Content relatability and racial standardized test gaps: Evidence from Texas*

Steven Lee Matthew Schaelling†

[PRELIMINARY: NOT FOR PUBLICATION OR DISTRIBUTION]

September 24, 2023

Abstract

One goal of standardized tests is to measure aptitude across heterogeneous students with minimal bias. Could differential relatability of students to test content be contributing to bias and widening 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 build race/ethnicity-specific measures of content relatability to passages 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 causes an 1.7pp increase in student performance. White students have higher average relatability than non-white students, of which 32% can be attributed to the selection of topics in passages. Our results suggest that equalizing the cultural content of questions could reduce both the Black-white and Hispanic-white test score gaps in Texas by 4%.

1 Introduction

Inequality along racial and economic dimensions is well-documented and widespread — it also begins at a very young age. Among the first instances of these gaps are observed among

*We appreciate valuable comments from Jesse Bruhn, John Friedman, Amy Handlan, Peter Hull, Stelios Michalopoulos, Matt Turner, and Neil Thakral, as well as feedback from participants at various seminars at Brown University and the Annenberg Institute. 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.

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

children in public school, especially notable in standardized testing (Fryer & Levitt, 2004; 2013). For example, in the state of Texas there is a 14 percentage point gap 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 the gap. 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.[1]

If these tests fail to accurately measure student achievement of students of all backgrounds, understanding the exact mechanisms by which it happens is crucial. A possible pathway is the interaction between cultural relatability and educational content. Most existing research has focused on the representation of different identity groups in educational materials, such as textbooks and assigned literature (Adukia et al., 2021), but has not examined the importance of relatability directly impacting educational outcomes. Furthermore, in reading comprehension passages, there is frequently not enough information to evaluate the identity of the author or main character. However, empirically analyzing the relationship between relatability and education performance is challenging. First, student background or culture is inherently qualitative and not typically well-captured in existing datasets. Second, even with a quantitative representation of background, there is the technical challenge of mapping and measuring that within educational materials.

This paper evaluates the impact of the content relatability of reading passages on standardized test performance. We determine the "relatability" of 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

[1]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).

end-of-year standardized tests for primary and secondary school students in Texas from 2013-2019, and we test whether content relatability causally improves student performance, leveraging conditional randomness of topics in passages. The causal effect of content relatability has several implications for the construction of standardized tests. We explore in particular whether the realized distribution of reading passages leads to the empirical racial- and ethnic-based test gaps wider than their true differences in aptitude.

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 topics in a passage leads to a 1.7pp increase in the probability of a correct answer. Attempting to falsify the model, we consider a variety of placebo outcomes with the same model specification, including prior student performance and demographic composition, and find no effect. This impact is robust to alternative estimation strategies, such as changing the methodology for constructing the relatability measure. We find that white students have higher average relatability than non-white students, which is driven partially by the selection of more relatable topics in passages. Our results imply that equalizing content relatability across groups would lead to a 4% smaller Black-white test gap and 4% smaller Hispanic-white test gap. Merely reducing the tendency to select white-relatable topics could close the gap by 1% for black and Hispanic students.

We contribute to existing work studying cultural 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 overrepresented in textbooks using text analysis (Adukia et al., 2021; 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). 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). 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. In a closely related study, 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. 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 in Duquennois (2022) relies on random assignment of digital homework questions or exam booklets.

Our paper relates to two additional strands of literature. First, test score gaps, especially across race, has been a much-discussed topic among researchers and policymakers alike (see Fryer & Levitt, 2004; 2013; Freedle, 2010; Harvard Educational Review, 2010). 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).

We organize the rest of the paper as follows. Section 2 describes the student test data and administrative survey data. Section 3 explains how we construct our measure of sociocultural relatability. Section 4 provides an overview of our estimation framework. Section 5 describes our main results. Section 6 extends our results. Finally, section 7 concludes.

2 Data

2.1 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

focus on the State of Texas Assessments of Academic Readiness (STAAR) which are a set of end-of-year assessments taken by public school students in grades 3 through 8 and end-of-course assessments taken by students in high school. While tests are different across grade/course, most students receive the same test within a year. 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.

Within the set of STAAR assessments, we limit our attention to 2013-2019 reading comprehension exams from grades 3 to 8.[2] The reading exam format is relatively standard across grades; students are expected to read a handful of text passages and to 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.[3]

We use the item-level student responses and the corresponding answer key 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.[4] Ultimately, we obtain 205 unique passages from 42 different exams.[5]

We supplement the testing data with additional student demographic information including the race of a student, whether they are Hispanic, and their economic disadvantage status. 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. We also construct a

[2]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.

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

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

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

Table 1: Summary statistics for student sample

	Mean	SD
Demographics		
Hispanic	0.46	
Non-Hispanic		
Asian	0.05	
Black	0.15	
White	0.34	
Economically disadv.	0.56	
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).

binary measure of economic disadvantage, which is 1 if a student is on free- or reduced-price lunch or is otherwise indicated to have economic need and 0 otherwise.

Our sample consists of 13,180,138 student-exam observations (henceforth, just called "students").[6] 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.[7] Among non-Hispanic students, the majority are white (34% of total) compared to Black (15%) or Asian (5%). Fifty-six percent of students are considered economically disadvantaged. In our sample, we have an average of 4.8 passages per exam, corresponding to just under 42 test items. Students answer roughly two-thirds of these questions correctly.

[6]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.

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

2.2 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 designed to provide 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 on their activities in the prior day to a phone interviewer. The interviewer then takes this information and formally assign each reported activity to one of 424 six-digit classification codes. Each activity code encompasses 3-tiers of detail regarding the activity, with 17 major categories making up the highest tier.[8] We exclude certain activity codes from our analysis and collapse all others into 26 "topics." More information on topics can be found in Section 3 and 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. For each interviewed individual, we observe the number of minutes spent on every ATUS activity, as well as their race, household size, household income range, and other demographic characteristics. We align the racial/ethnic categorization in this data as closely as possible to that of the student data, by excluding multi-race individuals and categorizing Hispanic respondents into their own category.

Table 2 summarizes ATUS data we use. On average, a respondent does 19 activities per day, spending 138 minutes per activity. Further, an average respondent does activities within 2 to 3 of our defined topics. This accounts on average to 14% 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. Further,

[8]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."

Table 2: Summary statistics for ATUS sample

	Mean	SD	5%	95%
Activities				
Number of activities	18.76	7.30	8	32
Number of min spent per done activity	137.74	64.19	68	260
Number of topics	2.50	1.32	1	5
Share of time spent in topics	0.14	0.10	0	0
Demographics				
Asian	0.04	0.20	0	0
Black	0.14	0.35	0	1
White	0.66	0.47	0	1
Hispanic	0.15	0.35	0	1
<100% of poverty line	0.11	0.32	0	1
<185% of poverty line	0.28	0.45	0	1
<200% of poverty line	0.30	0.46	0	1
Observations	83061			

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.

we find that the ATUS sample is also relatively higher in SES than the student sample; only 28% of respondents are in households below 185% of the poverty line, significantly lower than the economically disadvantaged share of students in our Texas sample.

3 Measuring relatability

3.1 Motivation

Our approach to measuring relatability of a reading passage includes predicting familiarity with a topic given an individual's race based on survey response data in the ATUS. This is weighted by a measure of topic prevalence in the reading passage and summed across all topics. Our goal is to determine $r_{d,p}$, a measure of relatability, focusing solely on the degree to which a passage is relatable to a student of demographic group d, and thereby estimate some parameter which gives a measure of how predictive question relatability is on performance.

3.2 Demographic topic exposure

We construct a measure of the extent to which individuals in each demographic group are exposed to a topic. First, we must generate a set of a topics. Using the ATUS, we filter the 424 activities down to 146 activities that could reasonably be linked to an activity category that would impact a child's awareness of the world (excluding banal activities such as sleeping, commuting to work, or storing household items). These activities are then grouped into 25 "topic" categories such as arts & crafts, animals, soccer, and winter sports (see Appendix Table A1).

Using differences in reported time use across race, we can create a measure of exposure to topics for a given race. We can imagine this procedure as a proxy for a more ideal measure—directly surveying an individual regarding their familiarity with any topic that could appear in a reading passage. Given that the majority of ATUS respondents are not school-aged, we also assume that adult exposure is similar to child exposure since students are exposed to their parent's activities. Formally, we generate $e_{d,t}$ as the fraction of respondents of demographic group $d \in \mathcal{D}$ who report an activity related to topic $t \in \mathcal{T}$:

$$e_{d,t} \equiv \frac{1}{N_d} \sum_{j:d(j)=d} \mathbb{1}[\tau_{j,t} > 0]$$

where j indexes respondents, N_d is the total number of respondents of group d, and $\tau_{j,t}$ is reported minutes of activity related to t for j.

We construct additional alternative measures of topic exposure through subsampling respondents by age, Census region, day of week, and others. Further details on subsampling is discussed in Appendix Section B.

We observe differential exposure to activities across sociocultural groups in the ATUS data. In Table A2, we present the exposure of different racial groups to topics in relation to that of white respondents. We observe wide sociocultural divergence in direct exposure to activities. For example, Black individuals are significantly less likely than white individ-

uals 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.

3.3 Topic salience in reading passages

We determine the content of the reading passages in the standardized tests. For each passage-topic pair, we use natural language processing (NLP) methods to systematically create an index that represents how strongly related a specific passage is to a given topic.

We follow a very intuitive algorithm, a type of dictionary method, which operates as follows. First, determine the set of words that we consider to be indicative of a topic: that is, a set B_t for each topic $t \in \mathcal{T}$. Second, for each passage $p \in \mathcal{P}$, we calculate the fraction of words in a passage that are in a given topic dictionary, giving a straightforward measure of 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 $B_t = B_{t,1} \cup B_{t,2}$. These dictionaries can be seen in Tables A3-A5. 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). We then "stem" the words using Porter's stemming algorithm (Willet 2006) to collapse all instances of a word to a shared stem.[9]

To measure topic salience, denoted $m_{t,p}$, we count the number of times some word w is in p and denote it $count(w, p)$; that is, $count(w, p) = \sum_{v \in W_p} \mathbb{1}[v = w]$, where W_p is the set of words in passage p (including repeated words). Then, we can calculate the *term-frequency*

[9]For example, this algorithm would transform *cats*, *catlike*, and *catty* to the stem *cat*.

as $tf(w,p) = \frac{count(w,p)}{|W_p|}$. As our main measure of topic salience, we use term-frequency and define:

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

We also consider other measures, discussed further in Appendix Section B.2.

Considering a binary threshold of $m_{t,p} \geq 0.01$, we find that the average passage contains 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. We present the frequency which a topic appears in the top 3 matches for any passage using our dictionary-based method in Figure 1. 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).

3.4 Relatability measure

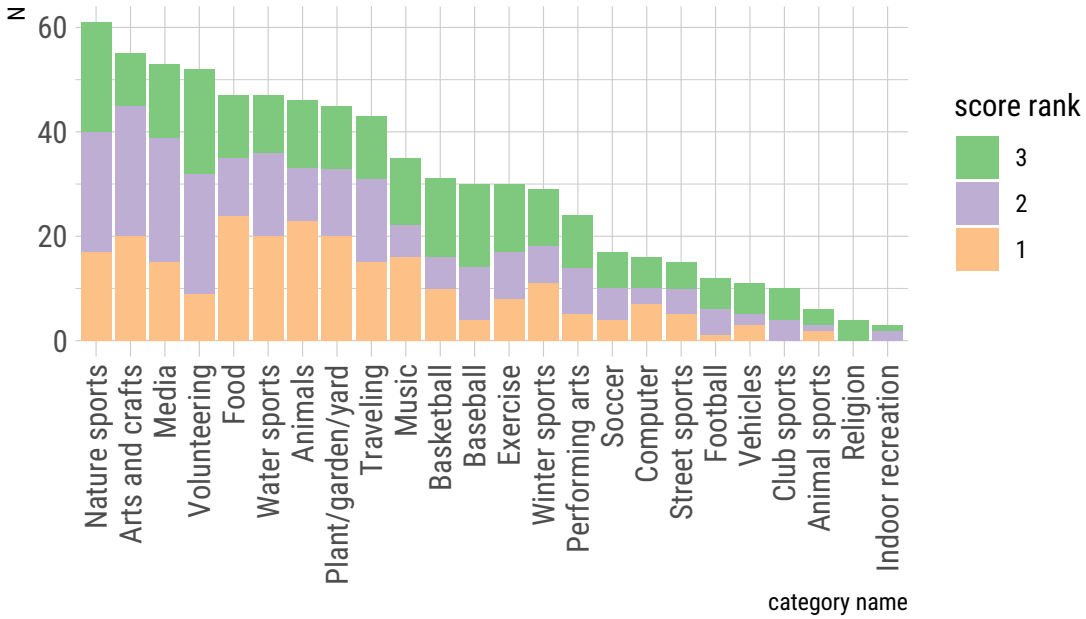
Our final measure of relatability, $r_{d,p}$ is a straightforward combination of the demographic exposure measure and the NLP measures from the previous subsections. Our preferred measure of relatability is defined as

$$r_{d,p} \equiv \sum_{t \in \mathcal{T}} e_{d,t} m_{t,p}. \tag{1}$$

Intuitively, $r_{d,p}$ is simply a weighted average of topic salience for passage p weighted by the topic exposure for demographic group d. For ease of interpretation, we also standardize $r_{d,p}$ to ensure they are mean zero with unit variance.

To show preliminary evidence that our relatability measure is predictive of the racial test gap, we plot the passage-level differential relatability and differential exam performance for Black vs. white students and Hispanic vs. white students in Figure 2. We observe

Figure 1: 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.

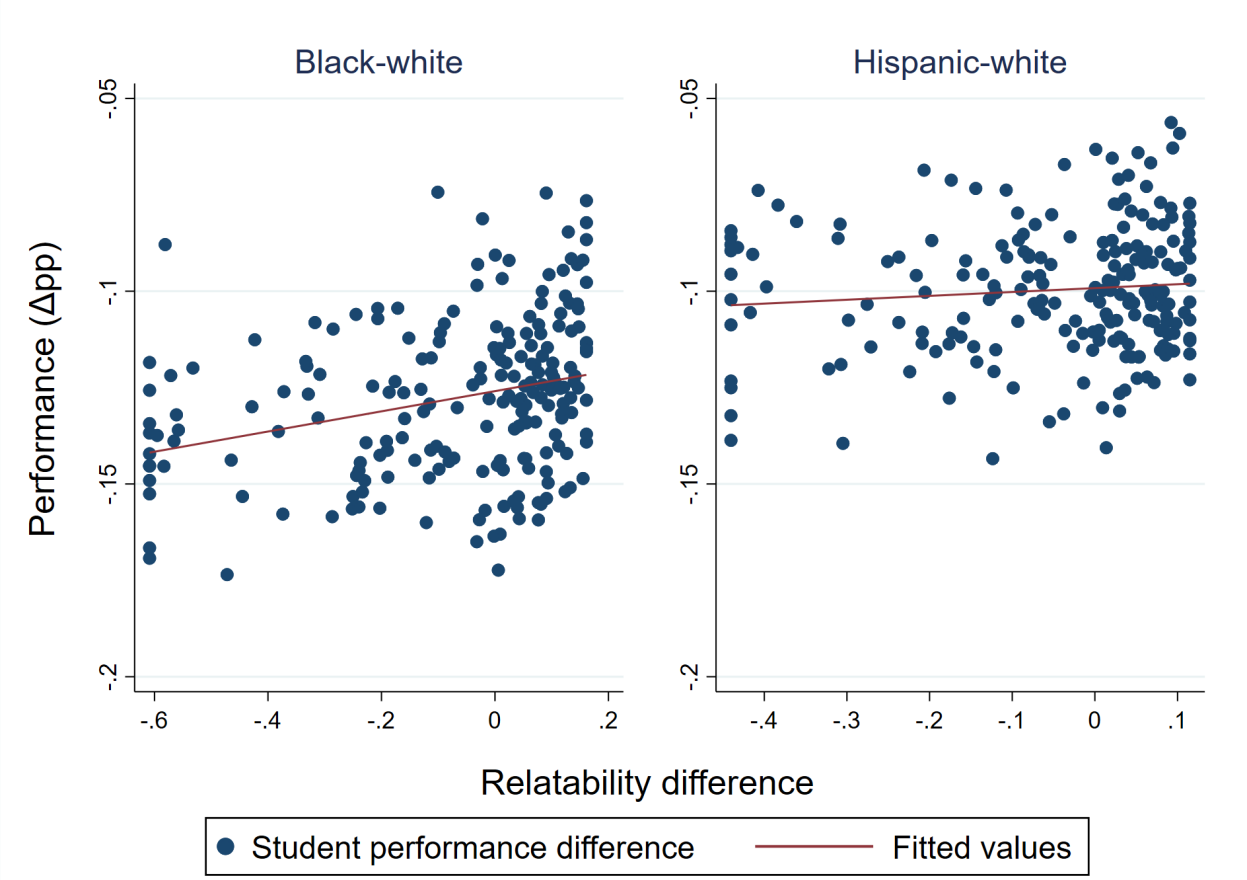
a clear positive relationship between Black-white relatability differences and Black-white performance differences, which is suggestive of an imapct of relatability on racial test gaps. However, we observe a less clear relationship between relatability and Hispanic-white test score gaps.

4 Estimation framework

4.1 Causal effect of relatability

Our estimation strategy relies on conditional randomness of topic salience to identify the causal effect of content relatability on student performance. To illustrate this, we start with

Figure 2: Non-white vs. white students performance gaps and relatability



Notes: Each graph plots average student performance differences between non-white and white students against relatability differences and the simple linear fit between these measures. 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.

the following simple potential outcomes model of exam performance on a passage at the student-level

$$Y_i(p) = f(\theta_i, \phi_p, \rho_{ip}, \eta_{ip}), \tag{2}$$

where i indexes students, p represents a passage, and $Y_i(p)$ is the share of questions associated with passage p answered correctly by student i. We assume there are four components

influencing student performance: an ability component θ_i, a passage characteristics component ϕ_p, a passage content relatedness component ρ_{ip}, and an idiosyncratic term η_{ip}. ρ_{ip} is designed to capture the degree of match between topics with which student i familiar and topics which appear in passage p. We make a simplifying assumption that the components in $f(\cdot)$ are linearly separable. Further, we assume that ρ_{ip} is known and determined according to equation (1) such that $\rho_{ip} = r_{d(i)p}$. The resulting model is

$$Y_i(p) = \theta_i + \phi_p + \beta r_{d(i)p} + \eta_{ip}. \tag{3}$$

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 and differential exposure to topics across racial groups. We face two main challenges with this approach. First, θ_i and $r_{d(i)p}$ may be correlated. For example, across grades, certain topics may be much more likely to appear than others. Given the possibility that θ_i is associated with race and grade level, if there is differential exposure to more prevalent topics then $r_{d(i)p}$ will be correlated with θ_i. Further, exposure $e_{d(i)t}$ can be directly correlated with θ_i. Second, ϕ_p may be correlated with $r_{d(i)p}$. 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 differential presence of topics conditional on a given topic-grade and a given passage. We assume that this variation is as-good-as-random with respect to other components in the model. Then, given the structure of $r_{d(i)p}$ as a linear interaction between "endogenous" shares ($e_{d(i)t}$) 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 expected relatability conditional on topic-grade and on passage.[10]

[10]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

In our setting, conditional expected relatability is spanned by controlling for race-grade fixed effects and passage fixed effects multiplied by $E_d \equiv \sum_t e_{dt}$. To see the former, consider first that topic exposure e_{dt} is passage-invariant. Thus, expected relatability in 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 passage 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 illustrate that these controls 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, group-grade fixed effects 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 passage fixed effects multiplied by exposure sum. Crucially, if we did not interact passage fixed effects with exposure sum, ϕ still predicts relatability through differential exposure sums between group A and group B.

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.

Thus, we estimate the following regression specification:

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

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 shares, E_d.[11] Since relatability varies by (d, p), we have aggregated Y_{ip} to the race-passage level and estimate the equation with weights representing the number of students and test items which make up (d, p).

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.[12] This approach is motivated by two factors: (a) topic salience is both negatively and positively correlated within a passage (e.g., basketball and baseball may be present together) and (b) topic salience is likely negatively correlated within an exam (e.g., testmakers 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 C.

4.2 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 racial test gaps.

[11] It is possible to re-estimate 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. We argue that the identification assumptions necessary for TWFE are not met in our setting; nevertheless, we obtain similar results under this specification.

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

We calculate the change in test score gaps if relatability were equalized for all students by predicting $\widehat{Y}_{dp}$ from equation (4) and predicting $\tilde{Y}_{dp}$ setting $r_{dp} = 0$. These predictions represent outcomes at average relatability and equalized relatability, respectively. We then generate conditional means by race:

$$\widehat{\mu}_d = \frac{1}{|I_d|} \sum_{i \in I_d} \widehat{Y}_{iq}, \qquad \tilde{\mu}_d = \frac{1}{|I_d|} \sum_{i \in I_d} \tilde{Y}_{iq}$$

where $I_d = \{i : d(i) = d\}$. 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.

Even if we detect that relatability contributes to test score gaps, it is important to distinguish between how much of this contribution is driven by passage selection rather than exposure carried into the test. From the perspective of the testmaker, if $E_d > E_{d'}$ such that members of group d generally have higher exposure than members of group d', the only way to equalize relatability would be to purge all relatable topics from passages or selectively choose topics favoring group d' to balance out differences in exposure. The former may be impossible and the latter may be inadvisable.

To approach this problem, we can decompose cross-group differences in average relatability as follows:

$$\bar{r}_d - \bar{r}_{d'} = (E_d - E_{d'})\bar{m} + \frac{\sum_p \sum_t (e_{dt} - e_{d't})\tilde{m}_{tp}}{|\mathcal{P}|} \tag{6}$$

where $\bar{m}$ is average topic salience across all topics and passages, $\tilde{m}_{tp}$ is the residual of m_{tp} from $\bar{m}$, and $|\mathcal{P}|$ is the number of passages. The first term of the right-hand side of equa-

tion (6) represents differences in relatability due to differences in overall levels of exposure. Testmakers still have control over this component through $\bar{m}$, but may find adjusting $\bar{m}$ challenging. The second term of the question represents differences in relatability due to selection of differentially favorable topics in passages. Note that since $\bar{m}_{tp}$ is a residual, it is possible for this term to be negative even if $e_{dt} - e_{d't} > 0, \ \forall t$. Intuitively, this term accounts for the association between overselected or underselected topics and differences in topic exposure. Finally, the share contribution of these terms to differences in relatability can be applied to the results in expression (5).

5 Results

5.1 Impact of relatability on performance

Table 3 shows theresults 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). Both specifications include passage-by-exposure sum interactions. We find in both specifications that relatability has a positive impact on test performance. According to our preferred specification, a standard deviation increase in relatability for a passage leads to a 1.7 percentage point increase in the share of items answered correctly for that passage. This effect is statistically significant at the .01 significance level.

As described in section 4.1, our specification relies on an orthogonality assumption between relatability and the unobserved error term after including race- and passage-based controls. This assumption would be violated if testmakers were selecting topics based on prior-year student performance or based on the composition of test takers for the upcoming exam.

We test these possibilities directly by replacing Y_{rp} with new "outcomes." Let $h(p)$ index passage position, which corresponds to the position which passage p holds in an exam, and

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
No. of students	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.

let $m(p)$ index the calendar year in which passage p appeared on an exam. Then for each race-passage dyad rp, we define $Y_{rp}^{prev_grade}$ and $Y_{rp}^{prev_cohort}$ which are student performance for race r for the same passage position $h(p)$ in last year's exam $m(p)-1$. The key difference is that in the former, grade $g' = g(p)$, while in the latter, grade $g' = g(p) - 1$. The final new outcome we define is $RaceShare_{rp}$ which is the share of test takers for an exam $g(p)m(p)$ that are of race d. These outcomes test for three potentially plausible ways in which testmakers select topics that would bias our estimates. Testmakers may observe racial heterogeneity in performance for a particular grade in one year and may seek to correct (or enhance) this for the same grade in the following year. Alternatively, they may seek to correct these results by adjusting topics for the same cohort of students. Testmakers may also be able to adjust topics for a given year's test after observing the enrolled students for that school year.[13]

Table A6 shows the results of estimating our preferred specification with these three falsification outcome measures. In all three columns, the coefficient estimates on relatability are close to 0, and we are unable to reject the null hypothesis of no effect. Since columns (1)

[13]Since our exams are held at the end of the school year and almost all students are required to take them, it is theoretically possible that six or seven months prior to test administration testmakers know the demographic breakdown of test takers.

and (2) show specifications that effectively use the lagged main outcome as the dependent variable, these two results can also be thought of as approximately equivalent to passing a "pre-trend" test that evaluates differential trends in the outcome prior to treatment.[14]

Our results are robust to alternative specifications and formulations of our relatability measure. Specifically, we re-run our main specification by employing alternative fixed effects, using various subsamples of the ATUS data to calculate e_{dt}, using different NLP methods and measures to calculate m_{tp}, and dropping topics from our calculation of r_{dp}. Details on the methodology and results of the robustness analysis is in Appendix Section D.

5.2 Estimating racial test gaps

We observe large test gaps between non-white and white students on reading comprehension tests. Table A8 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 gap which shrinks to 11.9 percentage points by grade 8. The gap for Hispanic students is slightly smaller: between 9 and 11 percentage points across grades.[15] 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 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

[14] Columns (1) and (2) also involve further restrictions to our estimation sample beyond those laid out in section 2.1. In table A7 we re-run our main specification with the restricted samples featured in columns (1) and (2) of table A6. We find that our main results remain largely unchanged in these smaller samples.

[15] 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.

students, racial test gaps are potentially smaller than can be detected using raw student performance. Relatability accounts for 4% of the test score gap between Black and white students and Hispanic and white students. Relatability is a larger contributor to test score gaps in early years, where in 3rd grade the Black-white and Hispanic-white test gaps are 5% and 6%, respectively, and in 8th grade both gaps are 2%.

As detailed in section 4.2, 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.

Using the decomposition represented in equation (6), we find that differences in exposure levels contributes to just over two-thirds of Black-white (68%) and Hispanic-white (69%) differences in average relatability. Applying this decomposition to the relatability contribution calculations, we find that more than 1% of nonwhite-white test gaps can be closed just by randomizing test passage performance.

6 Extensions

6.1 Reconsidering student test standards

Our analyses thus far show that relatability is predictive of performance on reading comprehension tests which may lead to incorrect calculations of the Black-white test gap. We now consider whether differences in passage relatability across tests can impact students directly. In particular, we calculate whether adjusting for passage relatability could change the number of Black students meeting STAAR performance standards.

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.

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 approach. 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 estimate

$$score'_i = score_i + N_{g(i)m(i)}\hat{\beta}\left(\tilde{r}_{g(i),d(i)} - \bar{r}_i\right), \tag{7}$$

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 $\bar{r}_i$ 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 their respective performance standards category and calculate the change from what their adjusted performance standards were originally.

Tables 4 and 5 summarizes the results of applying this exercise to a 1% sample of Black students and Hispanic students, respectively.[16] 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

[16]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 constraints.

Table 4: 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 5: 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.

Black and Hispanic students being classified as not meeting any 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 we find effects are generally smaller for higher grades, relatability still has large impacts for some exams in higher grades; for one 7th grade exam, 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.

The results suggest that even with small adjustments in average scores at the aggregate, the effect of relatability is strong enough to reclassify students as meeting or not meeting performance standards. To the extent that the STAAR tests are used to assign future classes for students or prevent grade promotion, this shows that subtle changes in test content could

meaningfully change the learning trajectory of students.

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. According to our estimates, differences in relatability accounts for 4% of the Black-white and Hispanic-white test gaps.

Our results have implications both for test writers and education policymakers, more generally. First, it highlights that in order to write balanced assessments, test makers should take into account 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 considering achievement gaps along demographic dimensions, one additional place to examine may be the standardized tests used to calculate those achievement gaps as we find that they may exacerbate the magnitude of the issue. 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 students on average have lower performance on tests than white students.

Finally, part of the end goal of educators may include teaching students to be adaptable to different environments and new concepts and/or to be somewhat knowledgeable about a variety of content. In our setting, this would involve having students read passages on topics they are unfamiliar with or assuming that students have some baseline content knowledge

that would allow them to answer questions effectively. Under this scenario, one would have to reinterpret our results as not about bias in standardized testing but merely confirming that knowledge is important and that certain demographic groups are less equipped to meet the standards of "adaptability" or having a wide breadth of knowledge. While we acknowledge this possibility, to results suggest that this is not at play in our setting. First, in our analysis we find that white students are just as affected by relatability as Black students; white students perform worse on questions with lower relatability just as much as Black students do. Second, we are unable to identify any standards associated with reading comprehension tests that identify this as a goal of the test. Insofar as this is a goal for educators, it is not publicly labelled as such.

References

Adukia, A., Eble, A., Harrison, E., Runesha, H. B., & Szasz, T. (2021). What We Teach About Race and Gender: Representation in Images and Text of Children's Books. (SSRN Scholarly Paper No. 3825080). Social Science Research Network. https://doi.org/10.2139/ssrn.3825080.

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

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.

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

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.

Nickell, S. (1981). Biases in Dynamic Models with Fixed Effects. *Econometrica, 49*(6). https://www.jstor.org/stable/1911408

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/10.1037/0022-3514.69.5.797.

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

Table A1: 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 A2: Relative exposure of groups to topics

	Asian-white	Black-white	Hispanic-white
Animal sports	0.000	0.024	0.540
Animals	0.325	0.316	0.488
Arts and crafts	0.599	0.431	0.528
Baseball	0.337	0.254	0.747
Basketball	1.153	3.044	0.923
Computer	1.486	0.659	0.787
Exercise	1.366	0.721	0.858
Food	1.016	0.852	0.910
Football	0.272	0.918	1.698
Indoor recreation	0.510	0.731	0.486
Media	0.878	0.994	0.971
Misc sport	1.093	0.665	0.841
Music	1.158	1.464	1.324
Nature sports	0.276	0.263	0.404
Performing arts	0.792	0.715	0.779
Plant/garden/yard	0.698	0.464	0.688
Religion	1.179	1.900	1.187
Club sports	0.899	0.211	0.250
Soccer	1.145	0.677	3.834
Street sports	1.641	0.820	0.983
Traveling	0.877	0.872	0.958
Vehicles	0.546	0.692	0.940
Volunteering	0.513	0.674	0.553
Water sports	0.791	0.215	0.517
Winter sports	0.621	0.253	0.052

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.

Table A3: Dictionary of terms related to each topic (1/3)

Arts and crafts	**Food**	**Plant/garden/yard**	**Animals**	**Vehicles**	**Media**	**Music**	**Computer**	**Performing arts**
paint	bake	grow	cat	car	movie	guitar	web	theater
sew	cook	plant	dog	vehicle	film	music	computer	dance
knit	dice	garden	kitten	engine	watch	piano	screen	play
wool	cut	yard	puppy	oil	see	song	desktop	act
yarn	saute	grass	veterinary	mechanic	blockbuster	band	monitor	actor
fabric	oven	lawn	pet	repair	actor	sing	keyboard	stage
decorate	pan	green	animal	tire	television	artist	mouse	lines
style	knife	soil	feed	drive	tv	radio	click	practice
draw	knead	dirt	kennel	steer	screen	violin	type	musical
craft	dough	flower	horse	automobile	channel	string	program	encore
create	flour	petal	litter	suv	stream	sound	software	choir
mold	sugar	seed	litterbox	truck	theater	vocal	hardware	routine
clay	spices	root	reptile	sedan	plot	compose	internet	choreography
furniture	salt	fertilize	sniff	coupe	camera	drum	email	perform
design	taste	pollenate	train	gear	media	rhythm	app	
rug	flavor	pollen	bird		popcorn	beat	application	
weave	ingredients	harvest	rabbit		video	tune	install	
color	fry	plough	donkey			melody	download	
watercolor	cream	tree	mule			harmony	webpage	
carve	vegetable		leash			rap	website	
woodwork	fruit					rhyme	online	
quilt	bread					lyric		
art	pastry					flute		
	food					trumpet		
	butter							
	cake							

Table A4: Dictionary of terms related to each topic (2/3)

Baseball	**Basketball**	**Street sports**	**Water sports**	**Football**	**Winter sports**	**Soccer**	**Religion**	**Volunteering**	**Traveling**
baseball	basketball	bike	water	football	ski	soccer	religion	volunteer	travel
glove	dribble	cycle	boat	pads	snow	ball	church	help	drive
mit	guard	bicycle	wave	helmet	snowboard	field	chapel	community	fly
pitcher ¢er	gear	lake	field	gloves	goalie	worship	organize	airplane	
catcher	court	shift	pool	yard	coat	penalty	pray	civic	car
base	shoot	rollerblade	tub	down	ice	referee	praise	fundraise	highway
shortstop	basket	rollerskate	pond	quarterback	mountain	forward	sing	public	airport
midfield	pass	pads	swim	lineman	hill	striker	hymn	serve	bus
field	foul	helmet	swimsuit	sack	lift	goal	priest	event	train
outfield	dunk	roll	sail	coach	hockey	midfield	minister	teach	platform
dugout	layup	walk	ancor	team	puck	pass	preach	lead	gate
cleats	rebound	pedal	kayak	tackle	stick	offsides	reverend	protest	station
pitch	guard	brake	canoe	touchdown	blades	team	bishop	provide	vacation
out	backboard	chain	padle	kick	skates	coach	pew		roadtrip
bat	bench	wheel	lifejacket	goal	goalie	pitch	bible		freeway
homerun	team	inline	float	fumble	penalty	goalkeeper	scripture		
strike	coach	skate	surf	cleats	cold	knockout	spirit		
strikeout			waterski	flag	winter	captain	mosque		
mound				intercept	boot	defender	islam		
softball					binding		muslim		
umpire					alpine		christian		
safe					crosscountry		jew		
catch					nordic		temple		
inning							buddhist		
team							synagogue		
coach							sermon		
							altar		
							heaven		
							clergy		

Table A5: Dictionary of terms related to each topic (3/3)

Exercise	**Indoor recreation**	**Animal sports**	**Rich sports**	**Nature sports**
aerobics	bowl	rodeo	golf	climbing
yoga	billiard	horse	club	rock
mat	pool	ride	driver	cave
stretch	cue	jockey	iron	fish
hold	pin	cowboy	putter	fly
practice	strike	boots	green	rod
run	spare	hat	faraway	bait
work	gutter	polo	tee	tackle
weights	stripe	saddle	par	worm
lift	solid	gallop	eagle	catch
strength	eightball	ride	tennis	hike
weightlift	lane	mule	bogey	trail
sweat		donkey	racquetball	vista
workout			squash	views
exercise			racquet	hill
cardio				mountain
fitness				dirt
				hunt
				elk
				deer
				rabbit
				cast
				trout
				bass
				salmon
				rifle

Table A6: Effect of relatability on falsification outcome variables

	(1) $Y_{dp}^{prev_grade}$	(2) $Y_{dp}^{prev_cohort}$	(3) $RaceShare$
Relatability (r_{dp})	0.003	0.006	-0.003
	(0.006)	(0.008)	(0.002)
# of topic-passages	4,300	3,550	5,125

Notes: $^{*}p < .10$,$^{**}p < .05$,$^{***}p < .01$. Each specification is an OLS regression of race-passage level outcomes on content relatability. Unreported controls include race-by-grade fixed effects and differential slopes on 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. The outcome in column (1) is student performance for race d on the prior year exam for the same grade and passage position. The outcome in column (2) is student performance for race d on the prior year exam for the same cohort and passage position. The outcome in column (3) is the share of test takers of race d on the exam containing passage p. Observations are at the race-passage level. The estimation sample differs across each column; 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. For column (1) this excludes all exams from 2013. For column (2) this excludes all exams from 2013 and grade 3.

Table A7: Effect of relatability on falsification outcome variables

	(1)	(2)	(3)
Relatability (r_{dp})	0.0187***	0.0172***	0.0166***
	(0.005)	(0.006)	(0.005)
# of topic-passages	4,300	3,550	5,125

Notes: $^{*}p < .10$,$^{**}p < .05$,$^{***}p < .01$. Each specification is an OLS regression of student performance on content relatability. Unreported controls include race-by-grade fixed effects and differential slopes on 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 differs across each column; 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. For column (1) this excludes all exams from 2013. For column (2) this excludes all exams from 2013 and grade 3. These sample restrictions are analogous to those found in Table A6

Table A8: 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 demographic familiarity measures

In addition to our standard measure of demographic 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 demograhpic familiarity from that subsample, using the process described in section 3.2.

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[17] 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 re-calculate 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},

[17] 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 leaving the words unstemmed and using only nouns. 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.

C 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, we should calculate standard errors in a regression at the topic-passage level, which is the level in which the treatment varies. 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.[18] 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 4.1.

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

[18] 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

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

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.

D Robustness

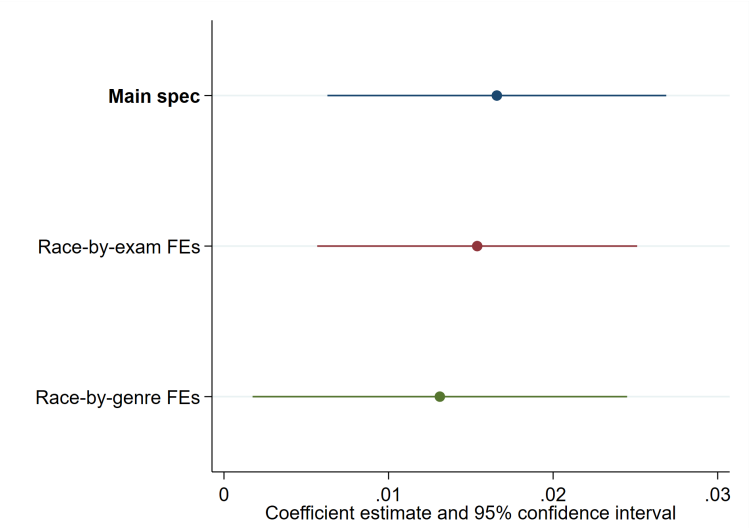
Our estimates are downstream of modeling and data choices that are baked into our race-based relatability measure and model specification. First, we show that the estimates are qualitatively the same with alternative fixed effect specifications. Second, we show that our results are not sensitive to specific 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 D3). 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 D4 we demonstrate that the coefficient is remarkably 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 10% 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 D5 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 D3: Robustness of baseline effect to different levels of saturation in the specifications of fixed effects

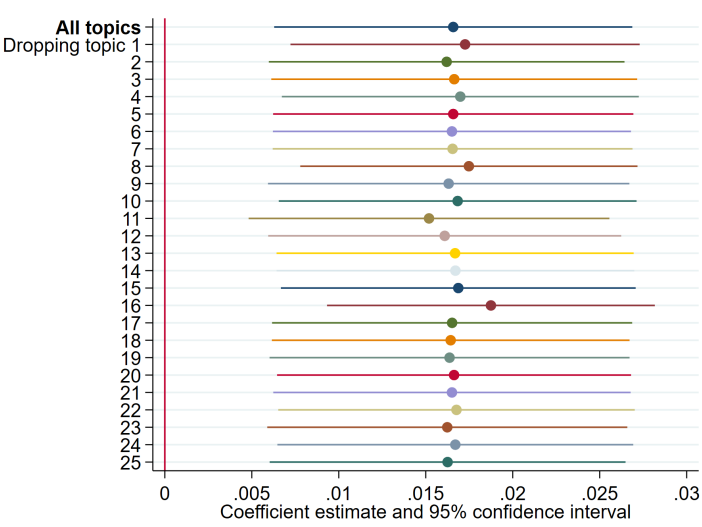


Notes: Each specification is a regression of the share of items answered correctly on a passage on content relatability. Unreported controls include: (1) "unit" fixed effects at the level indicated in the legend and (2) exposure-sum-by-passage fixed effects. Standard errors clustered by exam are used 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.

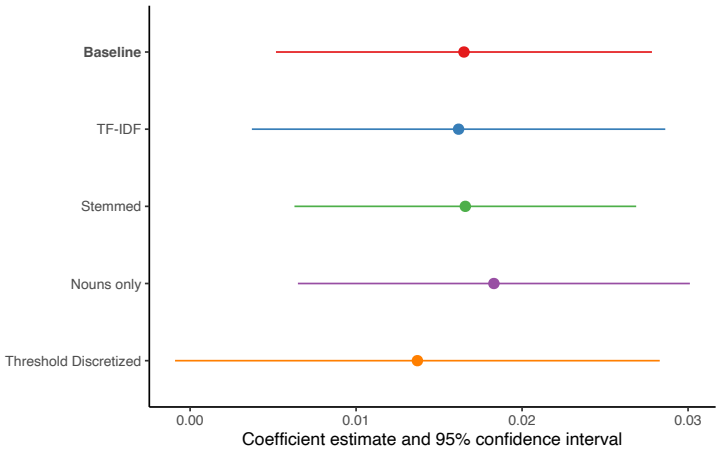
Lastly, we consider the importance of the ATUS respondent sample we use to construct the race-based exposure scores. In our main specification, we use the entire ATUS data to construct exposure, but here we consider filtering down to subgroups that may be more representative (but have smaller sample size), discussed in section B.1. We see in Figure D6 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 data as-is.

Figure D4: Robustness to topic set



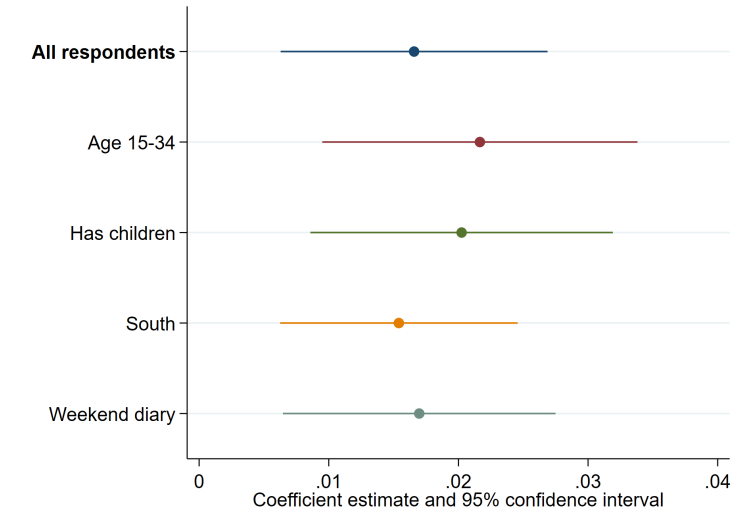
Notes: Each row after the first is a separate regression of the fraction of questions correct for a passage on relatability, after leaving the indicated topic category out. Unreported controls include: (1) race-grade fixed effects and (2) exposure-sum-by-passage fixed effects. Standard errors clustered 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.

Figure D5: 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 D6: 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.