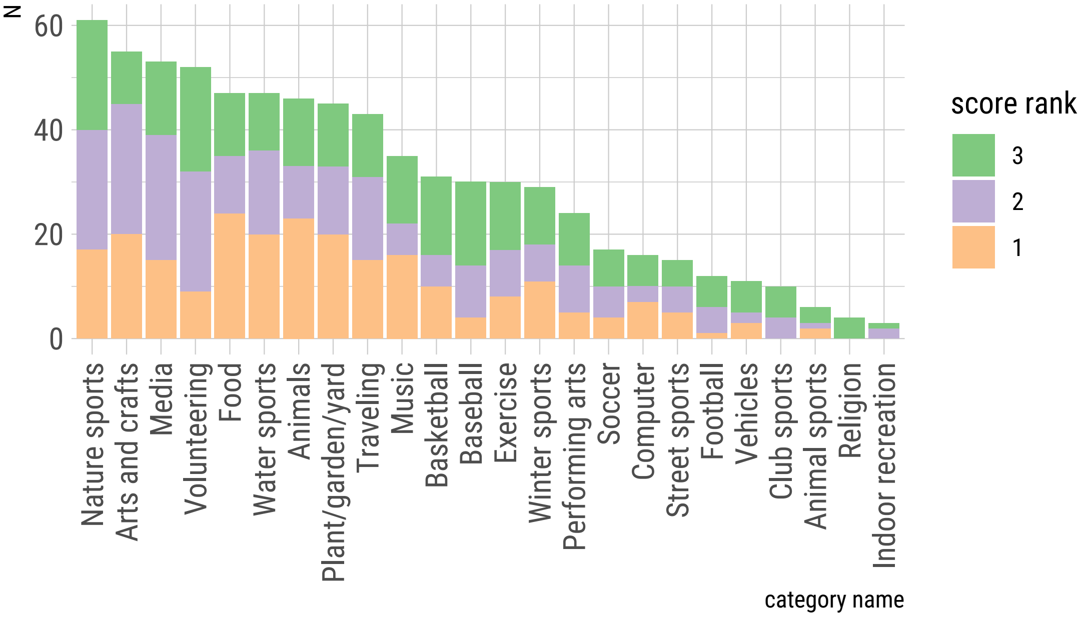
Each cell represents a ratio of topic exposures between two demographic groups. Exposure is the share of ATUS diaries for a demograhpic group reporting participating in any activity in the topic. For racial groups, exposure for white respondents is used as the comparison group. Economic disadvantage is defined as being below 185% of the poverty line. Topics are formed combining 6-digit ATUS activity codes. Topics are mutually exclusive but not completely exhaustive. The sample includes all ATUS respondents from 2013–2019.

Figure A1: Histogram of the topic salience score



Notes: This histogram is constructed using topic-passage level observations. Details on the term-frequency metric used here can be found in Section 3.3. The sample of passages include grade 3 to 8 STAAR reading comprehension tests from 2013–2019.

Figure A2: Most frequent topics in STAAR reading passages



Notes: Reading passage measures are calculated for each passage-topic pair by the term-frequency metric discussed in Section 3.3. Keeping the three topics with the highest score for each passage, this histogram shows how frequently each topic is detected in the passages. The sample of passages include grade 3 to 8 STAAR reading comprehension tests from 2013–2019.

Table A4: Effect of relatability on falsification outcome variables

	Coeff.	SE	N of race-passage dyads
Prior year performance...			
by grade-race	-0.00286	(0.00307)	700
by cohort-race	0.00293	(0.00281)	592
by grade-race-passage position	0.00385	(0.00604)	688
by cohort-race-passage position	0.00520	(0.00804)	568
Previous passage perf.	0.00406	(0.00706)	640
Subsequent passage perf.	0.000301	(0.00561)	640
Population of race by exam	-0.00347	(0.00212)	820
Exam year (continuous)	-0.0694	(0.0927)	820
Passage position (continuous)	0.0346	(0.0593)	820
Passage word count	5.637	(8.353)	820
Literary passage	0.0625***	(0.0160)	820

Notes: $^{*}p < .10,^{**}p < .05,^{***}p < .01$. Each row reports the coefficient from a regression of the specified outcome variable on relatability. Each specification includes race-by-grade fixed effects and passage fixed effects interacted with exposure sums. Standard errors reported in parentheses adjacent to the estimates are (a) obtained using topic-passage-level regressions following Borusyak et al. (2022) and (b) clustered by exam. The estimation sample differs for each specification; it includes all students grades 3 to 8 from 2013–2019 taking the standard STAAR reading comprehension test for which the outcome measure is available.

Table A5: Black-white and Hispanic-white test score gaps by grade

	Grade					
	3	4	5	6	7	8
Black-white test gap (pp.)	-0.136	-0.138	-0.128	-0.128	-0.122	-0.119
Hispanic-white test gap (pp.)	-0.094	-0.094	-0.094	-0.111	-0.104	-0.100
No. of students	9,800,300	10,283,570	10,705,640	11,498,840	11,100,380	10,964,120

Notes: Each column displays the test score gap between different racial/ethnic groups for a given grade. Test score gaps are calculated as the percentage point difference between two groups' average performance on every test item. The sample includes all students grades 3 to 8 from 2013–2019 taking the standard STAAR reading comprehension test.

B Construction of additional variables

B.1 Additional racial familiarity measures

In addition to our standard measure of racial familiarity, we construct additional measures through various restrictions to the ATUS sample. We follow a simple procedure which takes a subsample of ATUS respondents and recalculates content familiarity from that subsample, using the process described in section 4.1.1.

We create subsamples of the ATUS sample across four dimensions: (a) weekend/weekday survey response, (b) U.S. Census region, (c) age, and (d) parental status. First, since ATUS respondents give a diary of a single day, if there is heterogeneity in activities between the weekend and weekday, this may affect our relatability measure. We split the sample by whether the respondent's diary day is for Saturday or Sunday (weekend) or Monday through Friday (weekday). Second, while we leverage data respondents from across the country for predictive power, Texans may behave substantially differently than other Americans. To account for this possibility, we split the sample by whether or not respondents are in the "South" U.S. Census Region which includes Texas[31] Third, there may be differences across age in activity participation. We create the following five subsamples by age: 15-18, 15-34, 35-49, 50-64, and 65-85. The first subsample reflects data from school-age individuals, while the latter four subsamples are mutually excluive, almost completely exhaustive age ranges that roughly split the sample into quartiles. Finally, if we are using adults to proxy for the familiarity of children, our prediction may be more accurate for adults with children. Thus, we identify respondents with children and respondents without children.

Ultimately, after creating exposure measures for the subsamples described above, we recalculate relatability measure, r. We ultimately create 12 alternative relatability measures: $r^{weekend}$, $r^{weekday}$, r^{south}, $r^{nonsouth}$, $r^{childage}$, $r^{age15-34}$, $r^{age35-49}$, $r^{age50-64}$, $r^{age65-85}$, r^{parent},

[31] Other states in the South Census Region include Oklahoma, Arkansas, Louisiana, Mississippi, Tennessee, Kentucky, Alabama, Florida, Georgia, North Carolina, Virginia, South Carolina, West Virginia, Maryland, and Delware, as well as Washington, D.C.

$r^{parent|childage}$, $r^{nonparent\&nonchild}$.

B.2 Alternative measures of salience

While term frequency is our preferred definition of topic salience, we also consider two other definitions. The simpler being the *count* metric, which is simply $m_{t,p} \equiv \sum_{w \in B_t} \text{count}(w,p)$ where $\text{count}(w,p) \equiv \sum_{v \in W_p} \mathbb{1}[v = w]$ and W_p is the set of words in passage p (including repeated words). Another we consider adds a weight for "uniqueness" of a word. We can multiply the term frequency by an *inverse document frequency* measure, defined as $\text{idf}(w) \equiv \log\left(\frac{|\mathcal{P}|}{\sum_{p \in \mathcal{P}} \mathbb{1}[w \in W_p]}\right)$, which is zero for words that appear in all passages and emphasizes words that are less commonly used across passages. Here, we define $m_{t,p} \equiv \sum_{w \in B_t} \text{tf}(w,p) \cdot \text{idf}(w)$.

We also consider many variations of the dictionary method such as "stemming" the words, using only nouns, and removing words from the dictionary.[32] Futher, while our empirical strategy leverages the intensive margin variation in the methods described so far, we do consider discrete shocks of topics by setting a topic shock to 1 if it is in the top 1/25th of m_{tp} within the grade-level and 0 otherwise. Thus, on average each passage will be about one topic but some passages will be about nothing and others will have multiple topics.

B.3 Analysis leveraging non-racial relatability differences

The ATUS data and student data have additional measures of individual demograhpic characteristics beyond race: economic disadvantage, urbanicity, and sex. We outline below a process which allows us to leverage this non-racial variation to construct new measures of exposure and relatability. These additional measures can be used as suggestive evidence that the identifying variation in our race-based relatability measure is not correlated with unobservable confounders.

We start by bringing demographic definitions in the ATUS in close concordance to the level

[32] We stem the words using Porter's stemming algorithm (Willet 2006) to collapse all instances of a word to a shared stem. For example, this algorithm would transform *cats*, *catlike*, and *catty* to the stem *cat*.

of variation that exists in the student data. For economic disadvantage, the item-level testing data contains an indicator for whether a student is on free lunch, on reduced-price lunch, or is on another social insurance programs provided by the state or the federal government. We collapse this measure into a binary variable of economic disadvantage indicating whether or not the student participates in any program. We do not observe directly in the ATUS data whether the respondent lives in a household which participates in any social insurance program or has a child on free or reduced-price lunch. Instead, we observe a respondent's household income range and household size. Since household participation in state or federal assistance programs—free and reduced-price lunch included—is often tied to being at or below federal poverty line thresholds, we use the income range, household size, and federal poverty line tables to determine a respondent's distance to the poverty line. Respondents whose household income is at or below 200% of the federal poverty line are classified as economically disadvantaged.[33] For urbanicity, the ATUS data is more limiting than the student data. We only observe whether respondents live in a metropolitan area or not, while for students we know exactly which school they are attending during the school year. e determine the county location of each Texas school and assign all students in a given school to the metropolitan status of its county, using U.S. Census determination of county metropolitan status. One pitfall of this approach is that we do not observe the physical address of students, which may differ from the county of school attendance. However, we regard this to be a relatively minor concern given that our sample consists of students attending public school, a setting in which attendance is overwhelmingly determined by proximity. Finally, both the ATUS and the student data report sex as male or female for all individuals. We accept these classifications with no modifications.

Next, we calculate topic exposure much in the same as we do before. Let $c \in \mathcal{C}$ be respondents in a demograhic group for a demographic grouping scheme and let t index

[33] Since income is reported as falling within a range, it is not possible to classify some respondents using this method. We drop such respondents from the data when calculating exposure by economic disadvantage due to the ambiguity.

topics as before. Then e_{ct} is simply the share of respondents in group c which reported any number of minutes participating in activities related to t. Once we obtain e_{ct} $\forall c, t$, we can calculate both E_c, the sum of topic exposure for c, and r_{cp}, the relatability of passage p to group c.

Finally, we estimate an analogue of equation (4),

$$Y_{cp} = \delta_{c,g(p)} + \pi_p E_c + \beta^{\mathcal{C}} r_{cp} + \nu_{cp},$$

where Y_{cp} is the mean performance of students of group c on passage p. The identification assumptions for this specification follow similarly as the assumptions for equation (4). As such, we calculate standard errors at the topic-passage-level as before and cluster them at the exam-level. We estimate this specification for different $\mathcal{C}$ and compare the resulting $\beta^{\mathcal{C}}$ with β from equation (4).

C Topic salience validation

We validate our topic salience metric with manual topic labeling by two student research assistants. The labeling task is designed to capture the relative salience across different topics within passage as well as the intensive margin variation in a topic's presence across passages. That is, the ordinal and cardinal properties of passage topics. To do this, each research assistant was instructed to read the entirety of a passage (ignoring the contents of the related question items) and perform two labeling activities. First, having the list of topics we consider for our analysis, the research assistants may select between 0 and 3 topics that appear in the passage and rank them ordinally in terms of relative salience.[34] Second, if they reported at least one topic as appearing in the passage, they would categorize the topic they ranked as number one to be either high, medium, or low salience in this passage.

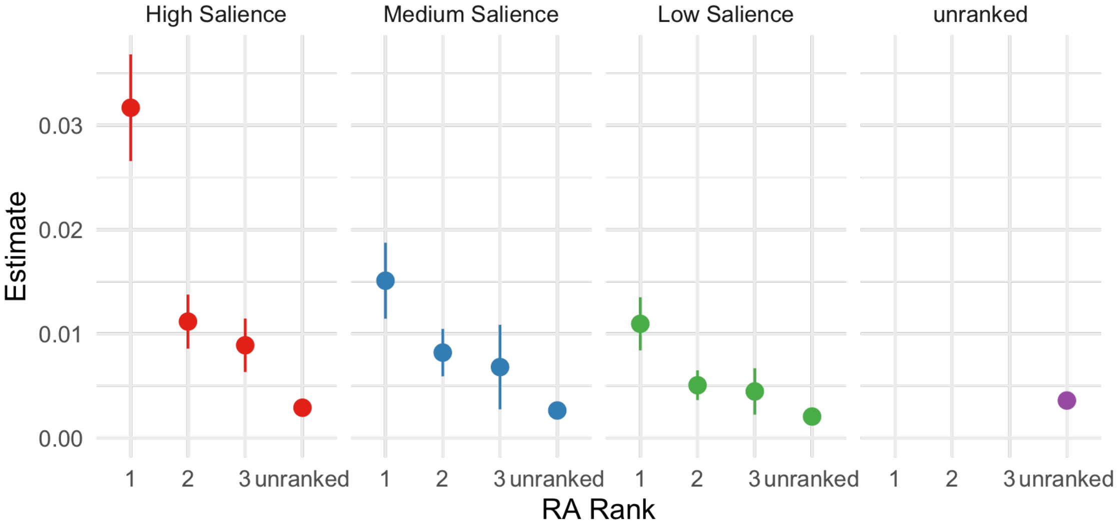
Now, consider how our topic salience metric m_{tp} should relate to the human labeling results. We would expect that when a research assistant ranks topics t, t', t'', respectively, as the 1st, 2nd and 3rd most salience topics in a passage, that $m_{tp} \geq m_{t'p} \geq m_{t''p}$. Further, if three passages p, p', p'' are labeled as a high, medium and low salience passage with respect to top-ranked topic t, we would expect that $m_{tp} \geq m_{tp'} \geq m_{tp''}$.

To test whether our topic salience metric aligns with these expectations, consider the regression

$$\begin{aligned} m_{tpi} = \sum_{k \in \{1,2,3,\text{unranked}\}} \Big(& \alpha_k \times \mathbb{1}[rank_{tpi} = k] \times \mathbb{1}[salience_{pi} = \text{high}] \\ & + \beta_k \times \mathbb{1}[rank_{tpi} = k] \times \mathbb{1}[salience_{pi} = \text{medium}] \\ & + \gamma_k \times \mathbb{1}[rank_{tpi} = k] \times \mathbb{1}[salience_{pi} = \text{small}] \Big) \\ & + \mu \times \mathbb{1}[salience_{pi} = \text{unranked}] + \varepsilon_{tpi} \end{aligned}$$

[34] The research assistants received no information about the methodology we used to classify the passages nor do they know about the dictionaries used in our NLP approach.

Figure C1: Regression of the topic salience scores m_{tp} on research assistant labeling



Notes: Standard errors clustered by passage are used to construct the 95% confidence intervals. Observations are at the topic-passage-RA level, while the outcome only varies at the topic-passage level. The estimation sample is all students grades 3 to 8 from 2013–2019 taking the standard STAAR reading comprehension test.

where $rank_{tpi}$ is the rank research assistant i assigned topic t in passage p and $salience_{pi}$ is the research assistant i's assigned salience level of the topic with the highest ranking in passage p. We first test whether $\theta_1 \geq \theta_2 \geq \theta_3 \geq \theta_{\text{unknown}}$ for each $\theta \in \{\alpha, \beta, \gamma\}$. To verify our NLP scores capture the intensive margin variation well, we next test $\alpha_1 \geq \beta_1 \geq \gamma_1$. The estimated coefficients displayed in Figure C1 exhibit the expected properties, suggesting that our dictionary-based methodology accurately reflects how a typical person may describe the relevant characteristics of a passage. We conduct this same analysis separately for each research assistant, and the results are qualitatively similar.

D Proper estimation of standard errors

Our estimation strategy, which closely follows Borusyak et al. (2022), relies on quasi-random variation in topic salience to identify the causal effect of relatability on student performance. We present our main specification as a regression in "standard" form, that is, at the race-passage level in which we observe our data. However, in actuality, given that our true, identifying variation occurs at the topic-passage level, we calculate standard errors in a regression at the topic-passage level. This is because since students receive common topic salience "shocks," we must account for the possibility that relatability and the residual may be correlated across racial groups. We detail below exactly how we obtain these "exposure"-robust standard errors using our main specification.[35] While we illustrate this approach for our main specification, the approach straightforwardly applies to all additional regressions relying on our core identification assumptions laid out in section 4.2.

We first define N as the number of student-item dyads which make up our underlying estimation sample and N_{dp} as the number of student-item dyads which correspond to race d and passage p. This allows us to formally define our regression weights $w_{dp} \equiv \frac{N_{dp}}{N}$.

Next, we separately residualize Y_{dp} and r_{dp} on the same controls variables in equation (4) through a regression with w_{dp} weights, obtaining residuals of $Y_{dp}^{\perp}$ and $r_{dp}^{\perp}$, respectively. In order to convert the outcome and independent variables from the race-passage level, for each topic t in passage p, we calculate a weighted average of each variable weighted by number of underlying observations w_{dp} and exposure e_{dt}:

$$\bar{v}_{tp}^{\perp} = \frac{\sum_d w_{dp} e_{dt} v_{dp}^{\perp}}{\sum_d w_{dp} e_{dt}} \tag{8}$$

with $v \in \{Y, r\}$.

[35] As discussed, the approach for calculating "exposure"-robust standard errors was formalized in Borusyak et al. (2022) and adapated for our specific setting.

Finally, we estimate an IV model with second-stage equation

$$\bar{Y}_{tp}^{\perp} = \varepsilon + \beta \bar{r}_{tp}^{\perp} + q_{tp}'\psi + \zeta_{tp} \tag{9}$$

where in the first-stage equation we instrument $\bar{r}_{tp}^{\perp}$ using m_{tp}, $q_{tp}'\psi$ includes topic-grade fixed effects and passage fixed effects, and the equations are weighted by $\sum_d w_{dp}e_{dt}$. Standard errors are clustered at the exam-level, which reflects potential positive and negative correlation between topic salience within passage and across passsages for the same exam. Borusyak et al. (2022) shows the coefficient on $\bar{r}_{tp}^{\perp}$ using this specification is equivalent to the coefficient obtained from estimating the "standard" form regression.

To make sense of the form of this final specification, we intuitively explain each step of this calculation process. First, we need to purge all non-identifying variation from both student performance and relatability. Then, we effectively "unpack" our data by recognizing that (t, p) underlies (d, p). After disaggregating the data to the (d, t, p)-level, we collapse the variables across race, but emphasizing observations with more underlying student and item data (which contribute more to our "standard" regression estimates) and racial groups with higher exposure to the topic. The two-stage IV strategy isolates just the variation in relatability that is driven by topic salience. Further, the topic-grade fixed effects and passage fixed effects are exactly analogous to the race-grade fixed effects and passage-by-exposure sum fixed effects in the "standard" regression; when disaggregating from the race-passage level and aggregating to the topic-passage level, race fixed effects become topic fixed effects and passage-by-exposure sum effects collapse to passage fixed effects.

E Robustness

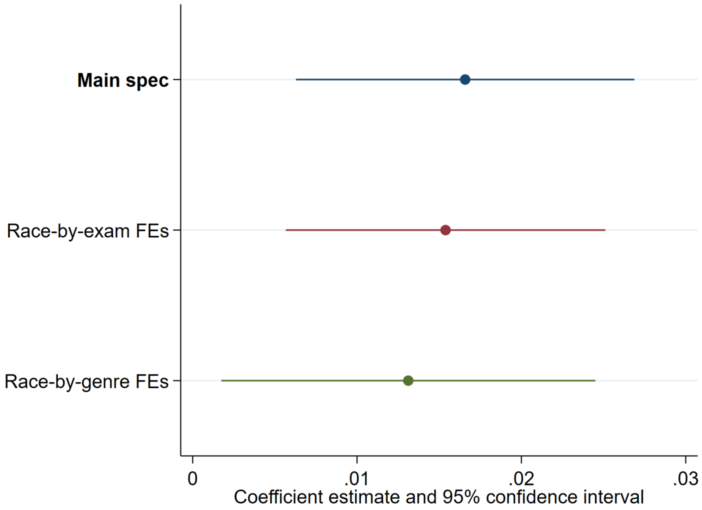
Our estimates are downstream of modeling and data choices that are baked into our race-based relatability measure and model specification. First, we show that the estimates are qualitatively the same with alternative fixed effect specifications. Second, we show that our results are not sensitive to specific choices pertaining to the relatability measure, considering topic categories to include from the ATUS, the natural language processing algorithms or metrics, and the ATUS respondent sample construction.

Given our empirical strategy, intuitively, the model's fixed effects are intended to condition sufficiently such that we have plausibly exogenous topic salience measures drawn for each passage. Accordingly, our baseline specification uses race-by-grade fixed effects. We do consider more and less granular specifications (see Figure E1). Using race-by-exam fixed effects (i.e. race-by-grade-by-year) gives similar results to baseline with a slightly attenuated coefficient yet tighter standard errors. However, using race-by-genre fixed effects, which is conditioning on the type of passage such as "fiction poetry" or "literary non-fiction", results in a qualitatively similar coefficient while making the comparison group less restrictive.

Examining the relatability measure, we first consider the choice of which topic categories we include. The baseline estimate has 25 topic categories, and in Figure E2 we demonstrate that the coefficient is relatively stable to the removal of any individual category. Quantitatively, the handful of outliers that swing up or down the most (relative to the baseline) are only about a 30% increase or decrease in the point estimate.

Next, we consider using alternative measures of topic salience. We deviate from the baseline metric of term frequency and consider the term frequency-inverse document frequency (tf-idf) measure (see Appendix Section B.2 for details). Further, we consider a variety of changes to the NLP data processing steps such as leaving the words unstemmed, using only the nouns, and discretizing the shocks. We see in Figure E3 that the results are qualitatively similar, with all significant except for the discritized shocks which is relying only on the extensive margin variation and ignoring the intensive margin variation that the NLP methods

Figure E1: Robustness of baseline effect to different levels of saturation in the specifications of fixed effects

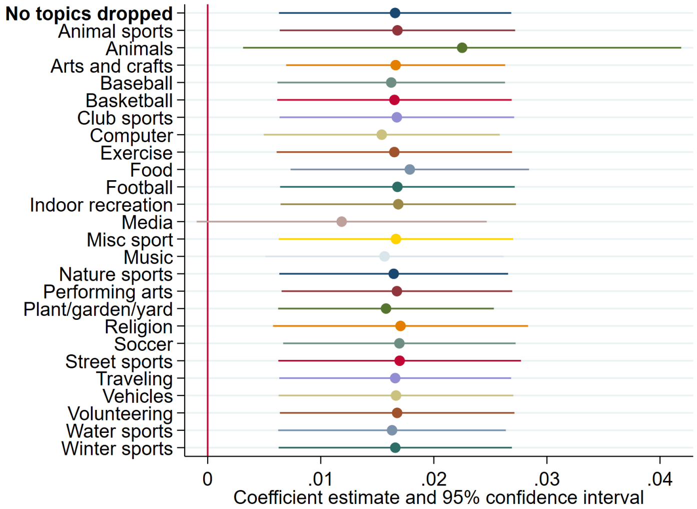


Notes: Each specification is a regression of the share of items answered correctly on a passage on content relatability. Unreported controls include: (1) "unit" fixed effects at the level indicated in the legend and (2) exposure-sum-by-passage fixed effects. Standard errors clustered by exam are used to construct the 95% confidence intervals. Observations are at the race-passage level. The estimation sample is all students grades 3 to 8 from 2013–2019 taking the standard STAAR reading comprehension test.

allow us to leverage. We also demonstrate that our results are not reliant on specific words in any of the topic dictionaries by generating 1000 permutations of the dictionary set, leaving out one word at random for each topic in each permutation. These estimates are compared to our baseline estimate in Figure E4.

Lastly, we consider the importance of the ATUS respondent sample we use to construct the race-based exposure scores. In our main specification, we use the entire ATUS data to construct exposure, but here we consider filtering down to subgroups that may be more representative (but have smaller sample size), discussed in section B.1. We see in Figure E5 that using the more tailored age group and respondents with children provides a more predictive point estimate. Further, restricting to Southerners seems to slightly attenuate the estimate and only using weekend responses has no difference. Nonetheless, these differences are minor and collectively point to affirming the baseline estimates that just use the ATUS

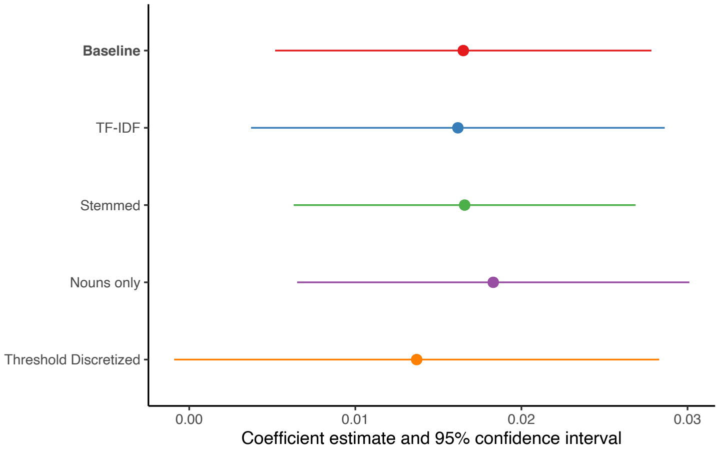
Figure E2: Robustness to topic set



Notes: Each row after the first is a separate regression of the fraction of questions correct for a passage on relatability, after leaving the indicated topic category out. Unreported controls include: (1) race-grade fixed effects and (2) exposure-sum-by-passage fixed effects. Standard errors clustered by exam are used to construct the 95% confidence intervals. Observations are at the race-passage level. The estimation sample is all students grades 3 to 8 from 2013–2019 taking the standard STAAR reading comprehension test.

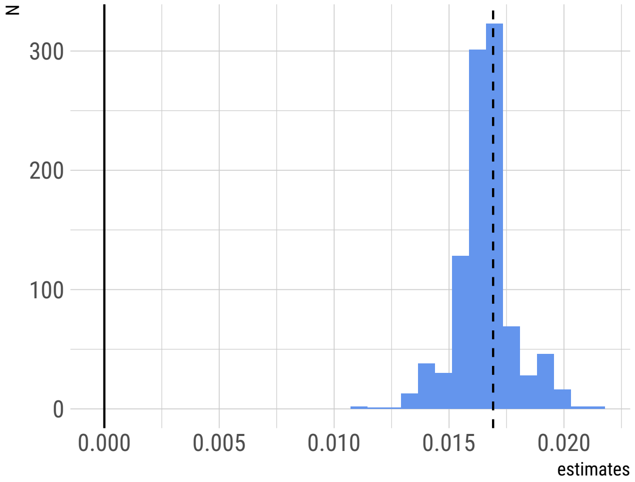
data as-is.

Figure E3: Robustness to different NLP measures for topic salience



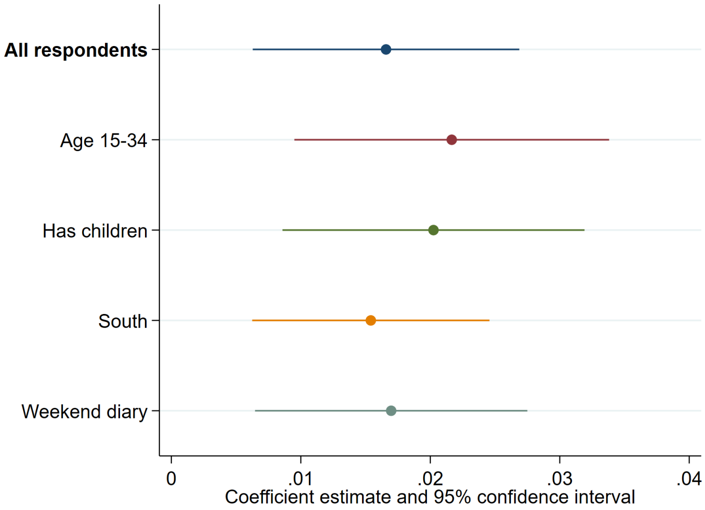
Notes: Each row is a separate regression of the fraction of questions correct for a passage on relatability. Unreported controls include: (1) race-grade fixed effects and (2) exposure-sum-by-passage fixed effects. Standard errors clustered by exam are used to construct the 95% confidence intervals. Observations are at the race-passage level. The estimation sample is all students grades 3 to 8 from 2013–2019 taking the standard STAAR reading comprehension test. The "TF-IDF" specification uses inverse document frequency weights to calculate the topic salience. The "stemmed" specification uses stemmed words when matching between the dictionaries and passages. The "Nouns only" specification filters down to nouns when matching between the dictionaries and passages. Lastly, the "Threshold Discretized" specification sets the top 1/25th topic salience values to 1 and the rest to zero, within grade-level. See details in Appendix Section B.2.

Figure E4: Histogram of estimates after removal of one word from each topic's dictionary



Notes: Each of the 1000 observations is a separate regression of the fraction of questions correct for a passage on relatability. Each regression uses a different estimate of relatability obtained after removing a single word at random from each topic's dictionary when constructing topic salience measures. Unreported controls include: (1) race-grade fixed effects and (2) exposure-sum-by-passage fixed effects. The level of observation for each regression is at the race-passage level. The estimation sample is all students grades 3 to 8 from 2013–2019 taking the standard STAAR reading comprehension test.

Figure E5: Robustness to different ATUS samples for calculating demographic exposure



Notes: Each row is a separate regression of the faction of questions correct for a passage on relatability. Each row calculates relatability using a different subsample of the ATUS sample. Point estimates and 95% confidence intervals are reported. Unreported controls include: (1) race-grade fixed effects and (2) exposure-sum-by-passage fixed effects. Standard errors clustered by exam are used to construct the 95% confidence intervals. Observations are at the race-passage level. The estimation sample is all students grades 3 to 8 from 2013–2019 taking the standard STAAR reading comprehension test. "All respondents" corresponds to no ATUS sample restriction. "Age 15-34" corresponds to respondent age. "Has children" corresponds to respondents which have a child in the household. "South" corresponds to the Southern U.S. Census Regions."Weekend diary" corresponds to using only the ATUS sample by which respondents' diary day is on the weekend.

F STAAR standards harmonization

Every year, the TEA sets STAAR performance standards which are meant to link STAAR results to state-mandated curriculum standards. Raw test scores are converted to scale scores based on the overall difficulty of the test, then performance standards categories are based on the scaled scores. Over the course of STAAR testing, the TEA has continually changed the overall framework for determining performance standards. From the 2012–13 to 2015–16 academic year, students were in one of three categories: "Unsatisfactory," "Satisfactory," and "Advanced." The TEA originally intended to gradually raise the threshold for "Satisfactory" to a long-run, pre-announced level ("Satisfactory: recommended"), but only did so for the 2015-16 academcic year. Starting in the 2016–17 academic year, the TEA instead switched to a four-tier system: "Did Not Meet," "Approaches," "Meets," and "Masters." Students who would have been classified in the lowest and highest categories would continue to do so across the three-tier and four-tier system. However, the "Satisfactory" category was split into two categories, for students who were below the "Satisfactory: recommended" threshold and students who were above it.

In order to make consistent predictions across years, we create four categories across all years. For the 2016–17 to 2018–19 testing years, we maintain the TEA-designated categories. For the 2012-13 to 2015-16 years, we split the "Satisfactory" category into two groups based on the "Satisfactory: recommended" threshold. While this adjustment may not exactly reflect how students were actually classified in these years, we argue that this approach is reasonable. First, this procedure reflects precisely how the TEA modified the three-tier system to the four-tier system and allows for easier comparisons across time. Second, the TEA had already announced the threshold for "Satisfactory: recommended" before the 2012–13 school year, meaning that it could have potentially been used as an unofficial benchmark by parents, teachers, and schools.