AER Disclosure Statement for "Content Relatability and Standardized Testing: Evidence from Texas"

Steven Lee

1. This article received financial support from the Population Studies and Training Center at Brown University, the Bravo Center for Economic Research at Brown University, the Annenberg Institute at Brown University, and the American Institute for Boys and Men.
2. I have not received significant financial support from any interested party in the past three years.
3. I do not have any paid or unpaid positions from relevant organizations.
4. No close relative or partner has any conflicts to disclose.
5. The Texas Education Research Center (ERC), which provided the main dataset which underlies the analyses in our article, requires research products to be reviewed for FERPA compliance. We have already received confirmation of compliance for all exhibits contained within the article.

The research protocol contained in our article has been determined to be exempt from 45 CFR 46 by the Institutional Review Board at Brown University. A copy of the exempt determination memorandum can be provided upon request.

AER Disclosure Statement for "Content Relatability and Standardized Testing: Evidence from Texas"

Matthew Schaelling

1. This article received financial support from the Population Studies and Training Center at Brown University, the Bravo Center for Economic Research at Brown University, the Annenberg Institute at Brown University and the American Institute for Boys and Men.
2. I have not received significant financial support from any interested party in the past three years.
3. I do not have any paid or unpaid positions from relevant organizations.
4. No close relative or partner has any conflicts to disclose.
5. The Texas Education Research Center (ERC), which provided the main dataset which underlies the analyses in our article, requires research products to be reviewed for FERPA compliance. We have already received confirmation of compliance for all exhibits contained within the article.

The research protocol contained in our article has been determined to be exempt from 45 CFR 46 by the Institutional Review Board at Brown University. A copy of the exempt determination memorandum can be provided upon request.

Content Relatability and Standardized Testing: Evidence from Texas*

Steven Lee† Matthew Schaelling‡

January 1, 2026
(click here for latest version)

Abstract

One goal of standardized tests is to measure aptitude across heterogeneous students with minimal bias. That different students relate in different ways to topics and characters in exam content could, however, lead to differential test scores even when their aptitude is the same. We study how differential content relatability can affect test scores using item-level data from reading-comprehension exams for primary and secondary school students in Texas. Using time-use data and natural language processing techniques, we first build a novel measure of race- and gender-based relatability to topics in the exams' text passages. A 1 standard deviation increase in exam-level topic relatability across race predicts a 0.05 standard deviation change in exam performance, with null effects for gender. We also find that test scores improve on passages with a higher share of characters matching either the student's race or gender. Our estimates suggest that equalizing the relatability of passages in these standardized tests could reduce the Black-White and Hispanic-White testing gaps by up to 8 and 10 percent, respectively. We then counterfactually estimate that during our sample period, close to 12,000 Black students and 34,000 Hispanic students would have been classified at a higher reading-comprehension level had relatability been more equal.

*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 in 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 general support (P2C HD041020). We additionally acknowledge funding support from the Bravo Center for Economic Research at Brown University, the Annenberg Institute for School Reform at Brown University, and the American Institute for Boys and Men. 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.

†Brown University, Department of Economics, Box B, Providence, RI 02912; steven_lee@brown.edu

‡Brown University, Department of Economics, Box B, Providence, RI 02912; matthew_schaelling@brown.edu

1 Introduction

Inequality along demographic dimensions is well documented and widespread in education. In the United States, achievement disparities 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). For example, there is a 14 percentage point (pp) difference between White and Black students and a 3 pp difference between female and male students on third- and fourth-grade reading-comprehension exams in the U.S. state of Texas. In response, some observers and policymakers have called for a deeper understanding of this testing gap and the mechanisms behind it. Some have advocated abandoning standardized testing entirely, claiming that it systematically disadvantages some groups, while others insist that standardized measurement is essential to accountability and progress toward racial and economic equality in education.[1]

If standardized tests fail to measure achievement consistently across student backgrounds, it is essential to identify the test attributes that contribute to such mismeasurement and quantify their impact. One possible factor is the degree to which educational content is *relatable* to certain groups of students, as students might learn or perform better when encountering topics that are familiar and interesting or characters with whom they identify. Studies in educational psychology show that interest in a topic can affect performance on reading-comprehension tests and that these interests diverge by race and gender (Bray & Barron, 2004; Asher, 1979). Other research documents race and gender gaps in identity representation in educational materials, such as children's literature (Adukia et al., 2023).[2]

Empirically analyzing the relationship between test performance and the relatability of content poses several challenges. First, student familiarity with or interest in topics is inherently qualitative and typically unobserved in existing datasets. Even when quantitative representations of students' topic familiarity or identity representation are available, they are likely correlated with other test constructs that exam designers intentionally vary, such as vocabulary difficulty or text length. Second, students' topic familiarity and interest may correlate with ability, which poses a threat to identification. Finally, topics and character identities may covary, so examining either attribute in isolation risks omitted variable bias.

[1] See, for example, discussions by the National Education Association ("The Racist Beginnings of Standardized Testing," https://www.nea.org/nea-today/all-news-articles/racist-beginnings-standardized-testing) and in *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 academic discussion of such topics, including by psychometricians (Boykin, 2023) and economists (Card & Giuliano, 2016).

[2] This work, along with ours, contributes to a growing literature using text analysis in causal social science research (Gentzkow & Shapiro, 2010; Loughran & McDonald, 2011; Hassan et al., 2017).

In this paper, we study whether and how differential relatability to exam content across students affects exam performance and estimation of demographic test-score gaps. Our two measures of content relatability reflect a student's connection to topics and character identities in the text, which may vary across race and gender. We build a novel measure of *topic relatability* using time-use data and natural language processing techniques. We complement this with a measure of *identity relatability*, constructed with predictions of the race and gender of characters in the text. Applying these measures to text passages in reading-comprehension exams in Texas, we obtain causal estimates of topic and identity relatability on test scores. Using these estimates, we calculate the share of demographic test-score gaps explained by relatability.

The core concern motivating our inquiry is that two students of equal ability but different demographic backgrounds may receive different test scores. This discrepancy arises from the interaction between a student's characteristics and the passage-text attributes, specifically those attributes that the exam designer does not intend to influence test performance. We focus on multiple factors in this systematic student-passage interaction, such as personal interest in passage topics, familiarity with those topics, and shared identity with passage characters. Topic relatability corresponds to both familiarity and interest factors, while identity relatability captures the demographic representativeness of a passage.

The estimation data cover the universe of third- to eighth-grade public school students from 2013 to 2019 in Texas. The administrative data contain reading-comprehension exam responses for every student and test question. The questions are directly linked to reading-passage text, providing data on passage-level test performance for each student. We also observe the race and gender of every student, which we use to build our relatability measures. We supplement these testing data with detailed time-use information from the American Time Use Survey (ATUS) and name/demographic-group databases from the US Social Security Administration and state voter registration files.

Our topic-relatability measure is constructed by combining demographic-level exposure to topics with the salience of topics in a passage. We start by selecting a set of leisure activities from the ATUS and group them into topic categories (e.g., basketball, gardening). Topic exposure for a race or gender group is the share of ATUS respondents in that group who spend any time on activities related to that topic. The salience of a topic in a passage is the share of the passage's words directly related to the topic. We join topic exposure and topic salience together to create the demographic-passage-level topic-relatability measure. To obtain the causal effect of topic relatability on test scores, we isolate variation in relatability that is as good as random. This leads directly to our shift-share empirical strategy, which is premised on the idea that while topic exposure (the "shares") is nonrandom, only topic salience (the

"shifts") needs to be quasi-random subject to appropriate controls (see Borusyak, Hull, and Jaravel 2022). For example, topic exposure to gardening across races is not random, but it is random whether Black fourth graders see a basketball passage in one year and a soccer passage in another.

We find that the race-based topic relatability of a reading-comprehension passage causally raises student performance on questions connected to the passage. A one standard deviation (SD) increase in race-based topic relatability in a passage leads to a 1.9 pp increase in the share of correct answers on that passage. This is equivalent to a 0.05 SD increase in test scores from a one SD increase in exam-level relatability. While these effect sizes are small in absolute terms, they are moderately sized in comparison to other factors affecting test outcomes.[3] Race-based topic-relatability effects are strongest for lower-achievement students. We also find suggestive evidence that race-based topic-relatability effects are strongest for students in more racially homogenous schools—those who are less likely to be exposed to a diverse set of topics. In contrast to the race-based topic-relatability results, gender-based topic relatability is not predictive of test scores. Given that cross-gender topic exposure is higher than cross-racial exposure, the school and gender results highlight the potential importance of peer effects in mediating the impact of topic relatability. Our results are robust to alternative topic-relatability measures, including using time-use data for children to construct the measure.

We supplement the topic-relatability results by assessing the impact of identity relatability. We define identity relatability as the share of characters matching the student's race or gender. To construct this measure, we first use a large language model (LLM) to identify all characters in a text. We predict race and gender for each character. For race, this is done by linking first and last names to SSA data; for gender, the LLM imputes gender based on pronouns or titles. The results for both race and gender are similar. For race, we show that moving from zero own-race characters to all own-race characters results in a 1.1 pp change in a student's test performance. For gender, we show that moving from zero same-gender to all same-gender characters yields a 0.8 pp change in test performance. These results are robust to using different LLMs, using different imputation methods for race and gender, and weighting characters by number of mentions when calculating the race or gender share.

We consider the total combined effect of relatability across our two measures. To fill the gaps left by using our structured measures of relatability, we invite respondents to an online survey platform to provide an unstructured assessment of passage relatability. This flexible survey-based measure is positively associated with identity relatability and predictive of

[3]For instance, Chetty, Friedman, and Rockoff (2014) find that a 1 SD increase in teacher value added raises English test scores by 0.1 SD.

test scores. Estimating the test-score impacts of topic relatability, identity relatability, and survey-based measures in one regression, we find that the topic- and identity-relatability measures maintain their effect sizes, suggesting that these effects are separate from one another.

We investigate the extent to which content relatability contributes to disparities in test outcomes across demographic groups, focusing particularly on differences across race. We find that simply setting topic and identity relatability to be the same for all racial groups would lead to a 8% smaller Black–White test gap and a 10% smaller Hispanic–White test gap. Around one-third of the estimated topic-relatability effect on test gaps is due to the selection of more White-relatable topics in our sample of passages. The results for average test-score gaps mask the role that test-score-based thresholds play in exacerbating racial disparities for individual students. We examine how each student's state-determined reading-performance category would change if, counterfactually, they had received a more relatable exam from our sample. We find that around 12,000 Black students and 34,000 Hispanic students would have been assessed at a higher reading-comprehension standard if relatability had been more equal across groups.

We contribute to the literature documenting and explaining demographic gaps in educational outcomes, a core issue in the economics of education. One strand of this literature examines race, asking what covariates might explain the gap that begins in very early childhood (Fryer & Leavitt 2004; Fryer & Leavitt 2013) and examines the sensitivity of group differences in test scores to scaling decisions (Bond & Lang, 2013; Bond & Lang, 2018; Nielsen, 2023). The relationship between segregation and the racial SAT test gap is examined in Card and Rothstein (2007), raising questions about peer effects that we address. Another strand of this literature examines gender, documenting that female students perform better than male students on reading exams, a finding that is consistent across geographic location (Pope & Sydnor 2010), present within socioeconomic status (Cobb-Clark & Moschion 2017), and robust to many other controls (Lundberg 2020). Recent work by Brown et al. (2022) considers the role of cognitive endurance in explaining the socioeconomic test-score gap. While these papers have documented many factors that explain variation in the demographic testing gaps in question, we propose another variable that may improve the accuracy of measurement of racial or gendered testing gaps. In particular, we consider whether test content itself may affect estimated gaps on exams administered with real-world stakes.[4]

The literature has also explored the representation of race, culture, and gender in ed-

[4]There is evidence suggesting the importance of a test's setting when estimating performance gaps: Ofek-Shanny (2024) demonstrates with a field experiment on Israeli students that estimation of gaps is more credible on higher-stakes exams. This motivates our decision to use real-world exams with real-world consequences.

ucational materials, as well as how it might affect student success. Recent papers have documented essential facts about race and gender in learning materials, such as the underrepresentation of certain identities, using text-analysis and computer-vision tools (Adukia et al., 2023; Lucy et al., 2020) or using novel survey methods for detecting stereotypes (Baldazzi et al., 2025). Researchers also find that racial- or ethnic-coded content in educational text can affect students' outcomes and beliefs (Dee & Penner, 2017; Cantoni et al., 2017). Other papers study exams directly. Dobrescu et al. (2021) experimentally varies a standardized test's cultural context in Australia, while Dee and Domingue (2021) tests Steele and Aronson's (1995) stereotype-threat theory by looking at the impact of a culturally insensitive test questions in the state of Massachusetts. For gender, Good et al. (2010) finds symmetric gender testing effects when educational materials feature a gender identity match, but Cohen et al. (2023) finds gender-neutral language only improves women's performance. Finally, in a study closely related to our own, Duquennois (2022) finds that students of low socioeconomic status do worse on money-themed math practice and test questions.

We add to this literature in several ways. First, we construct a rich and novel measure of relatability that explores demographic dimensions of both race and gender. This measure not only captures widely studied features such as identity representation, but considers less prominent features such as differential topic familiarity and interest across demographic groups. We provide a way to assess these features using time-use data. To the best of our understanding, we are the first to use time-use data for this purpose. Second, we obtain our estimates from a large-scale standardized exam, providing a direct measure of the impact of educational material on a student outcome with real-world stakes. Finally, we propose an innovative identification strategy suited to large-scale exams administered uniformly on a population, whereas other approaches rely on random assignment of exam booklets or focus on the impact of one test question ("item") out of many.

Last, our paper is complementary to the extensive psychometrics literature on differential item functioning (DIF) (see Zumbo (1999) for an overview). DIF models use exam outcomes to detect item- or test-level differential performance across student characteristics, conditional on ability. However, distinguishing between benign and adverse DIF is difficult (Douglas et al. 1996). There are no unambiguous quantitative methods for interpreting whether detected differences represent bias, and thus it is left to test makers to look at the ex post exam *outcomes* to determine whether a test item flagged by a DIF model has desirable or undesirable properties. We make progress on this issue by developing two transparent, structured measures that indicate potential sources of bias based on the ex ante exam *content*. In principle, this allows test makers to tailor exams to their institutional objectives prior to field testing or administration. Further, our approach can capture differential impacts across

groups that might fall below conventional DIF thresholds.

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 the time-use data. Section 4 discusses the topic-relatability estimation strategy and results. Section 5 describes the identity-relatability estimation strategy and results. We bring the topic-relatability and identity-relatability results together in Section 6. Section 7 extends our results to study how relatability differentially affects students. Finally, Section 8 concludes.

2 Conceptual framework

Essential to our exercise is distinguishing between a student's reading ability and the extent to which they relate to a text. We present a simple model to illustrate how relatability influences test scores, followed by a practical discussion of what factors may affect relatability.

2.1 Model

The primary objective of exams is to measure the ability and progress of students. However, the signal observed via standardized testing may be a function of learning *and* other factors that the testing administrator might not want to consider. For example, consider a biographical excerpt about a sailor. Comparing two students who have identical reading-comprehension ability but differ in exposure to boating and the sea, we might not be surprised to find that the student with greater exposure performs better on questions regarding this passage. 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

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

Here, i indexes a student and p indexes a passage.[5] 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. Now, we parse ρ_{ip} into two parts: one that represents

[5]For ease of exposition, we omit from the model that each passage is accompanied by multiple questions, as is the case in most exams.

systematic variation, and one that is idiosyncratic and uncorrelated with θ_i. That is,

$$\rho_{ip} = \underbrace{\vec{\varepsilon}_i' B \vec{\mu}_p}_{\substack{systematic \\ \text{or "relatability"}}} + \underbrace{\nu_{ip}}_{idiosyncratic}$$

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

Educators only observe y_{ip}, and they use it to draw inferences about individual-student learning outcomes. For instance, educators 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 (for example, the entirety of a single reading-comprehension exam). Educators may also use test results to compare learning levels across groups, such as classrooms, schools, regions, or demographic groups. It follows that educators may set $\tilde{\theta}_d \equiv \frac{1}{N_d} \sum_{i \in d} \tilde{\theta}_i$ as the group-level performance indicator for some group d (where N_d is the number of students in group d). Then $\tilde{\theta}_{d_0} - \tilde{\theta}_{d_1}$ is the difference between groups d_0 and d_1.

The testing-outcomes model we propose suggests that these simple estimates (based only on y_{ip}) of student- and group-level ability may be biased because of 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 in a given grade and school year—as is often the case—there is no bias when comparing students or groups. The potential correlation between components of $\vec{\varepsilon}_i$ and θ_i, in contrast, poses a threat to interpreting standardized test outcomes $\tilde{\theta}$. Consider again a reading passage about a sailor. We represent the presence of nautical themes in this passage as an element of $\vec{\mu}_p$ such that $\mu_{p,sea} = 1$. Student exposure to nautical activities may similarly be a component of $\vec{\varepsilon}_i$; thus, student i_1 has $\varepsilon_{i_1,sea} = 1$, while student i_0 has $\varepsilon_{i_0,sea} = 0$. Then our model suggests there is a systematic wedge in observed performance between students i_1 and i_0 of $\varepsilon_{i_1,sea} B_{sea,sea} \mu_{p,sea} - \varepsilon_{i_0,sea} B_{sea,sea} \mu_{p,sea} = B_{sea,sea}$. If attributes like topic exposure are correlated with demographic groups, such as race or income level, then conclusions about group-level outcomes are also affected by this issue.[6]

This model framework and its implications drive our definition of content relatability and our estimation strategy. We refer to the interaction between $\vec{\varepsilon}_i$ and $\vec{\mu}_p$ as content relatability. First, we limit our focus to attributes of the text corresponding with elements of $\vec{\varepsilon}_i$ and $\vec{\mu}_p$, which are arguably outside the scope of evaluation for reading-comprehension exam

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

designers. In our setting, this includes selecting topics, such as sports or arts and crafts, that students are differentially exposed to that are orthogonal to reading comprehension. This also includes the names of characters in the text, which may evoke different relatability across students.

We qualitatively evaluate this by considering the testing standards that educators in our setting explicitly seek to assess through their standardized tests. The standards assessed in the state of Texas during our sample period (2013–2019) predominantly focus on reading ability, neutral to the topics or the character identities we study (Texas Administrative Code §§ 110.10–110.20 2008). The standards focus on vocabulary, grammar, genre-specific structure, and the ability to make comparisons across multiple texts. In particular, students are not expected to have background knowledge about the subject of the text or its topics. Still, a handful of tested provisions related to *culture* deserves particular attention. In grade 8, for example, students must "...make inferences and draw conclusions about theme and genre in different cultural, historical, and contemporary contexts..." (Texas Administrative Code §§ 110.10–110.20 2008). These and related standards could arguably overlap with topics and character identities.

However, we argue that the small number of provisions related to "culture" has limited impact on the implications of our results. First, many of the topics we study do not carry cultural or historical markers. For example, we include cooking and music as topics, which cut across all cultural contexts. Second, to the extent that some of our topics or identities are associated with different histories and cultures, the standards do not specify which histories and cultures should be studied. In that way, even if content relatability reflects some ability for students to engage with questions that bridge different cultural contexts, it remains true that the selection of certain cultural contexts or historical contexts over others reflects an advantage or disadvantage to certain groups of students.

Next, our empirical strategy focuses on how test attributes change test scores at the race/ethnicity and gender levels, as they are of primary interest to educators and researchers, though we acknowledge the potential for variation in the presence of these topics to cause differential performance at non-demographic-group levels (for example, students interested in baking compared to their peers). Given our interest in race- and gender-level estimation, we use race- and gender-level data when constructing estimates of elements of $\vec{\varepsilon}_i$. 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 7.

2.2 Components of content relatability

We conceptualize content relatability as comprising three components: interest, familiarity, and identity. Each of these components can be represented as a match between student characteristics $\vec{\varepsilon_i}$ and passage characteristics $\vec{\mu_p}$. For example, included in $\vec{\varepsilon_i}$ is student i's interest in baking, while $\vec{\mu_p}$ includes the degree to which baking appears in passage p. Similarly, $\vec{\varepsilon_i}$ may include a term representing a student's identity, with an analogous term in $\vec{\mu_p}$ representing whether characters with that identity appear in passage p. We collapse familiarity and interest in creating and defining our topic-relatability measure (Section 4). We then propose a separate measure of identity relatability (Section 5).

Topic relatability is a function of personal interest in a topic, such as pets. For example, a student with strong interest in a passage's topic may find it easier to focus on the test. However, a student with no interest in it may still be quite *familiar* with it if it is popular among their family members or others in their community. Even with no interest in football, they may still know football-specific terms and concepts if their siblings or parents are fans. A carefully designed laboratory experiment would be necessary to credibly disentangle these concepts and their mechanisms. Instead, this paper focuses on quantifying these effects in observational data. A well-developed educational psychology literature directly speaks to the ideas of familiarity (Singer and Alexander, 2016; Johnston and Pearson, 1982; Davey and Kapnius, 1985; Norris et al., 2003) and interest (Norris et al., 2003; Shirey and Reynolds, 1988; Schraw et al., 1995) as they relate to reading comprehension. Given the close relationship between interest and familiarity, we collapse these concepts together into one measure. Through our empirical strategy, we seek to represent each student's topic exposure (interest or familiarity) and each exam passage's topic salience (presence of topics in the text).

Similarly to topic relatability, a student who *identifies* with characters in a text has high identity relatability. In this paper, we focus on identity relatability that arises from a student identifying with same-race or same-gender characters in the text. Identification with a character is not limited to race and gender—for example, a student may identify with a character that is shy or the oldest sibling. Nonetheless, we restrict our attention to race and gender since they are relatively observable characteristics in both the administrative data and the passage text and as they have been the focus of recent scholarly work on identity representation in educational material (Lucy et al., 2020; Adukia et al., 2023). As with our empirical strategy for topic relatability, we need data on student demographic characteristics and passage-character characteristics, which interact to form identity relatability.

3 Data

3.1 Texas student and assessment data

Our primary dataset comes from the Texas Education Agency (TEA), which administers standardized assessments to students in Texas. We study the State of Texas Assessments of Academic Readiness (STAAR), mandatory end-of-year assessments of public school students in grades 3–8. Students in the same grade or course receive the same test in a given year, with two exceptions: The TEA offers Spanish-language tests for 3rd to 5th grade students and alternative tests for students with cognitive limitations. We only consider students who take the standard English-language exam. To avoid concerns regarding the COVID-19 pandemic's effect on education, we limit our attention to 2013–19 reading-comprehension exams. The format of reading exams is standard across grades. Students read four to six text passages and answer multiple-choice questions ("items") regarding each passage, including its content, vocabulary, and grammar.

Using the item-level student responses, we define a binary outcome measure of whether a student answered a question correctly. We match each item to its corresponding reading-comprehension passage. If a set of items is associated with two pieces of text, both texts are combined and considered to be one passage for the purpose of analysis. The item-level data also include the TEA-designated reading standard associated with each item, which we use to define the genre of each passage. Our final dataset includes student responses and passage characteristics for 205 unique passages from 42 exams.[7] We provide a brief description of an example passage in Appendix Section B.1.

We supplement the testing data with students' demographic information, including their reported race, ethnicity, and sex.[8] From that information, we create four racial/ethnic categorizations: Asian, Black, Hispanic, and White. We remove from our sample students reporting as another race or multiple races because of small sample sizes or limited ability to link these data to external data. All Hispanic-identifying students are categorized as Hispanic, and the remainder of students are categorized according to their reported race. For brevity, we use *race* to mean both race and ethnicity for the remainder of this paper. Data on each student's free- or reduced-price lunch status and their school attended are used in supplementary analysis.

Our analysis sample consists of 13,180,138 student-exam observations (henceforth, "stu-

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

[8] The TEA data and all other administrative data used in this paper include only binary sex indicators and do not distinguish between sex and gender. Consequently, any references to gender in this paper should be interpreted as referring to recorded sex.

Table 1: Summary statistics for student sample

	Mean	SD
Demographics		
Hispanic	0.46	
Non-Hispanic		
Asian	0.05	
Black	0.15	
White	0.34	
Female	0.49	
Test		
Passages	5.05	0.44
Test items	43.17	5.32
Score (0–1)	0.67	0.20
N	13,180,138	

Notes: The table displays sample means and standard deviations of characteristics in the student sample. The sample includes all students in grades 3 to 8 from 2013–19 taking the standard STAAR reading-comprehension test.

dents"). Table 1 summarizes the student population. The plurality of students are Hispanic (46%), much higher than the US average.[9] Among non-Hispanic students, the majority are White (34% of total) rather than Black (15%) or Asian (5%). As expected, students are split roughly evenly by sex. Most exams contain five passages, corresponding to just over 43 test items, per Table 1. Students answer roughly two-thirds of these questions correctly.

3.2 Time-use data

3.2.1 American Time Use Survey (ATUS)

We use the ATUS to construct measures of topic exposure at the demographic level. The ATUS, sponsored by the Bureau of Labor Statistics and conducted by the Census Bureau, is a nationally representative survey that provides estimates of how Americans 15 years and older spend their time. Randomly selected individuals from a subset of households in the Current Population Survey report a diary of the prior day's activities to a phone interviewer. The interviewer assigns each activity to one of 442 six-digit classification codes, nested within 17 major categories. Each code reflects three levels of detail regarding the activity. For example, the activity "sewing" is classified as code 02, representing the major category "Household Activities"; 0201, representing the activity group "Housework" within

[9] According to US Census estimates for 2019, 20% of individuals aged 5 to 19 were Hispanic.

"Household Activities"; and 020103, representing the activity "Sewing, repairing, and maintaining textiles."

We reduce the dimensionality of the detailed activity data by excluding codes that are unrelated to leisure or reflect life-maintenance activities performed by virtually all individuals. Appendix Table A1 summarizes the excluded activity codes. Although we exclude almost two-thirds of codes, most excluded minutes (68%) are spent on commonly engaged-in activities such as sleeping, working, and eating. We classify the remaining 140 activity codes into 24 mutually exclusive topics, including categories such as arts and crafts, animals, soccer, and winter sports (see Appendix Table A2 for examples). A detailed mapping of activity codes to topics is available in the online code repository.

Our ATUS sample includes all respondents from 2013 to 2019, the same period as our student data. For each respondent, we observe minutes spent on each ATUS activity, as well as their race, household size, household income range, and other demographic characteristics. We follow the same racial/ethnic categorization in these data as we do for the student data: excluding multiracial respondents and placing Hispanic respondents in their own category.

Table 2 describes the individual time diaries and demographics of our 73,626 respondents. Each respondent is weighted by sampling weights provided by the US Census Bureau. On average, a respondent reports 11 activities in their diary, spending 146 minutes per activity.[10] Further, a respondent's daily activities typically correspond to two to three topics. Activities in these topics correspond to 20% of an average respondent's time. 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 a 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.

For the main analyses, we use data from all respondents in the ATUS sample period, which includes older adults and respondents outside of Texas. Incorporating data from a large sample reduces the risk that measurement error in the relatability measure will bias the coefficient estimates. However, using this larger group comes at the expense of keeping the ATUS sample representative of the estimation sample: children residing in Texas. To address this concern, we consider more relevant ATUS subsamples based on age, household composition, and geography in a supplementary analysis. We also later consider children's time-use data from the Child Development Supplement of the Panel Study of Income Dynamics (PSID-CDS). More details on the subsamples and PSID-CDS data can be found in

[10]Dividing the number of minutes in a day by the average number of reported activities does not yield 146 minutes per activity. This is because the ratio of the averages (average number of minutes divided by average number of activities) is not equivalent to the average of the ratios (averaging minutes per activity across respondents).

Table 2: Summary statistics for ATUS sample

	Mean	SD	5%	95%
Activities				
Number of activities	11.34	4.22	5	19
Number of min spent per done activity	145.96	63.05	76	286
Number of topics	2.32	1.22	1	4
Share of time spent in topics	0.20	0.15	0.00	0.48
Demographics				
Asian	0.05			
Black	0.12			
Hispanic	0.16			
White	0.66			
Female	0.52			
N	73,626			

Notes: The table displays sample means, standard deviations, and the 5th/95th percentile values for each category. Respondent observations are weighted by ATUS sampling weights provided by the Census Bureau. The sample includes all ATUS respondents from 2013 to 2019.

Appendix Section B.

3.2.2 Constructing demographic-based topic exposure

Using the ATUS sample, we construct a measure of each demographic group's expected exposure to topics. The key challenge is that we cannot directly observe test takers' interests or familiarity levels. To construct measures of topic exposure, we proxy for these variables using reported time use across race in the ATUS data. Although most ATUS respondents are not school-aged, children are likely to be exposed to the activities of adults in their households (for example, parents, siblings) (see Bandura 1977; Bisin and Verdier 2001).

Formally, we define topic exposure $e_{d,t}$ as the share of respondents in demographic group $d \in \mathcal{D}$ reporting any activity related to topic $t \in \mathcal{T}$, weighting respondents by ATUS sampling weights. We demonstrate differential exposure to activities across race in the ATUS data. In Figure 1, we present ratios of topic exposure by demographic group. We observe wide divergence in exposure to activities by race and gender. For example, in Panel A, Black respondents are significantly less likely than White respondents to have participated in water sports, but much more likely to have participated in basketball. Hispanic respondents are not as exposed to winter sports, but are much more likely to have been exposed to soccer. We also observe that there are substantial differences between the Black-White and Hispanic-White exposure ratios. In Panel B, we see that male respondents are much more likely to

report participating in nature sports or football, while female respondents are much more likely to report participating in arts and crafts and animal sports. We provide the raw exposure ratios for all topics in Table A3.

3.3 Passage-text data

We extract the passage text from publicly available PDFs of the STAAR exams. This corpus of text allows us to generate measures of topic salience for use in constructing measures of topic relatability (Section 4) and identity relatability (Section 5). The topic-salience measure and the categorization of character identity correspond to $\vec{\mu}_p$ in the conceptual framework: a set of passage characteristics that interacts with student characteristics to form relatability.

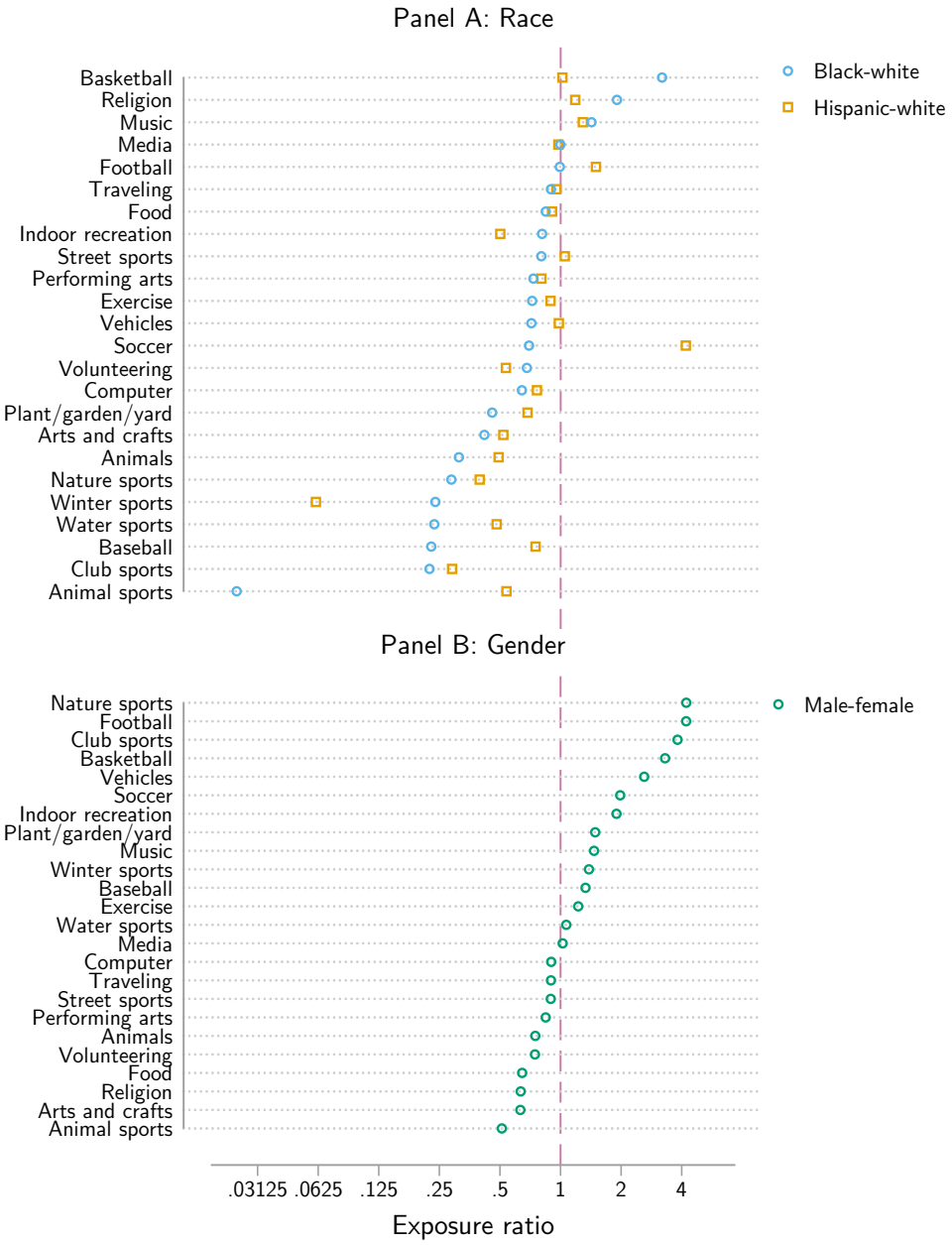
For the construction of topic salience, we score the content of each reading passage using an intuitive *dictionary method* that is standard in the literature on natural language processing. For each topic-passage pair, we measure how salient a topic is in a specific passage. The algorithm operates as follows. First, determine a set of words that 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 share of words in a passage in a given topic dictionary as our measure of topic salience. This is commonly referred to as *term frequency* in natural language processing analysis.

We 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 $t \in \mathcal{T}$. We then construct $B_t = B_{t,1} \cup B_{t,2}$.[11] 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). Our baseline specification also converts all words to their stem, such as converting "dogs" to "dog" or "running" to "run."

Our baseline measure of topic salience, denoted $m_{t,p}$, is simply the share of words in a passage p that is in dictionary B_t. The distribution of this measure, displayed in Figure A1, is heavily right-skewed and the modal value is zero. We provide an example of topic salience for a particular passage in Appendix Section B.1. We also consider other measures of topic salience, textual data cleaning, and expansions (with synonyms) or contractions (random subsamples) of the topic dictionaries, discussed in Appendix Section B.3.

[11] We created two dictionaries $B_{t,1}$ and $B_{t,2}$ prior to any analysis and have not edited the dictionaries since then. The dictionaries are provided in the online code repository. For example, for the activity of *basketball*, each author constructed a list of approximately 20 words related to basketball. These two lists combined to create the dictionary of words for basketball, which is this set: {backboard, ball, basket, basketball, bench, center, coach, court, dribble, dunk, foul, guard, hoop, jump, layup, pass, rebound, shoot, shot, team, tip-off, turnover}.

Figure 1: Differences in topic exposure by race and gender



Notes: The graph plots the ratio of topic exposure between demographic groups. Panel A plots the ratio between Black and White respondents and the ratio between Hispanic and White respondents. Topics are ordered by the Black–White exposure ratio. Panel B plots the ratio between male and female respondents. Topics are ordered by the male–female exposure ratio. The vertical dotted line at 1 corresponds to the value at which both groups have identical exposure. Topic exposure is calculated separately for each racial 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 to 2019.

After generating this measure of topic salience for each passage-topic combination, we consider how topics are distributed in passages. A majority (53%) of topic-passage pairs ($m_{t,p}$) are non-zero-valued—that is, at least one word from topic t's dictionary (B_t) appears in passage p. The average passage has 12.6 (out of 24) non-zero topics. However, most of the non-zero topics have a very low measure of topic salience: the median non-zero value of $m_{t,p}$ is 0.005. If we only count a topic t appearing in a passage p when above a threshold of $m_{t,p} \geq 0.01$, we find that the average passage contains 3.8 topics; alternatively, a threshold of 0.02 suggests the average passage contains 1.6 topics. Histograms illustrating how many topics appear in a passage using each of these thresholds are found in Figure A2.

We also present the frequency with which a topic appears in the top three matches for any passage using our dictionary-based method in Figure 2. This chart indicates that many passages relate to nature sports (for example, hiking, climbing, fishing); arts and crafts; animals; plants, gardening, and yards; music; and water sports (for example, swimming, boating). Last, we validate our topic-salience scores with human coders to manually label the reading passages, which demonstrates that our dictionary-based scores are capturing the intended variation (see Appendix Section C). [12]

4 Topic-relatability estimation and results

4.1 Constructing topic relatability

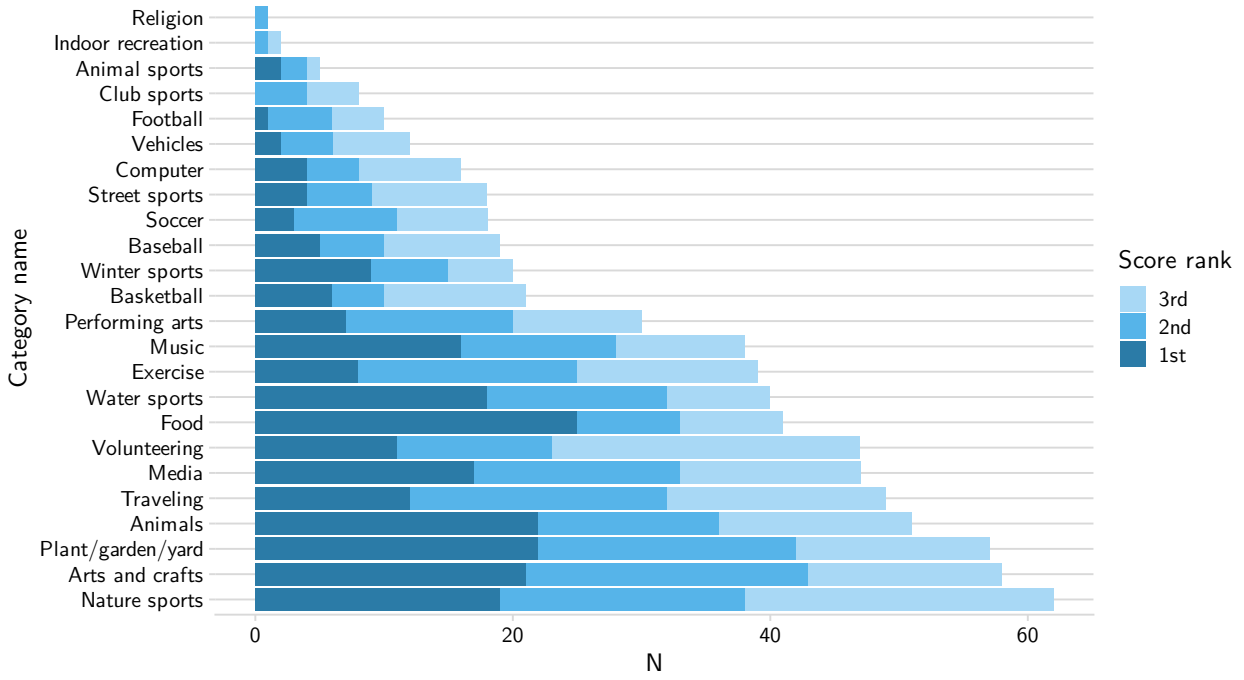
We construct a measure of topic relatability by joining the topic-exposure measure $e_{d,t}$ described in Section 3.2 and the topic-salience measure $m_{t,p}$ described in Section 3.3. We define relatability for demographic group d—defined over either race or gender—and passage p as

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

where $e_{d,t}$ is the topic exposure of demographic group d to topic t and $m_{t,p}$ is the topic salience of topic t in passage p. This definition assumes the following: (a) the increased salience of topic t_j operates through exposure to t_j and not through other topics t_k for $k \neq j$; and (b) exposure weights to topic t operate homogeneously for all demographic groups d and

[12] As part of this exercise, the human coders classified each passage's most salient topic as high-, medium-, or low-salience. The average passage classified as high-salience has $m_{t,p} = 0.046$. The same values for medium- and low-salience passages are 0.023 and 0.016, respectively. (These values are visualized in Figure C1.) This may help us to interpret the "salience threshold" of 0.02 as indicating topics that a human coder may consider high- or medium- salience. Hence, the right panel of Figure A2 depicts the number of medium- or high-salience topics per passage.

Figure 2: Most frequent topics in STAAR reading passages

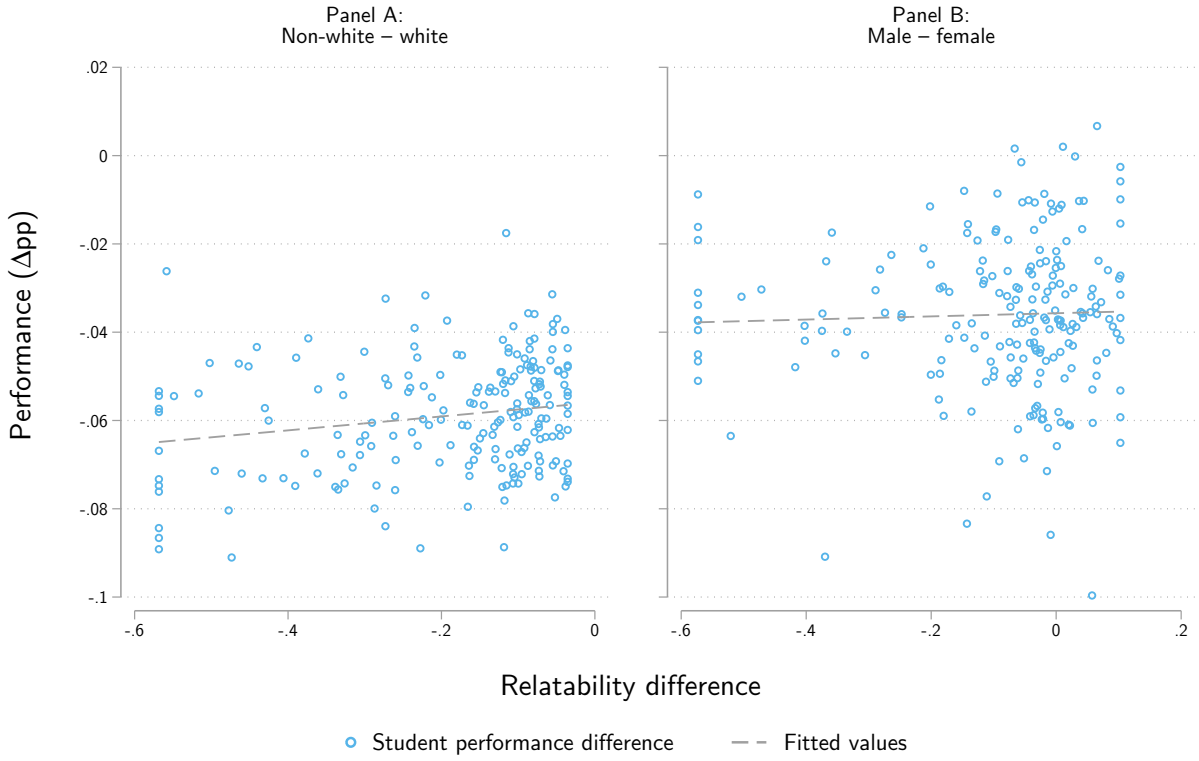


Notes: Reading passage measures are calculated for each topic-passage pair by the term-frequency metric ($m_{t,p}$) discussed in Section 3.3. This histogram shows how frequently each topic is detected in a passages' top three topics. The sample of passages includes grade 3 to 8 STAAR reading-comprehension tests from 2013 to 2019.

passages p. Intuitively, $r_{d,p}$ is a sum of topic salience for passage p weighted by group d's topic exposure. We standardize $r_{d,p}$ to unit variance across the estimation sample at the student-item level.

To show preliminary evidence that our relatability measure predicts test-performance differences between demographic groups, we plot exam outcomes against passage-level relatability differentials for non-White versus White students and male versus female students in Figure 3. Panel A shows a positive relationship between race-based topic-relatability differences and race-based test-outcome differences, which suggests topic relatability affects racial test gaps. However, we observe a less clear relationship between relatability and male–female test-score differences in Panel B. We proceed to a formal test of the relationship between topic relatability and test performance.

Figure 3: Demographic-level test-outcome differences and relatability



Notes: Each graph plots average 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. Panel A plots differences between non-White and White students while Panel B plots male–female student differences. Observations are at the passage level. Differential relatability is winsorized at the 5th and 95th percentiles. The estimation sample includes all students in grades 3 to 8 from 2013 to 2019 taking the standard STAAR reading-comprehension test.

4.2 Estimating the causal effect of topic relatability

To understand the underlying data-generating process and how it produces our identifying variation, consider a test maker who is responsible for creating a third-grade exam each year that evaluates student competency on a fixed rubric. Because standards for third grade are state mandated, she may need to include a poem, two fiction-prose and two nonfiction-prose passages each year with some fraction of vocabulary and comprehension questions. While she assembles five passages designed to appeal to third graders, each year there are some residual differences in the relatability of the topics across students, which differ by race and gender. Thus, while the topic-exposure levels of students are nonrandom, the topic-salience measures within a grade have an element of randomness across exam years within

a grade, as do passages within an exam. We then isolate this residual variation in our regression specification by selecting fixed effects that remove the mean expected relatability score for each demographic group at each grade level. As noted earlier, this approach follows a shift-share empirical strategy, in which we leverage quasi-random variation in shifts (topic salience) conditional on nonrandom and predetermined shares (topic exposure).

Formally, our estimation strategy relies on the conditional randomness of topic salience to identify the causal effect of content relatability on student performance. To illustrate this, consider a simple regression of test performance on relatability,

$$Y_{dp} = \alpha + \beta r_{dp} + \nu_{dp}, \tag{3}$$

where d indexes demographic groups and p indexes passages. Y_{dp} is the average share of questions answered correctly by students in demographic group d when faced with passage p. r_{dp} is topic relatability, defined as the interaction between topic exposure for demographic groups, e_{dt}, and topic salience in each passage, m_{tp}, as in equation (2). β is the coefficient of interest: the impact of demographic-based topic relatability on demographic-group-level student performance.

The non-randomness of r_{dp} complicates a causal interpretation of β in equation (3). Topic exposure, a component of topic relatability, is pre-determined and non-varying for each demographic group. Furthermore, it may be correlated with average achievement $\bar{\theta}_d$ across demographic groups, a key confounder. Thus, we can only rely on variation which comes from differences in topic salience across passages: the test makers selecting different topics for different passages. Still, it is difficult to argue that this variation in topics is totally random. We identify two key challenges. One, the test maker likely has institutional (or personal) preferences over topics, particularly across grades, $g(p)$. For example, the test maker may tend toward including topics in passages that are generally more relatable to Texans. The test maker may also include different topics across grades, believing certain topics, such as animals/pets, are more engaging to younger versus older students. Two, the test maker has institutional requirements to include different passage attributes in an exam, but passage attributes are likely correlated with the amount of topic-related text in an exam. Informational text may contain less topic-related content than literary text. Longer passages may contain more topic-related content than shorter passages.

We therefore leverage only the variation in topic relatability which is arguably random. Formally, we re-cast topic relatability r_{dp} as a "shift-share" variable and employ a shift-share estimation framework to directly estimate the causal effect of relatability (Borusyak, Hull,

and Jaravel 2022).[13] Under this framework, all students in a given grade-year face a series of common shocks through the topics contained in the exam passages (topic salience, m_{tp}). However, the effect of this shock is different across demographic groups, as each group has a different exposure share to each topic (topic exposure, e_{dt}). We argue that while the topic-exposure *shares* are nonrandom, the topic-salience *shifts* across passages are conditionally quasi-random. Specifically, as described above, we must account for the tendency for certain topics to have higher topic salience across grades and the tendency for certain passages to have higher overall topic salience than others.

We highlight the potential sources of endogeneity if we do not purge nonrandom variation in topic relatability. First, topic exposure may be correlated with achievement. Suppose students of some group d_1 have both higher exposure to topics and higher reading ability compared to students of other groups. Reading ability is then a confounder for the relationship between topic relatability and exam performance. We can purge this variation in topic relatability by including demographic-group-level controls. Second, we consider the role of differential topic salience across topic-grades. Suppose the test maker selects passages with topic t_1 more frequently at lower grades and demographic group d_1 has higher exposure to topic t_1 compared to other groups. Thus, group d_1 has higher topic relatability on average at lower grades compared to other groups. If group d_1 also has relatively higher reading ability at lower grades, then ability is again a confounding variable. We can simply include demographic-group-by-grade controls to account for this.

Finally, we must address nonrandom topic salience across passages. Let us suppose passages with attribute A (for example, fiction, difficult vocabulary, lengthy passage) tend to also have more topical text. We previously supposed group d_1 has higher exposure to topics overall, so it follows that group d_1 has higher topic relatability to passages with attribute A. Yet again, if group d_1 is relatively better at reading passages with attribute A, this confounds the relationship between relatability and test performance. Addressing this potential confounder requires a particular form of control. We must include passage attribute controls that are interacted with a measure of demographic groups' exposure to topics overall. In practice, we allow for the possibility of both observed and unobserved passage attributes to be correlated with topic salience by simply interacting passage fixed effects with demographic exposure intensity $E_d = \sum_t e_{dt}$.

[13] While analyses that have units with a vector of differential shares and a vector of common shifts typically feature the resulting exposure-weighted average of shocks serving as an instrument in an IV/2SLS analysis, Borusyak, Hull, and Jaravel (2022) note that the identification assumptions still apply if such objects are used in reduced-form analysis as is the case in our setting.

We formally estimate the following regression specification:

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

Here, $g(p)$ indexes the grade level of passage p and β is the coefficient of interest. $\delta_{d,g(p)}$ are demographic-group-by-grade fixed effects, included to address potential endogeneity stemming from nonrandom topic exposure across demographic groups and nonrandom topic salience across topic-grades. $\pi_p E_d$ are a full set of passage-specific slopes on E_d, included to address potential endogeneity stemming from nonrandom topic salience across passages. Since we have aggregated Y_{ip} to the group-passage level, we estimate the equation with weights representing the number of students and test items that make up (d, p).

We obtain exposure-robust standard errors for equation (4) by estimating an analogous topic-passage-level regression, following Borusyak, Hull, and Jaravel (2022). Further, since our identifying variation comes from the presence of topics in a passage, we must allow clustering of standard errors within a passage. In practice, we more flexibly allow clustering of standard errors within an exam. This approach is motivated by two factors: (a) Passages are the smallest unit for which the exogenous treatment varies, and (b) passages are expected to be correlated with other passages within an exam (for example, test makers are unlikely to include three poems in a five-passage exam). Further details on the exact procedure used to estimate the topic-passage-level regression and obtain standard errors are available in Appendix Section D.

The interpretation of β in equation (4) as the causal estimate of demographic-based topic relatability r_{dp} on test performance requires three assumptions. First, topic-salience shocks are quasi-random, conditional on topic-exposure shares and controls. Second, the shocks are numerous and conditionally uncorrelated with one another. The final assumption relates to construct validity: Topic-exposure differences across demographic groups d, such as race, are actually attributable to differences across d. We discuss each assumption separately.

Conditional quasi-random shock assignment. Assuming the topic-salience shocks are quasi-random is akin to an assumption of orthogonality between relatability and the unobserved error term after including demographic-group- and passage-based controls. Such an assumption may be violated in three ways. First, test makers might incorporate differential performance across groups when designing a test for a given year. If test makers observe large differences in test performance across groups, they may decide to change the way tests are administered or what is asked in the tests to adjust for this in subsequent years. More generally, test makers may feel that the observed test scores for a particular year does not

accurately reflect ability, meaning tests should be adjusted in the future. If this resulting adjustment jointly changes the topics that appear in the passages and non-topic-related attributes of passages (for example, passage difficulty or the types of questions asked), then our main estimates may be biased. Second, test makers may adjust the distribution of topics in response to changing underlying demographics of test takers. For instance, a projected increase in the number of Hispanic students in Texas may simultaneously 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 that also affect outcomes.

We assess the plausibility of this assumption through a falsification test regressing potential confounders on our relatability measure for race using equation (4). We define these confounders consistent with the three violations discussed above. To test for test-maker responsiveness to prior-year performance, we compute one-year lagged average performance for each group-passage dyad dp. Lagged performance is calculated by fixing grade or cohort.[14] This allows for flexibility in how test makers may respond to observing differential performance on a particular exam: They may seek to correct (or enhance) these differences in the same-grade exam next year, or they may correct the exam for the affected cohort of students next year. For each type of lagged performance, we compute the lag in two ways. In the first way, we aggregate the prior year's lagged performance to the exam level and apply it to all passages in the index year. In the second way, we match the prior year's lagged performance for a particular passage position to the sample passage in the index year (for example, the performance on passage #2 in exam $t-1$ would be matched to the relatability of passage #2 in the exam t). Next, we consider as a confounder the share of test takers for passage p that identify as group d. Finally, we consider passage characteristics as confounders: calendar year, passage position, word count, and passage category (literary text or informational text).

Table A4 shows the results of our falsification tests using these potential confounders. We fail to reject the null hypotheses that relatability is uncorrelated with 10 of 11 covariates at the 99% confidence level. The coefficient estimates of the lagged performance confounders are substantially smaller than that of our main regression, and we fail to reject the null hypothesis that the coefficient on relatability is equal to zero. Similarly, the student-population characteristics and most passage characteristics are not predicted by relatability. We do find that relatability is positively associated with the literary-passage indicator, and the joint

[14]For example, consider Black students taking the sixth-grade exam in 2017. Lagged performance *within grade* would correspond to performance of Black students taking the sixth-grade exam in 2016. Lagged performance *within cohort* would correspond to performance of Black students taking the fifth-grade in 2016.

F-test rejects the null that all covariates are jointly uncorrelated with relatability. At the same time, we fail to reject no-association for 10 of the 11 covariates at the 1% level, and a predicted-outcome regression using these covariates yields a coefficient on relatability that is statistically indistinguishable from zero. We also test sensitivity to the inclusion of passage-category fixed effects in a later analysis and find that the coefficients are largely stable.

Many uncorrelated shock residuals. To test for having too few salience shocks or highly concentrated exposure shares, we summarize the distribution of topic salience and aggregated topic exposure, both at the topic-passage level. Appendix Table A5 reports these summary statistics. Column (1) displays statistics without controls, and column (2) displays the distribution with our specified controls. We find significant remaining variation in topic salience even after including our specified controls. Next, we consider an inverse Herfindahl index (HHI) of aggregated exposure shares. This metric serves as a measure of effective sample size for our estimation sample. A low value would indicate a high degree of concentration in exposure shares to a handful of topics. We find an effective sample size of 925 across topics and passages.

As discussed previously, the main source of correlation across topic-passage shocks is within passages. Passages contain a finite number of words, so the presence of one topic may crowd out another topic. Topics that are thematically similar can also be positively correlated within passages. Beyond the passage level, since passages are put together to build an exam, we might also be wary of the potential for within-exam correlation of topics. We take a conservative approach and allow for clustering within exams, but our results are similar with either clustering approach.

Exposure shares attributable to group-level differences. A remaining concern is that the group-level topic-exposure measures are not actually attributable to exposure differences across groups $d \in \mathcal{D}$ but to a different set of groups $c \in \mathcal{C}$. Take racial groups as an example. In our setting and more broadly, there is a close correlation between race and a number of factors associated with both educational outcomes *and* topic interest/familiarity. For example, we observe wide disparities in aspects of socioeconomic status (SES) (for example, income, wealth, and employment) across racial groups. Across race, individuals differ in educational attainment and neighborhoods. While these individual and household characteristics have direct and indirect effects on students' educational outcomes, they also plausibly divide households by the topics they are familiar with and interested in. Exposure differences across racial groups could then be driven by these factors and misattributed to race.

Misattribution of $\mathcal{C}$-level differences to $\mathcal{D}$-level differences affects our estimates in two

ways. Differences across c may correlate with both exposure e_{dt} and test performance Y_{dp}. This confound is addressed by the assumption of quasi-random shock assignment conditional on race-grade fixed effects and passage-by-exposure-sum fixed effects. For example, if high-SES students are exposed to some topic t_0, and topic t_0 is more salient across passages, our estimation with race fixed effects accounts for this, as described earlier. If high-SES students have higher exposure to all topics, the passage-by-exposure-sum fixed effects account for this. Still, we later perform SES- and school-level heterogeneity analysis and show that our main estimates remain largely unchanged when accounting for these factors. Misattribution also affects the interpretation of our estimates. If racial topic relatability is closely linked to SES topic relatability, then our findings may be about SES differences and not racial differences. We find qualitatively that this is not likely to be the case. To demonstrate, we repeat the exercise in Figure 1 but replace race and gender differences in topic exposure with SES differences in topic exposure for Black and White respondents in the ATUS. Individuals with high SES are defined as those that are above the 200% poverty line.[15] Figure A4 displays these exposure ratios. Panel A orders the topic by SES exposure ratios for White individuals. We see some alignment between SES differences across race. The key comparison is between that panel and Panel B, in which we display the same exposure ratios but order them by the Black–White exposure ratio. The SES differences seem to be only weakly correlated with Black–White differences. However, we argue that even if our race- or gender-based topic-relatability measures are closely linked to topic relatability for other groups, our results can still meaningfully represent the disparate impact of test-passage choices across race and gender.

4.3 Results for race-based topic relatability

Our baseline results from regressing test outcomes on relatability at the race level are shown in Table 3. 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 topic relatability increases test performance. In our preferred specification, a 1 SD increase in relatability for a passage leads to a 1.87 pp 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 SD changes in student performance and find that our main effect is equivalent to a 0.08 SD increase in passage-level performance. To standardize the results at the exam level, we take into consideration the fact that variation in relatability is larger across passages than across exams. We find that a 1 SD increase in

[15] Details on how SES is determined in the ATUS sample can be found in Appendix B.

Table 3: Impact of race-based topic relatability on test performance

	(1)	(2)	(3)	(4)
Race topic relatability	0.0199***	0.0187***	0.0188***	0.0176***
	(0.0055)	(0.0058)	(0.0047)	(0.0049)
Race FE	✓		✓	
Race-Grade FE		✓		✓
Passage FE			✓	✓
Passage-by-Exposure-Sum FE	✓	✓		
N of topic-passages	4,920	4,920	4,920	4,920
N of student-passages	64,352,860	64,352,860	64,352,860	64,352,860

Notes: Each specification is a regression of the share of items answered correctly on a passage on race-based topic relatability. Standard errors reported in parentheses below the coefficient estimates are (a) obtained using a shock-level regression following Borusyak et al. (2022) and (b) clustered by exam. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$. For columns (1) and (2), race-based topic relatability is calculated as described in Section 4.1. For columns (3) and (4), race-based topic relatability is calculated after normalizing topic exposure to sum to one for each racial group. The estimation sample includes students in grades 3 to 8 from 2013 to 2019 taking the standard STAAR reading-comprehension test.

average relatability at the exam level predicts a 0.05 SD increase in exam-level student performance. Table 3 also reports the results of estimating the two-way fixed-effects model most analogous to our preferred specification. To estimate it, we recalculate race-based topic relatability, normalizing the topic exposures for a demographic group to sum to one. This purges all variation in race-based topic relatability that is driven by differences in aggregate levels of topic exposure across demographic groups. We find slightly smaller coefficients, but the differences are not statistically significant.[16]

4.3.1 Test-score gaps

While our main specification returns a coefficient that represents the average impact of relatability on test performance, it does not directly illuminate how relatability may contribute to observed racial test-score gaps. We observe large average test-score differences between non-White and White students on reading-comprehension tests. Table A6 displays

[16]While there are appealing estimation properties of directly employing a commonly used difference-in-differences estimation strategy to our original topic-relatability measure, the identification assumptions necessary for two-way fixed-effects estimation are not sensible in our setting. Crucially, treatment in our setting does not turn "on" and "off" consistently for any unit or passage (time) since r_{dp} is a passage-varying mixture of shocks. Accordingly, we also do not observe a unit that experiences a consistent level of relatability that might serve as part of an appropriate comparison group. Further, as discussed earlier, including only passage fixed effects would allow the coefficient estimate of r_{dp} to be confounded by differences in exposure sums across race.

the Black–White and Hispanic–White test gaps by grade, calculated as the percentage point difference in the percentage of items answered correctly. We find that Black students in the third grade have a 13.6 pp lower share of questions answered correctly than White students, a difference that shrinks to 11.7 pp by grade 8. The difference for Hispanic students is slightly smaller: between 9 and 11 pp across grades. These gaps are economically meaningful, equating to 16%–19% and 13%–15% of the White student mean for Black students and Hispanic students, respectively.

We calculate what the change in test-scores would be if topic relatability were equalized for all students by predicting $\widehat{Y}_{dp}$ from equation (4) and calculating a counterfactual $\tilde{Y}_{dp}$ after setting $r_{dp} = 0$. These variables represent 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_0 and d_1 we can then compute the share of average test-score gaps explained by relatability:

$$1 - \frac{\tilde{\mu}_{d_0} - \tilde{\mu}_{d_1}}{\hat{\mu}_{d_0} - \hat{\mu}_{d_1}}. \tag{5}$$

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 the lower average relatability of Black and Hispanic students compared to White students, racial test gaps may be smaller than can be detected using raw student performance. Topic 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 third grade the Black–White and Hispanic–White shares of test gaps explained are 5% and 6%, respectively, while in eighth grade it falls to 2%.

The implications of these findings depend on the source of variation in relatability. If the primary driver of relatability differences is 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, then part of the issue is exacerbated by the selection of topics that are more relatable to one group than the other. We discuss these issues further in Section 7.2.

4.3.2 Heterogeneity

Our main estimates mask potential heterogeneity in topic relatability's effect within racial groups. Certain characteristics or environments may mitigate or exacerbate the effects of

relatability. For instance, very high-achieving and very low-achieving students may be unaffected by relatability, as the effect of reading-comprehension ability may be greater than the effect of relatability. Race-based topic relatability may also be correlated with household income, meaning that accounting for household income may reduce our estimated coefficient on race-based topic relatability. Students could also mitigate the effect of topic relatability through greater exposure to diverse peers at school. While their interests may be driven by household and cultural factors, exposure to peers of other racial groups may allow them to relate to a wider variety of topics. We consider heterogeneity along each of these dimensions.

Achievement. We explore how student achievement mediates the effect of topic relatability on test performance. Since our main outcome measure is reading-comprehension test scores, we seek alternative proxies of reading-comprehension ability.

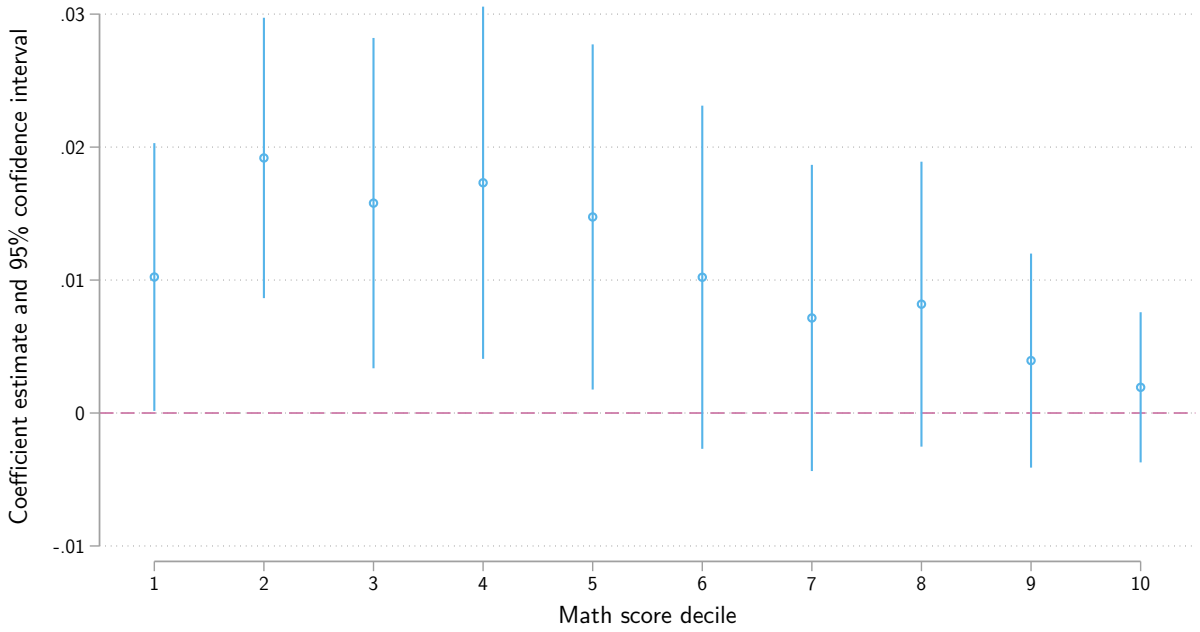
We consider two proxies for ability. Our preferred measure is a student's math score, taken from the same year as the reading-comprehension test. While math and reading-comprehension exams do not test for the same skills, they are highly correlated in practice. We also consider a student's reading score from the prior year (if available). We prefer the math-score proxy, as there is a full sample across grades—since there are no prior-year exams for third graders—and it does not exclude students who are held back in school. We use within-exam deciles of both proxies in the analysis.

Figure 4 plots the coefficients of estimating equation (4) within each decile of math score. We find that the coefficient on topic relatability is positive across deciles, but the effects are highest for, and statistically significant only for, students in the bottom half of the achievement distribution. The point estimate at decile 2 is more than double that of the estimate at decile 7, and this difference is statistically significant, using a stacked and pooled specification. The differences in point estimates are also statistically significant for deciles 1 and 2, suggesting attenuation at both ends of the achievement spectrum. We find a similar pattern using past reading score as the achievement proxy, displayed in Figure A3.

The achievement-heterogeneity results are consistent with the interpretation that students with higher levels of reading comprehension or test-taking ability are better able to offset the influence of topic relatability on performance. This pattern can be particularly pronounced because test performance is constructed using binary item responses: If students consistently answer nearly all items correctly (or incorrectly), the measure exhibits ceiling and floor effects, obscuring additional variation in performance at the tails of the ability distribution. For such high-achievement students, it may be appropriate to test the impact of relatability using exams with more difficult passages and harder questions.

These results also address potential concerns that differences in achievement across racial

Figure 4: Heterogeneity of race-based topic relatability effects by math achievement



Notes: Each coefficient estimate is from a separate specification that regresses the share of items answered correctly on a passage on race-based topic relatability. Unreported controls include (a) race-grade fixed effects and (b) exposure-sum-by-passage fixed effects. Each specification is run on a separate student subsample based on the decile of a student's math test score. Each student's math test score comes from the standard STAAR exam for mathematics. Score deciles are formed within exam. Exposure-robust standard errors clustered by exam are used to construct the 95% confidence intervals. Observations are at the race-passage level and are weighted by the number of student-items. The estimation sample includes all students in grades 3 to 8 from 2013 to 2019 taking the standard STAAR reading-comprehension test.

groups drive our main topic-relatability findings. Ability may be directly or indirectly associated with exposure to certain topics. If high-ability students perform better on passages with such topics, it may be because of relatability or their reading-comprehension ability. Given baseline test-score differences across racial groups, this raises a concern that relatability is correlated with ability. However, our results demonstrate that topic relatability still has a positive effect when comparing students with similar levels of predicted achievement.

Socioeconomic status. Race and SES are highly correlated in our student population. Over 56% of students in our sample are economically disadvantaged, defined as receiving free- or reduced-price lunch or being designated by the school to be experiencing economic disadvantage. Rates of economic disadvantage are much higher for Black (73%) and Hispanic students (75%) than Asian (29%) and White students (28%)

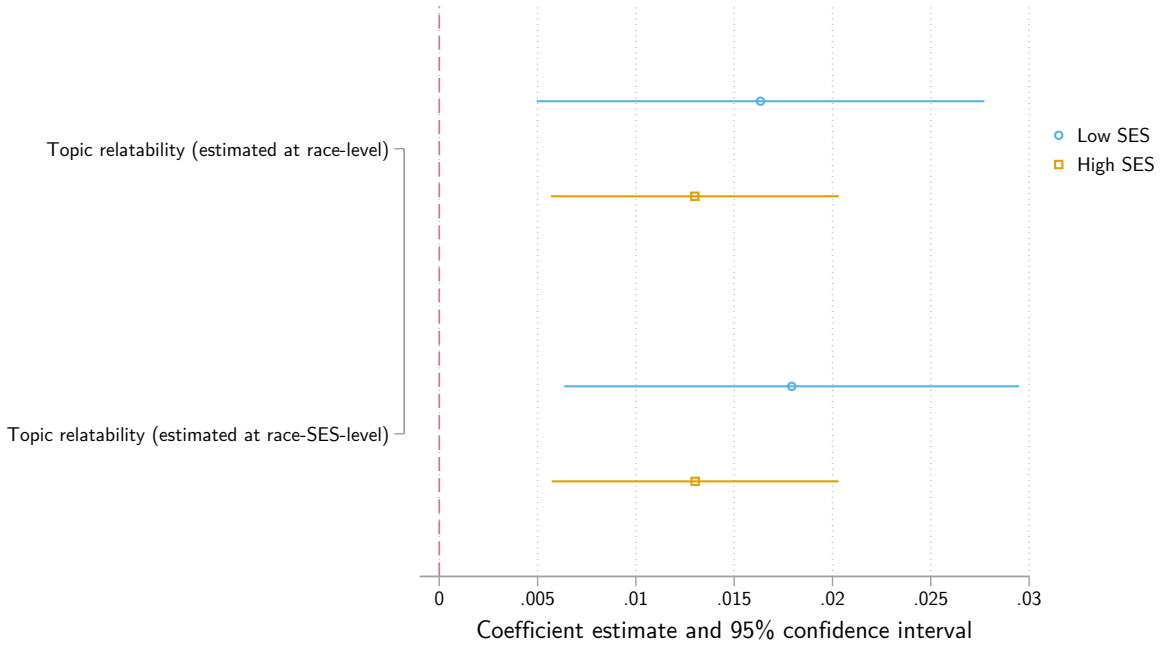
Given this relationship and the correlation between SES and school performance, we explore the extent to which topic-relatability responses differ by SES. We classify students as low or high SES based on their status in the administrative data as economically disadvantaged or not. Next, we find an appropriate adult sample group for this population in the ATUS data. We classify ATUS respondents as low or high SES based on whether their household income is below or above the 200% poverty line. We then estimate two versions of our main specification. First, we estimate the within-SES impact of race-based topic relatability—that is, we estimate equation (4) separately for low- and high-SES students using race-based topic relatability. Second, we repeat this exercise but replace race-based topic relatability with race-SES-based topic relatability—that is, using race-SES instead of race for demographic group d. Details on SES classification and the construction of race-SES topic-exposure measures can be found in Appendix Section B.

Figure 5 displays the coefficient estimates for topic relatability by race and by relatability estimation method. Focusing on the first set of coefficients, we can see that the effect of race-based relatability is slightly lower for high-SES students than for low-SES students, but we fail to reject the null hypothesis that the coefficient estimates are the same. The second set of coefficients show that shifting from race-based topic relatability to race-SES-based topic relatability does not significantly change the estimates.

These results show that race-based topic relatability matters for high- and low-SES students alike. We find suggestive evidence that the effect is smaller for high-SES-status students; we cannot rule out the possibility that observing SES at more granular levels would firmly establish such a trend. Such a relationship may be possible because high-SES students and families have more homogeneous interests and familiarity with topics or because of higher exposure to individuals outside their racial group. It may also reflect heterogeneity in the relatability effect across the achievement distribution. Since high-SES students likely perform better on tests, there is less scope for improvement on test performance. Also, a smaller race-based relatability effect for high-SES students would be inconsistent with the concern that high-SES students' exposure to certain topics are confounding our main estimate. If this were the case, we should observe the opposite effect: Our race-based relatability estimates should be higher for high-SES students.

Schools and peers. Peer effects may weaken or intensify the impact of race-based topic relatability on test outcomes. We have only considered a coarse topic-relatability measure that is constructed at the race level. Our approach is born out of data limitations: We are unable to directly observe the interest and familiarity of students in our sample. However, even from this crude starting point, we can make inferences about how the effect of topic

Figure 5: Heterogeneity of race-based topic-relatability effects by SES



Notes: Each coefficient estimate is from a separate specification that regresses the share of items answered correctly on a passage on topic relatability. Unreported controls include (a) race-grade fixed effects and (b) exposure-sum-by-passage fixed effects. Specifications differ by sample and the way topic relatability is calculated. The top set of coefficients is estimated using the baseline topic-relatability measure described in Section 4.1 at the race level. The bottom set of coefficients is estimated using race-by-SES topic relatability. The blue, circle coefficients are estimated for the low-SES student population, and the green, square coefficients are estimated for the high-SES student population. *Low SES* is defined as students who are enrolled in a free lunch program, a reduced-price lunch program, or some other income-based program. Exposure-robust standard errors clustered by exam are used to construct the 95% confidence intervals. Observations are at the race-passage level and are weighted by the number of student-items. The estimation sample includes all students in grades 3 to 8 from 2013 to 2019 taking the standard STAAR reading-comprehension test.

relatability may differ within race. Students who live, play, and study in more racially homogeneous settings are likely to have interests and familiarity, and thus topic relatability, similar to others of their race. Conversely, students living in racially diverse settings interact more with peers not like themselves. Even if their interests are not aligned with their school environment, this exposure to diverse peers may make them more familiar with a broad range of topics.

We proxy peer composition using school attended. First, we reestimate equation (4) but saturate the fixed effects with a school indicator to estimate the topic-relatability effect from only variation from students within schools. In Table A8, column (2) shows that the estimated impact of relatability is comparable to the main estimate. Thus, we show that

topic relatability affects test performance even when comparing students within schools.

We study heterogeneity along two peer-diversity measures. First, we compare estimates between more and less racially integrated schools. To calculate school integration for a given school, we use an HHI based on the school's population share of our four race categories, in which a higher value indicates more racial homogeneity. We estimate our main specification separately for each quartile of this index. Table A9 shows the results of the HHI analysis. The coefficient estimates are statistically indistinguishable from one another. These results are not particularly surprising, given that a school's HHI masks significant heterogeneity in the experiences of each racial group. For example, if a school's student body consists of 80% group A and 10% groups B and C, students of groups B and C in this school have a highly minoritized experience with high exposure to race A, but race A has a majority experience with low exposure to any other race.

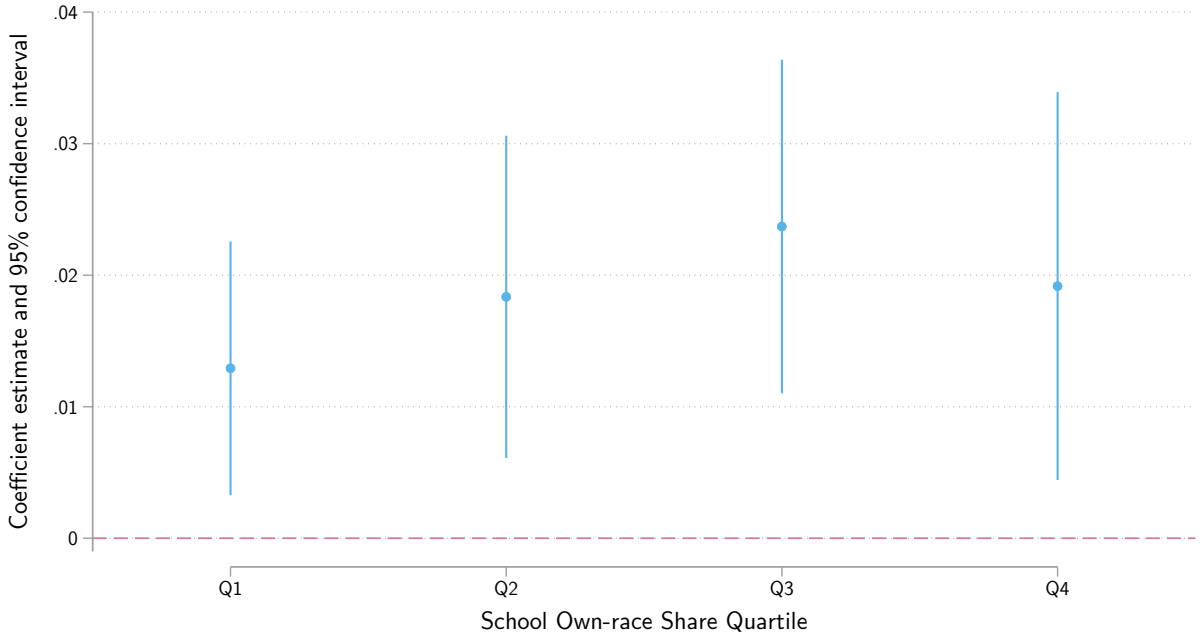
Second, we conduct an analysis that separates the student population into subgroups based on own-race share at a school. To do this, first we calculate for each student the share of their school that is their own race. Then, within each race group in the population, we divide students into quartiles relative to the distribution of this own-race share. Figure 6 shows the results of the school own-race-share analysis. Students who are surrounded by more students of their race—indicated by a higher quartile—have stronger relatability effects. The point estimate on topic relatability is 45% smaller for quartile 1 compared to quartile 3. The coefficient differences across quartiles 1, 2, and 3 are statistically significant at the 5% level or more. This result suggests higher exposure to individuals outside one's race moderates the impact of topic relatability.

4.3.3 Sensitivity and robustness

We test whether our results are sensitive to alternative specifications and alternative formulations of our main relatability measure. First, we rerun our main specification replacing race-grade fixed effects with alternative fixed effects. Second, we change our calculation of topic exposure e_{dt} by using subsamples of the ATUS data and by using time-diary data differently. Third, we employ different NLP methods and measures to calculate topic salience m_{tp} and estimate our main specification with the resulting relatability measure. Finally, we drop each topic when calculating the topic-relatability measure to test sensitivity to our construction of the topic set. We find that our main result is robust throughout this battery of exercises. Details on the methodology and results of the robustness analysis are in Appendix Section E.

One remaining concern is the extent to which adult time-use data is a good proxy for children's interests and familiarity. We show through topic-exposure sensitivity exercises in

Figure 6: Heterogeneity of race-based topic-relatability effects by school own-race share



Notes: Each coefficient is an estimate from a quartile of the population using our baseline specification regressing share of items answered correctly on a passage on race-based topic relatability. We calculate for each student the share of their school that matches their race. Then, for each race, we split the population into quartiles. Thus, quartile 1 is a minority in their school and higher quartiles are relative majorities. Unreported controls include (a) unit fixed effects at the level indicated in the legend and (b) exposure-sum-by-passage fixed effects. Exposure-robust standard errors clustered by exam are used to construct the 95% confidence intervals. Observations are at the race-passage level and are weighted by the number of student-items. The estimation sample includes all students in grades 3 to 8 from 2013 to 2019 taking the standard STAAR reading-comprehension test.

Appendix Section E that using the age 15–34 subsample of the ATUS produces a slightly higher coefficient estimate on race-based topic relatability, but this estimate is statistically indistinguishable from the main estimate. We take this analysis one step further and directly compare our estimates to estimates based on children's time-use data.

We use data from the Child Development Supplement of the Panel Study of Income Dynamics (PSID-CDS), which includes relatively detailed time-use data on children aged 0–17. The CDS surveys caregivers of children and older children of adults who are surveyed in the PSID, a panel survey of individuals and families in the United States since 1968. The CDS crucially includes a weekday and weekend time diary of each surveyed child. We use data from the 2014 and 2019 waves of the PSID-CDS, for children aged 8 to 17. Our final PSID sample includes 1,555 children. More details on the PSID-CDS data and our PSID

empirical methodology are in Appendix Section B.

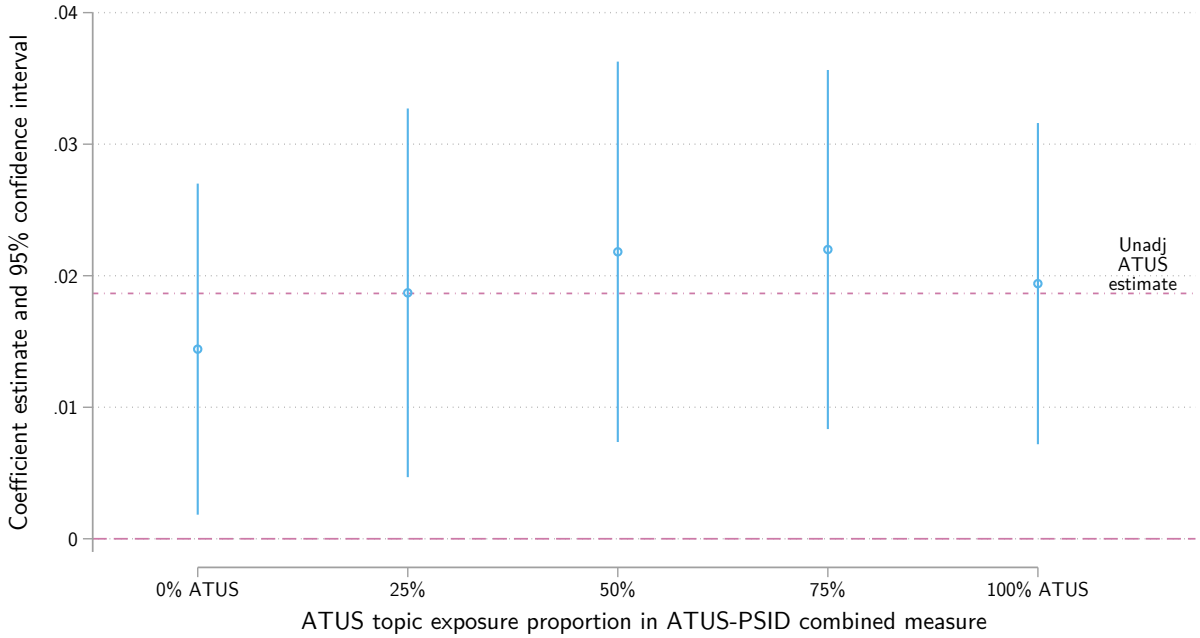
We use the PSID-CDS time diaries to construct a topic-relatability measure analogous to our ATUS-based topic-relatability measure, and we estimate its impact on test performance using our main specification. Table A7 displays the results for PSID- and ATUS-based topic relatability. We show the standardized topic-relatability point estimates as before, but we also show the effect sizes scaled by the Black–White and Hispanic–White topic-relatability gaps. Columns (1) and (2) show that the effect of a 1 SD increase in relatability is substantively different based on the underlying survey used: The ATUS-based effect is almost two times greater than the PSID effect. However, this is partially a result of the small sample sizes among certain groups in the PSID. Columns (3) and (4) show the same regressions estimated without Asian students, which tightens the difference in coefficients across the two measures.[17] The scaled coefficient estimates corroborate this story. Across columns (1) and (2) and across columns (3) and (4), we fail to reject the null hypothesis that the Black–White scaled effects of the relatability estimates are the same.

We test the extent to which different mixes of the ATUS and PSID produce different coefficient estimates. Given the concerns with small sample sizes in the PSID, we apply an empirical Bayes shrinkage estimator (Morris 1983, Walters 2024) to both the ATUS and PSID topic-exposure shares, shrinking estimates to the population topic-level mean. Column (5) in Table A7 shows the effect of creating a composite relatability measure based on a simple average of the shrunk ATUS and PSID topic exposures. The point estimate is larger than our main ATUS-based estimate, but the difference is not statistically significant at the 5% level. Figure 7 plots coefficient estimates using different weighted averages of the ATUS and PSID measures. There is suggestive evidence that a composite measure may better predict test performance by reducing measurement error from two sources: (a) using adult time-use data to approximate children's interest and familiarity (in the ATUS), and (b) using a survey with limited observations (in the PSID). We find more evidence of the latter source of error, as we find a statistically significant difference between the PSID-only measure and the 25% ATUS and 50% ATUS composite measures when we use a stacked regression and a t-test for equality between the coefficient estimates.

Figure 7 shows two additional results. First, the effect of the shrunk PSID-based relatability measure is much closer to the ATUS measure than the raw PSID-based measure, again suggesting that measurement error due to lower observation counts in certain cells causes attenuation in the initial PSID-based estimates. Second, the effect of the shrunk ATUS measure is almost exactly the same as that of the ATUS measure, suggesting that classical sampling error in the ATUS does not appear to generate noticeable attenuation bias in our

[17] Our PSID sample contains only 15 Asian respondents.

Figure 7: Race-based topic-relatability effect on test performance using different ATUS and PSID combinations



Notes: Each coefficient estimate is from a separate specification that regresses the share of items answered correctly on a passage on race-based topic relatability. Unreported controls include (a) unit fixed effects at the level indicated in the legend and (b) exposure-sum-by-passage fixed effects. Each specification uses a different average of race-based topic exposure based on ATUS and race-based topic exposure based on PSID. Estimates are ordered from left to right based on the weight on the ATUS topic exposure. Race-based topic exposures are adjusted using an empirical Bayes shrinkage estimator, which shrinks race-based topic-exposure measures to the population topic-exposure means. Our baseline coefficient estimate, based on the unadjusted ATUS topic-exposure measure, is indicated by the dot-dash line. Exposure-robust standard errors clustered by exam are used to construct the 95% confidence intervals. Observations are at the race-passage level and are weighted by the number of student-items. The estimation sample includes all students in grades 3 to 8 from 2013 to 2019 taking the standard STAAR reading-comprehension test.

estimates.[18]

[18]We further assess the stability of our main estimates using a split-sample instrumental variables approach. We randomly split the ATUS sample into two halves and construct topic-relatability measures separately for each subsample. These measures can be viewed as noisy but independent proxies for the true underlying relatability measure of interest. In a shift-share estimation framework with noisy proxies, consistency can be achieved by estimating a two-stage-least-squares regression of the outcome on one proxy, using the other proxy as an instrument. The resulting estimates are very similar to our main results, indicating that attenuation bias from sampling error in the ATUS has a negligible impact on our estimates.

Table 4: Impact of gender-based topic relatability on test performance

	(1)	(2)	(3)	(4)
Gender topic relatability	-0.0009	-0.0001	-0.0023	-0.0007
	(0.0057)	(0.0055)	(0.0053)	(0.0052)
Gender FE	✓		✓	
Gender-Grade FE		✓		✓
Passage FE			✓	✓
Passage-by-Exposure-Sum FE	✓	✓		
N of topic-passages	4,920	4,920	4,920	4,920
N of student-passages	64,352,860	64,352,860	64,352,860	64,352,860

Notes: Each specification is a regression of the share of items answered correctly on a passage on gender-based topic relatability. 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. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$. For columns (1) and (2), gender-based topic relatability is calculated as described in Section 4.1. For columns (3) and (4), gender-based topic relatability is calculated after adjusting topic exposure by exposure sums for boys and girls. The estimation sample includes students in grades 3 to 8 from 2013 to 2019 taking the standard STAAR reading-comprehension test.

4.4 Results for gender-based topic relatability

We conduct the same analysis as before, but focus now on gender. In Table 4, across all specifications, we fail to reject a null effect of gender-based topic relatability on student test performance. Even if these point estimates are taken at face value, the share-of-gap-explained calculations are one or two orders of magnitude smaller than the race-gap calculations. This is due to not only the smaller point estimates in magnitude but smaller gender-based topic-relatability differences compared to Black–White or Hispanic–White topic-relatability differences. In addition to our baseline ATUS measure, we generate a topic-relatability measure using the PSID data on children's time use. We still fail to reject a null effect with this alternative time-use data (see Table A10).

Our results point to heterogeneity in the importance of different dimensions of topic relatability for race and gender. In our discussion of the subcomponents of relatability in Section 2.2, we outline that topic relatability bundles two distinct effects that cannot be precisely distinguished in our setting: interest in a topic and familiarity with a topic. This framework is a useful lens for thinking through the different results we observe between gender and race.

Suppose test outcomes are a function of both interest in the passage topic and familiarity with the passage topic. Demographic group differences are then driven by average differences in interest *and* familiarity. Holding personal interest fixed, a student's topic

familiarity should increase when family or community members engage in topic-adjacent activities. However, students' exposure across gender lines is effectively universal, whether in the classroom or—through opposite-gender siblings and parents—in the home. Because of well-documented residential sorting across racial and ethnic lines, many students have much lower exposure to individuals of different races or ethnicities relative to cross-gender exposure. This suggests that cross-gender comparisons of test outcomes on the same passages should be driven more by variation in interest, whereas cross-race comparisons of test outcomes are likely further apart on both interest *and* familiarity.

This argument suggests, all else equal, our estimated coefficient will grow in the face of larger familiarity differences. This is supported by our peer-effects estimates using race-based topic relatability (see Figure 6). That analysis found that students in schools with a higher share of same-race peers exhibited higher relatability estimates. These effects are statistically distinguishable for each of the bottom three quartiles, demonstrating that students who are a relative minority, and thus have more cross-race familiarity, have a point estimate that is 45% smaller than students in a more homogeneous environment.

5 Identity-relatability estimation and results

Students might engage more deeply with passages containing characters that share their gender or racial background. Accordingly, we develop measures of identity relatability to complement our topic-relatability measures. Our method is intuitive: First, identify each character in a reading passage; second, impute gender and race for each character from their name (or, for gender, context clues such as pronouns); third, calculate the predicted gender and racial shares of characters. Below we confirm the robustness of each of these steps. For gender, we also consider a dictionary-based, rather than character-based, measure of a passage's gender congruence, adapted directly from Adukia et al. (2023), which yields very similar results. We measure the test-score impact of identity relatability separately for race and gender and calculate their contribution to race and gender test gaps.

5.1 Data and methodology

Our core identity-relatability measures are based on the share of human characters (in the reading passages) that match a student's identity. Our first approach is to use LLMs for named entity recognition (NER). For each passage, we instruct OpenAI's GPT-5-mini model to read it and return structured output for each character. This output includes the character's first name, last name, and title, as well as the number of times they are mentioned

by name. We verify that all names appear in the text to guard against hallucinations. We also check the LLM-provided count against a string-match count for the first- or last-name, which agree with each other over 96% of the time. We exclude from the analysis the 9% of passages for which the LLM doesn't identify any characters with a first or last name.

Next, we impute demographic characteristics for each character. Inferring gender is most straightforward: We instruct the LLM to infer (binary) gender of characters from pronouns, titles, and other context clues, directing it to leave this field blank if unclear (0% missing). To validate this method, we also use the US Social Security Administration database of baby names by gender to impute gender for characters with first names. For this, we simply calculate the share of all babies born since 1950 with a given name that are (fe)male. Race is also imputed using two name databases. The first contains last names matched to Social Security Administration databases with the same four racial categories we have considered. The second has first-name data from six southern states' voter files. Our baseline race prediction combines these first- and last-name ("FLN") data with a Bayesian prior (Imai & Khanna 2016; Imai et al. 2022; computed with `wru` R library). The prior is the racial share of the US population in 2020. We assume test administrators draw passages from a nationally representative distribution of passages, which makes this prior an appropriate first-order approximation of character identity. Some characters only have first names (22% of characters) or last names (12%). Thus, these characters are assigned to the Bayesian prior when we employ prediction methods that only use last or first names, respectively. Because we do not consider certain racial groups, such as Native Americans, or multiracial individuals, the race predictions we use may sum to less than unity. In Appendix Section B.1, we provide the race classification for characters in an example passage.

Finally, we aggregate to the passage-demographic-group level by averaging across characters. For gender, since the LLM yields universal labeling, it is as simple as taking the mean of binary gender indicators for each character in a passage. All race-based measures are predicted probabilities conditional on first or last name, so we aggregate over these predicted values. Thus, our passage-level score for race d represents the share of passage characters that are d in expectation. For robustness in this aggregation step, we also consider a mean weighted by the number of named mentions. This weighted-mean measure gives more weight to protagonists and less weight to minor characters in a passage.

Our baseline race- and gender-based identity-relatability measures use the simple-mean aggregation method with FLN data and LLM imputation, respectively. We plot the histograms of their distributions in Appendix Figures A5 and A6. For race-based identity relatability, we find that most passages have a substantial share of White-predicted characters. For each non-White racial group, we find that some passages predominantly features

characters from that non-White race. Passages are also more likely to be male skewed than female skewed: An all-male-character passage is 50% more likely than an all-female-character passage. However, a substantial number of passages are balanced or nearly balanced on gender. In contrast with our topic-relatability measures, our identity-relatability measures have a naturally interpretable scale: Moving from zero to one on either gender- or race-based identity relatability represents moving from a passage that contains no characters sharing your identity to one with only characters sharing your identity.

An alternative approach to the NER-centric methodology is a dictionary-based method that is computed directly at the passage level. Adukia et al. (2023) provide a list of gendered terms that may appear in text. We use their most expansive list of gendered terms as well as their most narrow, composed of just pronouns. For each set of words, we detect all gendered terms in a passage and calculate the passage-gender-level metric as the (fe)male word share of all gendered words, as defined by the dictionary.

With these identity-relatability measures in hand, we estimate the impact of identity relatability of test scores using a specification similar to equation (4),

$$Y_{dp} = \delta_{d,g(p)} + \pi_p + \beta^{identity} r_{dp}^{identity} + \nu_{dp}, \tag{6}$$

where $r_{dp}^{identity}$ is the identity relatability for group d and passage p. Since the econometric exercise for identity relatability differs from that for topic relatability, we do not use the shift-share empirical strategy. Most notably, we replace $\pi_p E_d$ from equation (4) with a passage fixed effect π_p. We cluster the standard errors at the exam level (that is, grade-year), as before.

Our choice of controls in the two-way fixed-effects specification in equation (6) follows from the general motivation for the controls in the shift-share specification in equation (4). One concern is that test-maker preferences for including some identity groups across passages may be spuriously correlated with test performance. For example, passages generally feature more White characters, and White students tend to have higher test outcomes. A second concern is that unobserved passage attributes are correlated with both character selection and test performance. The $\delta_{d,g(p)}$ and π_p fixed effects account for both these concerns.

5.2 Results

The results for identity relatability are qualitatively and quantitatively consistent for both race and gender measures, robustly maintaining conventional levels of statistical significance and coefficient stability across choices considered in Section 5.1. The point estimates are larger for race than gender. Further, when factoring in differences in average relatability, the

Table 5: Impact of race-based identity relatability on test performance

Prediction Source	First + Last Names		Last Name		First Name	
	(1)	(2)	(3)	(4)	(5)	(6)
Race identity rel. (mean)	0.0109**		0.0092**		0.0103*	
	(0.0043)		(0.0038)		(0.0053)	
Race identity rel. (w.mean)		0.0112***		0.0098***		0.0100**
		(0.0037)		(0.0033)		(0.0045)
Reading Passage FE	✓	✓	✓	✓	✓	✓
Race-Grade FE	✓	✓	✓	✓	✓	✓
N of student–passages	59,305,495	59,305,495	59,305,495	59,305,495	59,305,495	59,305,495

Notes: Each specification is a regression of the share of items answered correctly on a passage on race-based identity relatability. Standard errors reported in parentheses below coefficient estimates are clustered by exam. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$. Identity relatability is defined as the share of characters matching a student's race. Columns (1) and (2) calculate identity relatability using last names from Social Security data and first names from six southern states. Columns (3) and (4) calculate it using only the last-name data. Columns (5) and (6) use only the first-name data. All of these use a prior distribution matching national demographics in 2020. Simple means give each character a weight of one, while weighted means use the number of named mentions for the weight. The estimation sample includes students in grades 3 to 8 from 2013 to 2019 taking the standard STAAR reading-comprehension test.

results explain a greater share of the test gaps for race than for gender.

Moving from all different race to all same race yields a 1.1 pp improvement in test performance using our preferred race-based identity-relatability measure leveraging both first- and last-name data. This corresponds to a 0.03 SD improvement at the exam level. We show in Table 5 that race-based identity-relatability results are also qualitatively consistent across imputation methods. All of these estimates are statistically indistinguishable in pairwise hypothesis tests from a stacked regression with pooled standard errors. Further, both measures that leverage last-name data span nearly the entire range from 0 to 1 for all races, suggesting there are racially homogeneous passages for all four of our considered racial groups. Average White identity relatability is near 0.69, whereas average Black relatability is 0.12 and average Hispanic relatability 0.08. This suggests that equalizing relatability in passages with named characters would close both the Black–White and Hispanic–White test gaps by about 0.6 pp, or 5% and 7% of the Black–White and Hispanic–White test gaps, respectively.

For our preferred gender measure (using LLM imputation), we find that moving from all opposite-gender to all same-gender characters in a passage results in a 0.76 pp improvement in test performance (see Table 6), which translates to a 0.02 SD increase at the exam level. Average female relatability is near 0.42, and average male relatability is near 0.58, suggesting that equalizing relatability would actually widen the male–female test gap by about 0.1 pp

Table 6: Impact of gender-based identity relatability on test performance

Type	Character				Dictionary	
Prediction Source	LLM		SSN		Adukia	Pronouns
	(1)	(2)	(3)	(4)	(5)	(6)
Gender identity rel. (mean)	0.0076***		0.0064***		0.0076***	0.0073***
	(0.0018)		(0.0019)		(0.0019)	(0.0018)
Gender identity rel. (w.mean)		0.0059***		0.0046**		
		(0.0017)		(0.0017)		
Reading Passage FE	✓	✓	✓	✓	✓	✓
Gender-Grade FE	✓	✓	✓	✓	✓	✓
N of student–passages	59,305,495	59,305,495	57,339,693	57,339,693	57,956,202	55,325,576

Notes: Each specification is a regression of the share of items answered correctly on a passage on gender-based identity relatability. Standard errors reported in parentheses below coefficient estimates are clustered by exam. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$. Identity relatability is defined as the share of characters matching a student's gender. Columns (1) and (2) calculate identity relatability using gender labels from a large language model, imputing them from pronouns and other context clues. Columns (3) and (4) calculate it using first-name data from nationally representative Social Security data. Columns (5) and (6) follow the dictionary-based approach of Adukia et al (2023). Simple means give each character a weight of one, while weighted means use number of named mentions for the weight. The estimation sample includes students in grades 3 to 8 from 2013 to 2019 taking the standard STAAR reading-comprehension test.

(3% of the 3.5 pp female-male gap). Both of these results are consistent across both NER-type and dictionary-type measures of gender-based identity relatability.

Additional robustness results are reported in Appendix Tables A11 and A12. First, we confirm that both our race and gender results hold when using Anthropic's Claude Sonnet 4 model, a leading competitor to OpenAI's GPT-5-mini. For the race prediction, we consider a Bayesian prior corresponding to Texas's 2020 racial shares, which may be desirable if passages are primarily sourced from Texas-specific writing or if the Texas prior better captures students' perceptions of race rather than actual demographics. We find that the race-based identity relatability results are robust to this change. To test whether our results are driven by the functional form of identity relatability, we estimate specifications using identity relatability in levels (the count of same-identity characters) rather than shares, incorporating both character counts and mention-weighted counts. These results suggest that the race-based identity-relatability results are primarily driven by the *share* of own-race characters (as with the main results), whereas the gender-based identity-relatability results are also responsive to the *number* of characters matching a student's gender. We further verify robustness by constructing indicator variables equal to one if a passage's identity relatability exceeds the median for that demographic group; this approach ensures comparability across demographic groups with different underlying distributions of relatability and yields statistically significant results for both race and gender.

We also consider heterogeneous effects based on passage characteristics, such as whether the passage contains historical figures, includes celebrities, includes children, or is a memoir (see Appendix Tables A13 and A14). We classify characters into each category using the LLM (further details are in Appendix B.6). The 24% of passages with at least one historical figure have statistically distinguishable larger effects for gender relatability: about 1.7 pp vs. 0.4 pp for the rest of the passages. A similar but weaker effect is found for individuals tagged by the LLM as plausibly recognizable as celebrities by a grade school student. The results further suggest that gender effects may be muted for memoir passages. Interestingly, the pattern of statistical significance in the heterogeneity results are precisely flipped for race: null results for all the effects described for gender, but a statistically significant reduction in the effect of race-based identity relatability when there are children in the passage. This 40% reduction in effect size may suggest that students find it easier to identify with child characters of any race, but the disconnect grows larger when reading about adults.

6 Combined effect of relatability

Up to this point, we have taken a structured approach to studying the effect of content relatability on test performance. We have distinguished two sources of content relatability: topic relatability and identity relatability. This delineation allows policymakers to more easily compare each measure's impact on test performance, which is especially important since the recent literature has focused on the latter factor. Importantly, these measures are quantified and could be applied to any exam. For topic relatability, we use a dataset that measures topic exposure based on revealed preference (time spent on topic-related activities). For identity relatability, we use character names, for which we can leverage extensive data on identity associations. Our analysis focuses on content attributes orthogonal to targeted reading skills (for example, leisure topics), rather than on constructs like vocabulary difficulty.

However, our structured approach has limitations. Our relatability constructs are not exhaustive because the topics and identities we observe are limited. There are likely attributes of test passages that are not captured by our relatability estimates and cause differential test performance across demographic groups. Our estimated effects are also not necessarily mutually exclusive. Many passages are drawn directly from existing texts, so character identities and passage topics may be tightly correlated.

We address these issues by supplementing our relatability measures with an external survey directly asking the general population how they assess the relatability of different passages. This offers a more flexible, less prescriptive measure than our structured approach

based on time-use data, NLP, and named-entity recognition. Additionally, it serves as a robustness check for these existing measures. Then, we consider the combined impact of all three relatability measures on test scores.

6.1 Survey-based relatability

We survey 1,066 adults residing in the United States. Participants are recruited from Prolific, a widely used online platform to host surveys for social science research. Respondents are selected to be representative of the US population by race, gender, and age range. In addition to these demographic characteristics, we collect information on respondents' state of residence, their parental status, and the frequency with which they interact with children. Further details on participant recruitment, participant characteristics, and survey procedures not described here can be found in Appendix Section F.

Each respondent read 10 passage summaries and assessed children's relatability to each text. For each passage in our estimation sample, we ask an LLM to produce an approximately 220-word summary at an eighth-grade reading level. We rely on these passage summaries rather than full-length text, for both pragmatic and methodological reasons. Passages in our corpus take at least three minutes to read and comprehend, which risks significant drop-off in study participation. Longer texts also restrict how many relatability assessments we can elicit. This is important in a setting in which repeated assessments may improve response quality. Methodologically, summaries also reduce sensitivity to textual features like length or vocabulary level; similarly, genre cues may persist but are dampened. The set of passages and the order in which they are delivered are randomized for each respondent.

We construct a survey-based measure of relatability by aggregating the relatability assessments across respondents. Respondents were provided a general definition of relatability, which includes topic familiarity, content interest, and identity alignment. We elicit race-specific rankings of relatability and likewise (but separately) for gender. For race, respondents must rank the relatability of the text from most relatable to least relatable. For gender, respondents simply answer whether they think the text is more relatable to girls or boys. Our preferred measure of survey-based relatability comes after fitting a rank-ordered logit choice model (an "exploded" logit) by passage, using respondents' rankings. We use the predicted probability that a group ranks first as that passage's relatability for the group.[19]

We estimate regressions of test performance on survey relatability including demographic-group-by-grade fixed effects and passage fixed effects, as in equation (6). Table A15 shows the

[19]Our results are robust to using a simpler measure: the share of respondents who ranked a demographic group first for a passage.

results of these regressions. A one-unit increase in the predicted probability that a passage is most relatable to a racial group is associated with a 1.2 pp increase in test performance. We observe slightly larger coefficients for gender, but the difference is not statistically significant. We scale coefficients by the between-group relatability gaps to obtain gap-relevant effect sizes. We find that the survey-based relatability effect explains less of the racial difference in test scores than topic and identity relatability. We also find that, similarly to identity-relatability effects, estimated survey-based relatability effects would widen, not close, test differences between male and female students.

6.2 Correlation across relatability measures

The interpretation of the estimated results across our three measures depends on the correlation among our three measures. If the topic- and identity-relatability measures are closely related, that may imply that the estimated effects of topic relatability are not necessarily attributable to familiarity or interest but to a correlation with identity. Given that survey-based relatability predicts performance, if it is uncorrelated with topic and identity relatability this would indicate that there are other factors embedded in exams that influence performance differences across demographic groups.

We first consider the relationship among our three measures by regressing one measure on the other measures using a version of equation (6). We standardize each measure by its standard deviation for ease of comparability across estimates. Appendix Tables A16 and A17 show the results of these cross-correlations for race and gender, respectively. We find little correlation between the topic- and identity-relatability measures (columns (1), (3), and (5) in both tables). However, the survey measure explains 10% of the residual variation in the racial-identity measure and 32% in the gender-identity measure (column (4) in both tables). These results yield two insights. First, the presence of relatable topics for a group does not necessarily imply that characters from that group appear in a passage. This finding is consistent with a test-making process that prioritizes including diverse characters across many contexts rather than diversifying the contexts themselves. Second, survey respondents appear to anchor more strongly to identity than to topics when assessing relatability of passages. Identity cues such as names and pronouns may be more salient and easier to detect, while topics require deeper thinking and interpretation. Similarly, identity cues may be easier to perceive than cross-group differences in topic interest. The respondents' behavior may also reflect an underlying belief that identity is the most important driver of relatability for a piece of text.

6.3 Joint estimation of relatability effects

We next compare the effect of all three measures of relatability together to explore whether each relatability measure estimates a separate effect and assess the magnitudes of the effects. We run this joint specification following our main topic-relatability specification (equation (4)), in which we include demographic-group-by-grade fixed effects and passage-exposure-sum fixed effects and all three relatability measures.[20] We continue to standardize each relatability measure by standard deviation to ensure comparable magnitudes. Table 7 displays the results of these regressions separately for race and gender. Column (1) shows that topic, identity, and survey relatability all have a positive and statistically significant effect on test performance for race. Topic relatability has the largest standardized coefficient. This likely reflects greater cross-passage dispersion in the construction of the topic measure relative to that of the identity and survey measures. When scaled by within-passage variation, the standardized topic and identity coefficients become similar in magnitude. Column (2) shows that only survey relatability has a positive and statistically significant effect on test performance for gender. Given the strong survey-identity correlation for gender, multicollinearity attenuates the identity coefficient once survey relatability is included.

As before, we also consider the total effect sizes scaled by the relatability gaps between demographic groups. For both race and gender, accounting for all three measures leads to a larger scaled effect than simply considering the estimated topic-relatability and identity-relatability effects. These differences are statistically significant, suggesting that the relatability differences across demographic groups identified by survey respondents have a meaningful impact on demographic test-score differences. However, given the correlation between the identity-relatability measure and the survey-based relatability measure, we proceed by considering only topic and identity relatability for counterfactual policy analysis. Closing gender relatability gaps would widen the male–female test-score gap, implying that content relatability is not a first-order driver of the observed gender gap.

7 Extensions

7.1 Reconsidering student test standards

Our analyses demonstrate that content relatability systematically affects reading-comprehension scores, inflating measured Black–White and Hispanic–White gaps. Beyond shifting average

[20] The results are largely similar when replacing passage-by-exposure-sum fixed effects with passage fixed effects.

Table 7: Joint estimation of relatability effects on test performance

	(1) Race	(2) Gender
Topic relatability	0.0194***	-0.0045
	(0.0069)	(0.0054)
Identity relatability	0.0038**	0.0005
	(0.0017)	(0.0007)
Survey relatability	0.0023**	0.0038***
	(0.0009)	(0.0008)
Race-Grade FE	✓	
Gender-Grade FE		✓
Passage-by-Exposure sum FE	✓	✓
N of topic-passages	4,536	4,536
N of student-pasages	59,305,495	59,305,495

Notes: Each specification is a regression of the share of items answered correctly on various measures of relatability. Standard errors reported in parentheses below coefficient estimates are clustered by exam. Standard errors for topic relatability are obtained using a shock-level regression. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$. Column (1) reports estimates for race, and column (2) reports estimates for gender. The full estimation sample includes students in grades 3 to 8 from 2013 to 2019 taking the standard STAAR reading-comprehension test for passages which contain estimates for topic, identity, and survey-based relatability.

scores, relatability may also directly disadvantage students. We consider whether Black and Hispanic students are misclassified as failing to meet key test-performance standards.

Whether a student meets performance standards can have an important impact on future learning. For students in all grades, misclassification based on the reading-comprehension test can lead to students being placed in an intensive program or a remedial class for that subject. During our study period, Texas law required school districts to provide "accelerated instruction" for students after nonsatisfactory performance on a standardized test (Texas Education Code § 28.0211 2013, 2019). School districts are also required to create "intensive programs of instructions" for such students (Texas Education Code § 28.0213 2013, 2019). At the other end of the spectrum, reaching higher test standards helps determine eligibility for gifted programs. These test standards also aggregate to the school and district levels, where the share of students meeting standards feeds directly into accountability indices. Misclassification due to exam performance can lead to bigger consequences for students in grades 5 and 8. In addition to all of the previously mentioned potential impacts, exams are used as a grade-promotion requirement during our sample period. In practice, we observe relatively few students repeating a grade. However, the requirement to retake an exam is close to universally binding. We identify clean discontinuities at the expected score cutoff;

answering just one additional question incorrectly at the cutoff increases the probability of a retaken test from 0.0% to 97.8% (see Appendix Figure A7).

We test what share of students would have been put in a higher standards category if topics and characters—and, therefore, relatability—for a given test had been different. For each grade, we identify the test in our sample that minimizes Black–White and Hispanic–White relatability differences. We use these tests as relatability benchmarks and adjust observed scores based on the relatability coefficient estimates obtained using equations (4) and (6). Using these adjusted scores, we assign students to new performance-standards categories. Further details on the institutional setting, standards categories, and new performance-standards assignment procedures can be found in Appendix Section G.

Appendix Tables A18 and A19 summarize the results of applying this exercise to the samples of Black and Hispanic students, respectively. 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 placed 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, but the magnitudes differ across grades. Across all standards, 1.4% more Black students in grade 3 would have been placed in a higher performance category with relatability adjustments, but almost no Black students in grade 6 would have been placed in a higher performance category. For Hispanic students, we find 0.9% underclassification for grade 6 on the high end and 0.3% underclassification for grade 5 on the low end. In total, almost 12,000 Black students and almost 34,000 Hispanic students might have achieved a higher reading-comprehension standard if relatability had been more equal across groups.

In short, relatability affects not only average scores but classification into consequential categories. Even modest shifts in relatability are large enough to push thousands of students into different instructional tracks. Further, even though our results suggest smaller relatability-based distortions for students in higher grades, 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.

7.2 Counterfactual policies

We consider a few policies that the test maker could implement, focusing on test-score differences across race. In light of what we have discussed thus far, the most obvious goal a policymaker may want to adopt is reducing the influence of content relatability on test scores, thereby more precisely measuring student ability θ_i. Given racial differences in relatability,

these changes to exams would ensure that estimates of θ_i do not systematically diverge by race.

We briefly formalize the decision the policymaker needs to make, borrowing notation from Section 2. Let $\vec{\mu}_p = (\vec{\mu}_p^{topic}, \vec{\mu}_p^{identity})$, which represents the relatability-relevant passage attributes of passage p and consists of two subvectors—representing topic salience, $\vec{\mu}_p^{topic} = (\mu_p^1, \dots \mu_p^t, \dots)$, and character-identity composition, $\vec{\mu}_p^{identity} = (\mu_p^1, \dots \mu_p^d, \dots)$. For each student in d, there is an analogous vector $\vec{\varepsilon}_d = (\vec{\varepsilon}_d^{topic}, \vec{\varepsilon}_d^{identity})$, giving the average degree of relatability of group d to topics and character identities. Taking the inner product and gathering terms yields our relatability measures: topic relatability, $\sum_t \varepsilon_{dt}^{topic} \mu_p^t$, and identity relatability, $\varepsilon_d^{identity} \mu_p^d$.[21] Our empirical strategy yields the test-score effect of topic relatability, $\hat{\beta}^{topic}$, and the test-score effect of identity relatability, $\hat{\beta}^{identity}$.

Given these estimates and relationships between student and potential passage attributes, the policymaker selects a set of passages $\mathcal{P}$. Formally, they choose a target $\vec{\mu}^*$, which is the average of $\vec{\mu}_p$ across the set of passages $\mathcal{P}$. We suppose in this case the policymaker is concerned about differential contribution of relatability across two racial groups d_0 and d_1. Then they select $\vec{\mu}^*$ such that

$$\vec{\mu}^* = \arg\min_{\vec{\mu}} \left| \underbrace{\hat{\beta}^{topic} \sum_t \left(\varepsilon_{d_0 t}^{topic} - \varepsilon_{d_1 t}^{topic} \right) \mu^t}_{\text{difference in topic relatability}} + \underbrace{\hat{\beta}^{identity} \left(\mu^{d_0} - \mu^{d_1} \right)}_{\text{difference in identity relatability}} \right|. \tag{7}$$

That is, the policymaker selects topics and character identities such that they minimize the differential impact of relatability on test scores across racial groups.

We now consider what would happen to test scores if a policymaker embarked on this minimization with various constraints. Before doing so, we note that leisure-related topics and character identities are not the only dimensions of content which may vary in relatability across demographic characteristics. Therefore, topic relatability and identity relatability as defined thus far may not be the only things a policymaker would want to change to more precisely measuring test performance. Even so, as long as our topic-relatability and identity-relatability measures are not correlated with these additional content dimensions, this exercise would represent an improvement to estimated test score gaps, regardless of other sources of mismeasurement in the exams.

[21] The latter is true because we assume that a student in group d only relates to characters of group d.

7.2.1 Unconstrained equalization across groups

Equation (7) shows that equalizing the effect of relatability means selecting passages such that differences in group exposures and character-identity effects cancel out. In theory, if the set of feasible $\vec{\mu}_p$ is unconstrained, test makers can achieve this by strategically selecting passages that generate the same average relatability across racial groups and ensure a similar proportion of character identities. This is also always trivially possible if the test maker sets $\vec{\mu}^* = (0, \ldots, 0)$. We find that if relatability were equalized in this way, policymakers could observe 8.2% smaller test-score gaps between Black and White students and 9.7% smaller test-score gaps between Hispanic and White students.

7.2.2 Equalization based on existing exams

In reality, however, test makers may face constraints on the available combinations of passage attributes, $\vec{\mu}_p$. For instance, our analysis thus far does not take into account the stock of passages from which test makers can choose when constructing exams or the presence of additional political or educational considerations that influence topic selection or character-identity selection in exams.[22] We can add an additional constraint to equation (7): $\vec{\mu}^*$ must be based on the observed $\vec{\mu}_p$ in our passage sample. Practically, we repeat a version of the exercise in Section 7.1. First, we search for the exam within each grade that minimizes the non-White–White gap in average content relatability and designate this race-level relatability as the best feasible relatability for their respective grades. Second, we predict counterfactual test scores under the scenario in which students in all years received their grade's best feasible relatability. We estimate only 1.8% smaller Black–White and 2.4% smaller Hispanic-White test gaps through this procedure. Since the relatability adjustments in this exercise reflect observed relatability measures from our sample, this demonstrates a lower bound of what policymakers may achieve even in the presence of existing constraints on passage-attribute selection.

We find differences between Black and Hispanic students in the primary source of the feasible gap closure. For Black students, much of the relatability differences stem from topic relatability, indicating large variance within grade on the topics that are more relatable to Black students. For Hispanic students, much of the difference stems from identity relatability, indicating that some tests feature significantly more Hispanic representation than others.

[22] 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 about, which may mean the corpus of selectable passages is already biased prior to selection by test makers. This may be exacerbated or ameliorated if there are considerations that prioritize Texan writers or stories that are based primarily in Texas.

7.2.3 Equalizing topic distributions

Alternatively, a test maker could object to the previous methods 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 that should be corrected (for example, differences in household financial resources), this test maker may want to be agnostic about that heterogeneity. Thus, their ideal topic distribution would be one in which all topics are equally likely: $\vec{\mu}^{topic^*} = (\mu^*, \mu^*, \dots)$. Under this policy, content relatability would be given by the difference in aggregate topic exposure across racial groups.

We can illustrate what this means in practice by decomposing the empirical topic-relatability differences across groups in our sample into the portion that is due to the topic distribution and that due to differences in topic exposure. For this exercise, we use the objects m_{tp}, e_{dt} (see Section 3), which are the empirical analogues to $\mu_{tp}, \varepsilon_{dt}$, respectively. We can decompose cross-group differences in average relatability as

$$\bar{r}_{d_0} - \bar{r}_{d_1} = (E_{d_0} - E_{d_1})\bar{m} + \frac{\sum_p \sum_t (e_{d_0 t} - e_{d_1 t})\tilde{m}_{tp}}{|\mathcal{P}|}\,, \tag{8}$$

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_0} and E_{d_1} are the sums of exposures across topics. The first term of the right-hand side of equation (8) 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.[23] Intuitively, the first term reflects 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 beyond this baseline difference must be due to topic selection being skewed toward one group over the other. It follows that setting $\vec{\mu}^{topic^*} = (\mu^*, \mu^*, \dots)$ is equivalent to allowing only the first term in equation (8) to affect test performance.

We can apply equation (8) to our sample to see the extent to which topic selection affects topic-relatability differences across students. We find that the topic-salience residuals $\tilde{m}_{tp}$ contribute to around one-third of Black–White (34%) and Hispanic–White (33%) differences in average relatability. Put differently, if test makers had counterfactually set $\vec{\mu}^{topic^*} = (\bar{m}, \bar{m}, \dots)$, topic-relatability differences would be almost a third smaller than currently observed relatability differences by race. This implies that 1.3% of Black–White gaps and

[23] Since $\tilde{m}_{tp}$ is a residual, it is possible for this term to be negative even if $e_{d_0 t} - e_{d_1 t} > 0$, $\forall t$.

1.2% of Hispanic–White gaps can be explained by the topics selected. Combined with our estimate of the effect of adjusting identity relatability, adjusting tests in this way could close roughly 6%–7% of the non-White–White test gap.

We collect all counterfactual estimates in Figure 8, including estimates of the relatability-explained part of the gender gap. Relatability explains a larger share of the Hispanic–White test-score gap than the share of the Black–White gap. As discussed earlier, relatability does not explain the male–female test-score gap. While identity relatability has a statistically significant impact on test performance, the underrepresentation of female characters in test passages suggests that the male–female gap is potentially *understated*.

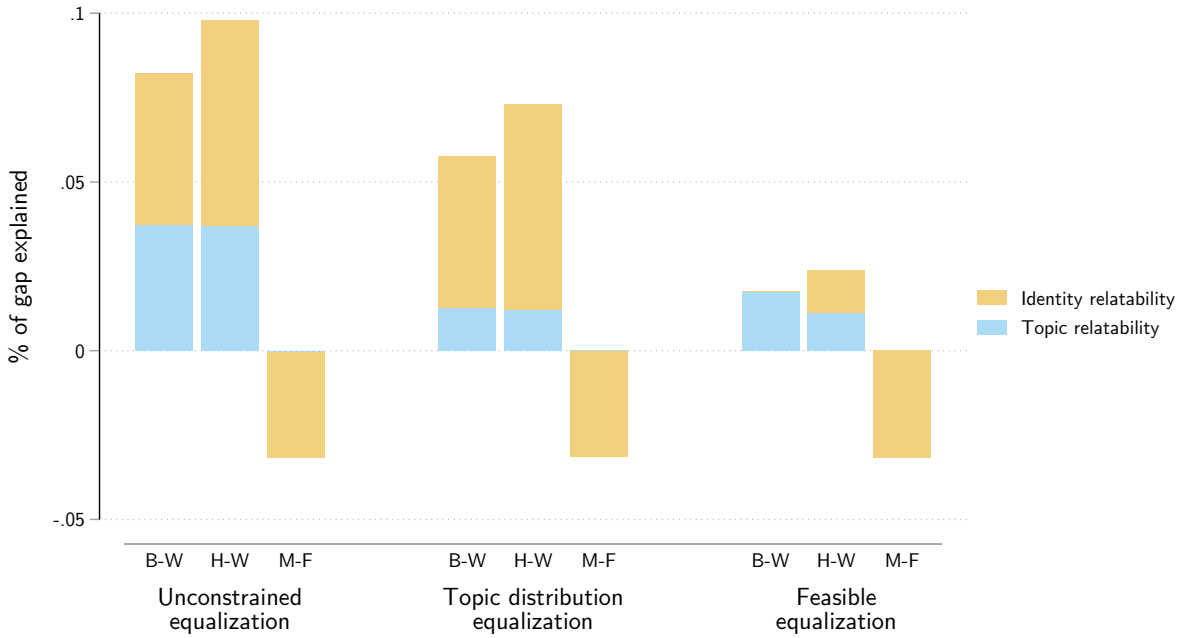
8 Conclusion

This paper demonstrates the extent to which relatable content in standardized tests impacts test performance and contributes to test-score disparities. We distinguish two sources of relatability for test takers: relatability stemming from interest or familiarity with topics in exam text and relatability stemming from shared identity with characters in exam text. Our methodology combines time-use data from the American Time Use Survey with natural language processing to form race- and gender-specific estimates of relatability to reading-passage topics. We find that our race-based measure is predictive of students' standardized test performance: a 1 SD higher race-based relatability for a passage leads to a 1.9pp increase in probability of answering questions for that passage correctly. We use large language models and name-demographic-group databases to build an analogous measure of identity relatability. Identity relatability has positive effects on test scores for both race and gender, and these estimates are independent of the effect of topic relatability on test scores. Given differences in average content relatability across race, we find that relatability accounts for 8% of Black–White test gaps and 10% of Hispanic–White test gaps in our student sample.

Our results have implications for both test writers and education policymakers. 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 test performance. Second, when policymakers consider outcome differences along demographic dimensions, one additional component to examine may 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.

Alternatively, one may interpret our results from the perspective that adaptability to

Figure 8: Estimated effect of equalizing relatability on demographic test score gaps



Notes: Each bar represents the share of the test-score gap explained by topic relatability (lower bars) and identity relatability (upper bars). "B-W" indicates the Black–White test gap, "H-W" indicates the Hispanic–White test gap, and "M-F" indicates the male–female test gap. "Unconstrained equalization" shows the share of the test gap that would be explained if relatability were equalized across the two groups. "Topic distribution equalization" shows the share of the test gap that would be explained if residual relatability (after accounting for the average topic salience and character identities) were balanced across the two groups. "Feasible equalization" shows the share of the test gap that would be explained if relatability for each grade were set to the smallest observed exam-level relatability-effect difference between the two groups. Since topic relatability and identity relatability are determined jointly within a passage and exam, the relatability-effect difference is calculated by averaging the effect stemming from the topic-relatability gap and the effect stemming from the identity-relatability gap. The coefficient estimates from equations (4) and (6) are used for topic relatability and identity relatability, respectively. The estimation sample includes all students in grades 3 to 8 from 2013 to 2019 taking the standard STAAR reading-comprehension test.

different environments and new concepts is an important student learning ability, one that should be tested in exams. Under this view, students should read passages regarding topics or identities with which they are unfamiliar or may find unrelatable. We discuss in Section 2.1, for instance, how the state standards in Texas expect students to gain insights from different cultural and historical contexts. We contend that our findings still have meaningful, policy-relevant implications under this view. Recast in this light, our main result demonstrates that students on average are not fully "adaptable" given that we are able to predict they will perform worse on topics with which they are less familiar. This suggests that edu-

cation curricula should place greater emphasis on teaching students skills to be "adaptable." Further, if test makers are deliberately including concepts or settings unrelatable to students as a way to assess reading-comprehension skills, they must still internalize the fact that familiarity or interest differs by demographic group: an unfamiliar topic to one group may be a familiar topic to another. Ultimately, the stated goals of most reading-comprehension exams do not include testing for breadth of topic knowledge or for adaptability to unknown topics, merely that students be able to read texts that carry cultural context. Insofar as stated testing standards reflect the knowledge and skills educators truly expect students to have, we take these standards seriously in our primary interpretation of our findings, setting aside any ancillary skills that educators would like to test.

We note general limitations of our research design, highlighting potential avenues for future research. First, we do not elicit the underlying determinants of topic relatability and identity relatability directly from test takers. Ideally, we would observe a student's interest in topics or the set of character identities which resonate with the student. This lack of visibility leads to coarse estimates of relatability for test takers. Next, our topic relatability measure is dependent on the activities delineated in the American Time Use Survey. These data provide sufficient delineation within certain topic areas such as sports, where we observe granular data for time spent playing tennis versus time spent playing volleyball. It does not, however, provide delineation within certain topic areas such as music, where all musical genres are collapsed into one activity. The reliance on time-use data confines us primarily to leisure-related topics, even though many other topics may be differentially relatable across race and gender. This measurement constraint implies our topic relatability estimates likely represent a lower bound on the total effect of familiarity with or interest in content on test performance. We face a related issue for identity relatability, where much of the variation for race is driven simply by a character's name. Finally, we have limited scope for studying the mechanisms behind why relatability affects student performance. Bridging this gap requires information on how much time students spend on each passage or the order in which they answer questions. As standardized exams move increasingly to online formats, these outcomes may become available in the future.

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.

Baldazzi, E., Biroli, P., Giusta, M.D., & Dubois, F. (2025). Seeing Stereotypes. arXiv preprint arXiv:2503.02146.

Bandura, A. (1977). *Social learning theory.* Englewood Cliffs, NJ: Prentice-Hall.

Bisin, A., & Verdier, T. (2001). The Economics of Cultural Transmission and the Dynamics of Preferences. *Journal of Economic Theory, 97*(2), 298–319. https://doi.org/10.1006/jeth.2000.2678.

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

Bond, T.N., & Lang, K. (2018). The Black–White Education Scaled Test-Score Gap in Grades K-7. *The Journal of Human Resources, 53*(4), 891 - 917.

Borusyak, K., Hull, P., & Jaravel, X. (2022). 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

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

Card, D. & Rothstein, J. (2007). Racial Segregation and the Black-White Test Score Gap.

Journal of Public Economics, 91(11-12), 2158-2184. https://doi.org/10.1016/j.jpubeco.2007.03.006

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.

Cobb-Clark, D.A., & Moschion, J. (2017). Gender Gaps in Early Educational Achievement. *Journal of Population Economics, 30*, 1093–1134. https://doi.org/10.1007/s00148-017-0638-z

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.

Davey, B., & Kapinus, B. A. (1985). Prior Knowledge and Recall of Unfamiliar Information: Reader and Text Factors. *The Journal of Educational Research, 78*(3), 147–151. https://doi.org/10.1080/00220671.1985.10885590

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.

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-Bundle DIF Hypothesis Testing: Identifying Suspect Bundles and Assessing Their Differential Functioning. *Journal of Educational Measurement, 33*(4), 465–484. http://www.jstor.org/stable/1435335

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.l05002505820401

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.

Good, J. J., Woodzicka, J. A., & Wingfield, L. C. (2010). The Effects of Gender Stereotypic and Counter-Stereotypic Textbook Images on Science Performance. *The Journal of Social Psychology, 150*(2), 132–147. https://doi.org/10.1080/00224540903366552

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.

Imai, K., & Khanna, K. (2016). Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records. *Political Analysis, 24*(2), 263–272. doi:10.1093/pan/mpw001

Imai, K., Olivella, S., & Rosenman, E. T. R. (2022). Addressing Census Data Problems in Race Imputation via Fully Bayesian Improved Surname Geocoding and Name Supplements. *Science Advances, 8*(49), eadc9824. https://doi.org/10.1126/sciadv.adc9824

Johnston, P. (1984). Prior Knowledge and Reading Comprehension Test Bias. *Reading Research Quarterly, 19*(2), 219–239. https://doi.org/10.2307/747364

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.

Lundberg S. (2020). Educational Gender Gaps. *Southern Economic Journal, (87)*, 416–439. https://doi.org/10.1002/soej.12460

Morris, C. (1983). Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association, 78*(381). https://doi.org/10.2307/2287098.

Norris, S. P., Phillips, L. M., & Korpan, C. A. (2003). University Students' Interpretation of Media Reports of Science and its Relationship to Background Knowledge, Interest, and Reading Difficulty. *Public Understanding of Science, 12*(2), 123-145. https://doi.org/10.1177/0963662503012200

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

Ofek-Shanny, Y. (2024). Measurements of Performance Gaps Are Sensitive to the Level of Test Stakes: Evidence from PISA and a Field Experiment. *Economics of Education Review,*

98, 102490. https://doi.org/10.1016/j.econedurev.2023.102490.

Pope, D. G., & Sydnor, J. R. (2010). Geographic Variation in the Gender Differences in Test Scores. *Journal of Economic Perspectives, 24*(2), 95–108. https://doi.org/10.1257/jep.24.2.95

Schraw, G., Bruning, R., & Svoboda, C. (1995). Sources of Situational Interest. *Journal of Reading Behavior, 27*(1), 1-17. https://doi.org/10.1080/10862969509547866

Shirey, L. L., & Reynolds, R. E. (1988). Effect of Interest on Attention and Learning. *Journal of Educational Psychology, 80*(2), 159–166. https://doi.org/10.1037/0022-0663.80.2.159

Singer, L. M.,& Alexander, P. A. (2016). Reading Across Mediums: Effects of Reading Digital and Print Texts on Comprehension and Calibration. *The Journal of Experimental Education, 85*(1), 155–172. https://doi.org/10.1080/00220973.2016.1143794

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.

Texas Administrative Code §§ 110.10–110.20. (2008). English Language Arts and Reading, Grades K–8, Beginning with School Year 2009-2010. In *Texas Administrative Code*.

Texas Education Code § 28.0211. (2013). Satisfactory Performance on Assessment Instruments Required; Accelerated Instruction. In *Texas Statutes*. https://statutes.capitol.texas.gov/StatutesBy

Texas Education Code § 28.0211. (2019). Satisfactory Performance on Assessment Instruments Required; Accelerated Instruction. In *Texas Statutes*. https://statutes.capitol.texas.gov/StatutesBy

Texas Education Code § 28.0213. (2013). Satisfactory Performance on Assessment Instruments Required; Accelerated Instruction. In *Texas Statutes*. https://statutes.capitol.texas.gov/StatutesBy

Texas Education Code § 28.0213. (2019). Satisfactory Performance on Assessment Instruments Required; Accelerated Instruction. In *Texas Statutes*. https://statutes.capitol.texas.gov/StatutesBy

Walters, C. (2024). Empirical Bayes methods in labor economics. *Handbook of Labor Economics, 5*, 183–260. https://doi.org/10.1016/bs.heslab.2024.11.001.

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

Zumbo, Bruno. (1999). A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and

Likert-Type (Ordinal) Item Scores.

For Online Publication

A Appendix tables and figures

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

	Mean Min./Respondent	# of activity codes
All activities	1427	442
Excluded activities	1113	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	170	110
Included activities	314	140
N	73,626	73,626

Notes: This table displays an overview of ATUS activity codes which are not included in our topic set. Each row represents the statistics for each group of activity codes. Column 1 represents the average reported minutes per respondents for the group of activities. Column 2 represents the total number of activity codes for the group of activities. Respondent observations are weighted by ATUS sampling weights provided by the U.S. Census Bureau. The sample includes all ATUS respondents from 2013–2019.

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

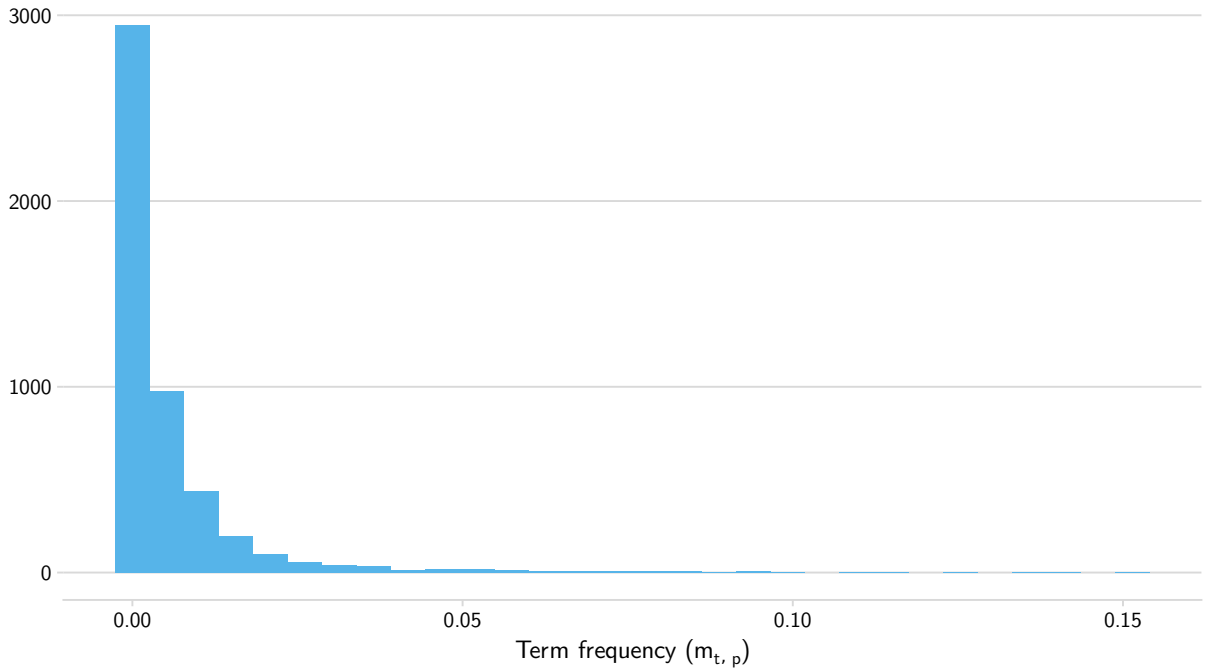
Notes: This table provides an overview of the ATUS activities which make up each topic. Each topic is associated with a set of mutually exclusive ATUS activities/activity codes. A full mapping of activities to topics are available in the online code repository.

Table A3: Relative exposure of groups to topics

	Asian-white	Black-white	Hispanic-white	Male-female
Animal sports	0.000	0.025	0.538	0.511
Animals	0.328	0.312	0.492	0.749
Arts and crafts	0.573	0.418	0.519	0.632
Baseball	0.310	0.228	0.751	1.330
Basketball	0.991	3.189	1.020	3.308
Club sports	0.851	0.223	0.289	3.813
Computer	1.491	0.642	0.763	0.899
Exercise	1.381	0.722	0.891	1.224
Food	1.023	0.843	0.906	0.645
Football	0.305	0.990	1.497	4.208
Indoor recreation	0.558	0.809	0.501	1.896
Media	0.883	0.997	0.975	1.024
Music	1.147	1.423	1.289	1.467
Nature sports	0.261	0.287	0.396	4.214
Performing arts	0.796	0.732	0.802	0.844
Plant/garden/yard	0.675	0.457	0.685	1.487
Religion	1.218	1.903	1.181	0.634
Soccer	1.371	0.697	4.184	1.979
Street sports	1.639	0.802	1.048	0.894
Traveling	0.858	0.895	0.952	0.896
Vehicles	0.598	0.716	0.979	2.604
Volunteering	0.507	0.679	0.535	0.746
Water sports	0.854	0.236	0.481	1.067
Winter sports	0.704	0.239	0.061	1.384

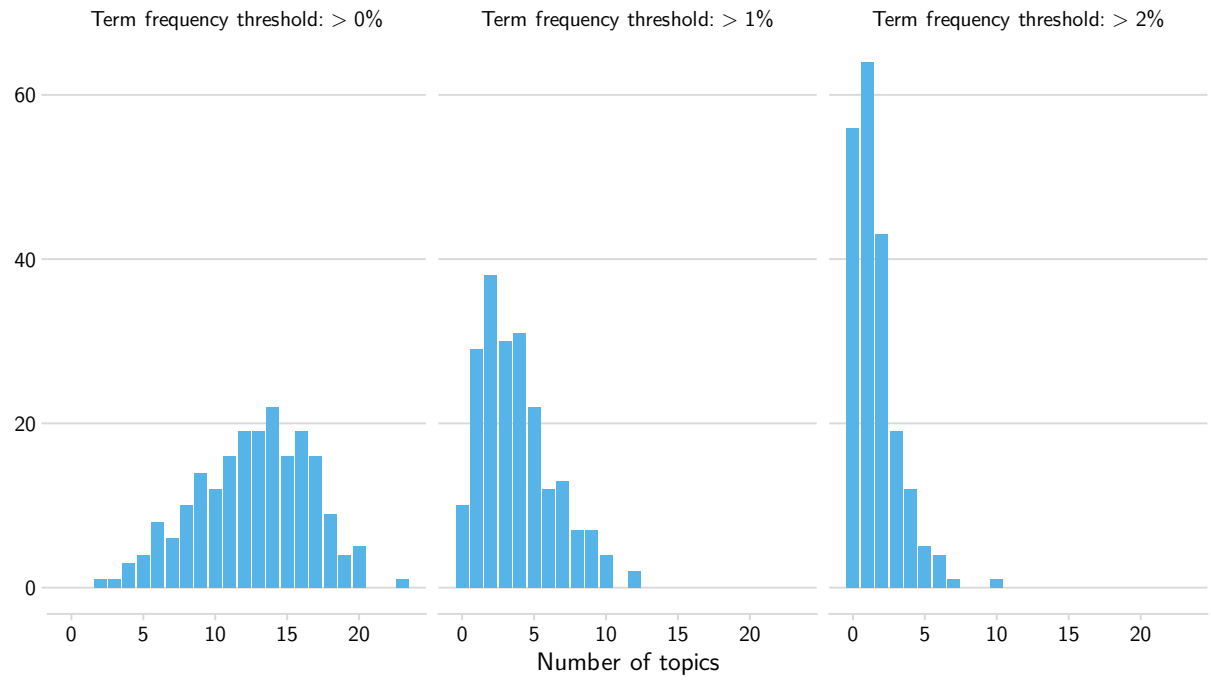
Notes: Each cell in this table represents a ratio of topic exposures between two demographic groups. Exposure is the share of ATUS diaries for a demograhpic group reporting participating in any activity in the topic. For racial groups, exposure for White respondents is used as the comparison group. Topics are formed combining six-digit ATUS activity codes. Respodnent observations are weighted by ATUS sampling weights provided by the U.S. Census Bureau. 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 ($m_{t,p}$) displayed 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: Number of topics appearing in each passage



Notes: The unit of observation in this figure is a passage. Reading passage topic-salience measures are calculated for each topic-passage pair with the term-frequency metric ($m_{t,p}$) 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, for which passages are not copyright restricted.

Table A4: Effect of relatability on falsification outcome variables

	Coeff.	SE	N (race-passage)
Prior year performance...			
by grade-race	-0.0031	(0.0035)	700
by cohort-race	0.0033	(0.0032)	592
by grade-race-passage position	0.0041	(0.0067)	688
by cohort-race-passage position	0.0059	(0.0091)	568
Previous passage perf.	0.0048	(0.0081)	640
Subsequent passage perf.	-0.0006	(0.0066)	640
Population of race by exam	-0.0039	(0.0024)	820
Exam year (continuous)	-0.0747	(0.1049)	820
Passage position (continuous)	0.0450	(0.0675)	820
Passage word count	6.2689	(9.6022)	820
Literary passage	0.0716***	(0.0178)	820
Predicted performance (all variables)	0.0067	(0.0073)	340
Predicted performance (non-missing variables)	0.0019	(0.0032)	820
	Statistic		p-value
Joint F-test of covariates	3.901		0.001

Notes: Each row reports the coefficient from a regression of the specified outcome variable on race-based topic relatability. Unreported controls in each specification include: (1) race-grade fixed effects and (2) exposure-sum-by-passage fixed effects. Standard errors reported in parentheses below coefficient estimates are (a) obtained using shock-level regression following Borusyak et al. (2022) and (b) clustered by exam. $^{*}\,p < .10$, $^{**}\,p < .05$, $^{***}\,p < .01$. Prior-year performance is the test performance for the listed group in the previous year. Previous and subsequent passage performances are scores on the preceding and subsequent passages, respectively. Population of race by exam is the share of test takers who are the same race as the student. Literary passage describes one of two passage categories that make up exams. Observations are at the race-passage level and are weighted by the number of student-item. 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: Topic-passage-level summary statistics

	(1)	(2)
Topic salience summary		
Mean	0.007	-0.000
Standard deviation	0.013	0.011
Interquartile range	0.009	0.007
Inverse HHI of topic exposure weights		
Across topics and passages	925.461	925.461
Across exams	41.504	41.504
Largest topic exposure weight		
Across topics and passages	0.003	0.003
Across exams	0.029	0.029
Residualized on $tg(p)$ and p FEs		✓
N of topic-passages	4,920	4,920
N of passages	205	205
N of topics	24	24

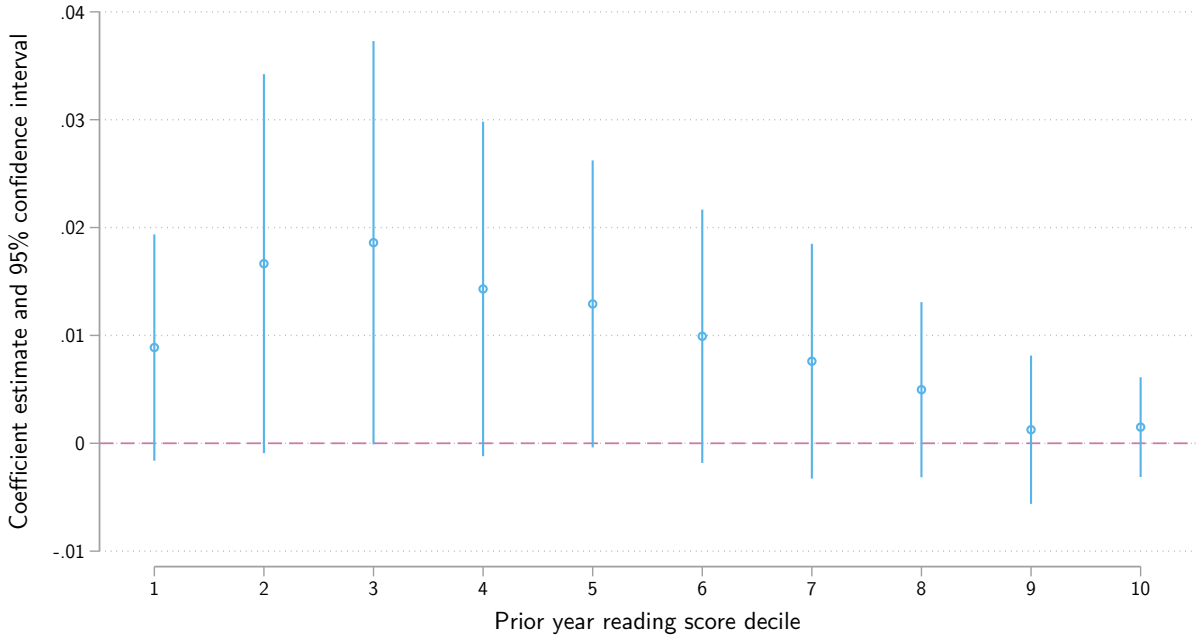
Notes: The table summarizes the distribution of topic-salience shocks g_{tp} and the aggregated topic-exposure weights at the topic-passage level. Topic-salience shocks are developed using NLP over the set of passages. Aggregated topic-exposure weights are developed based the topic-exposure shares calculated at the demographic group and passage-level. Topic-exposure shares are averaged within a topic-passage, with each demographic group topic-exposure share weighted by the number of student-items for that demographic group and passage. Then these aggregated values are normalized so that the weights sum to one across all topic-passage pairs. Statistics are weighted by these topic-exposure weights. Column (1) displays statistics with no controls. Column (2) displays statistics after residualization on topic-grade ($tg(p)$) and passage (p) FEs. The inverse HHI of topic-exposure weights is calculated as the inverse of the sum of squared aggregated exposure weights. Exposure weights are aggregated in two ways, within topic-passage pairs or within exam.

Table A6: Black-White and Hispanic-White test-score gaps by grade

	Grade					
	3	4	5	6	7	8
Black-white	-0.136	-0.137	-0.128	-0.127	-0.122	-0.118
Hispanic-white	-0.094	-0.093	-0.094	-0.109	-0.104	-0.100
N of students	2,155,843	2,175,995	2,202,413	2,232,987	2,220,076	2,192,824

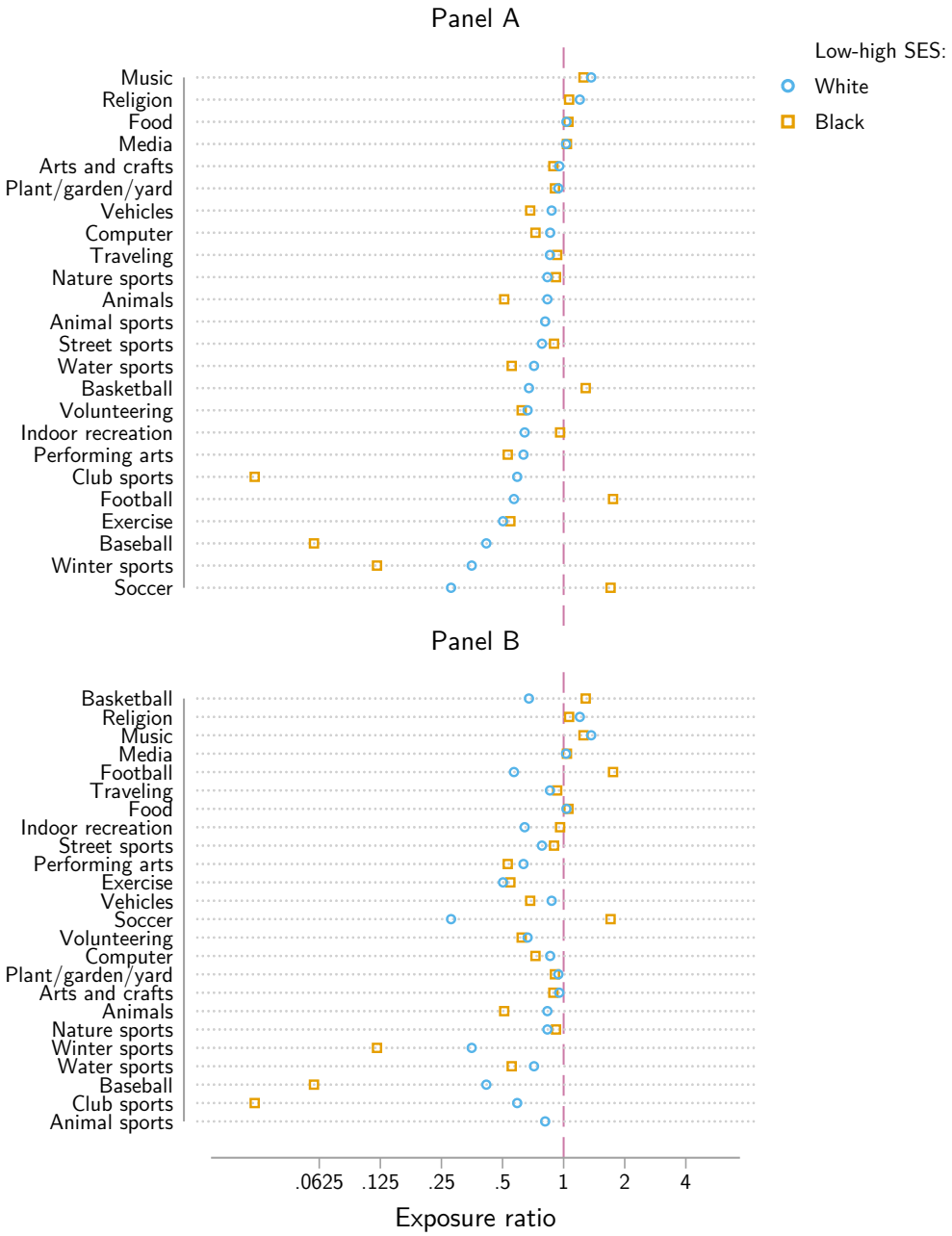
Notes: Each column displays the test-score gap between different racial roups for a given grade. Test-score gaps are calculated as the average difference between two groups' average share of questions correct at the test-level. The sample includes all students grades 3 to 8 from 2013–2019 taking the standard STAAR reading-comprehension test.

Figure A3: Heterogeneity of race-based topic relatability effects by past reading achievement



Notes: Each coefficient estimate is from a separate specification which regresses the share of items answered correctly on a passage on race-based topic relatability. Unreported controls include: (1) race-grade fixed effects and (2) exposure-sum-by-passage fixed effects. Each specification is run on a separate student subsample based on the decile of a student's prior-year reading test score. Each student's prior-year reading test score is from the standard STAAR exam for reading comprehension. Score deciles are formed within exam. Exposure-robust standard errors clustered by exam are used to construct the 95% confidence intervals. Observations are at the race-passage level and are weighted by the number of student-items. The estimation sample includes all students grades 4 to 8 from 2014–2019 taking the standard STAAR reading-comprehension test. This differs from the estimation sample in the rest of this manuscript, as prior-year test scores are not available for 2013 and grade 3.

Figure A4: Differences in topic exposure by SES



Notes: Graph plots the ratio of topic exposures between low SES and high SES individuals separately by race. Panel A orders topics by SES exposure ratio for White individuals. Plot B orders topics by the Black-White exposure ratio. The vertical dotted line at 1 corresponds to the value at which both groups have identical exposure. 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.

Table A7: Impact of race-based topic relatability on test performance by time-use sample

	(1)	(2)	(3)	(4)	(5)
Race topic rel. (PSID)	0.0103***		0.0162***		
	(0.0034)		(0.0059)		
Race topic rel. (ATUS)		0.0187***		0.0204***	
		(0.0058)		(0.0071)	
ATUS-PSID average					0.0218***
					(0.0072)
Black-white rel. diff.	-0.0045	-0.0048	-0.0069	-0.0052	-0.0068
	(0.0015)	(0.0015)	(0.0025)	(0.0018)	(0.0022)
Hispanic-white rel. diff.	-0.0014	-0.0037	-0.0022	-0.0040	-0.0037
	(0.0005)	(0.0011)	(0.0008)	(0.0014)	(0.0012)
Race-Grade FE	✓	✓	✓	✓	✓
Passage-by-Exposure sum FE	✓	✓	✓	✓	✓
Sample	All	All	Excl. Asian	Excl. Asian	All
N of topic-passages	4,920	4,920	4,920	4,920	4,920
N of student-passages	64,352,860	64,352,860	61,120,843	61,120,843	64,352,860

Notes: Each specification is a regression of the share of items answered correctly on a passage on race-based topic relatability. Standard errors reported in parentheses below coefficient estimates are (a) obtained using a shock-level regression and (b) clustered by exam. $^{*}\ p < .10$, $^{**}\ p < .05$, $^{***}\ p < .01$. PSID race-based topic relatability uses the Panel Study of Income Dynamics-Child Development Supplement to calculate race-level topic exposures. ATUS-PSID average uses race-level topic exposure calculated as an average of the ATUS-based topic-exposure measure and the PSID-based topic-exposure measure, after both topic-exposure measures are adjusted using an empirical Bayes shrinkage estimator which shrinks race topic-exposure measures to the population-level topic-exposure means. Coefficient estimates (and standard errors) alternatively scaled by racial differences in topic relatability are displayed in the bottom rows. This scaling is done by dividing the coefficient estimates by the average Black-White and Hispanic-White relatability gap. Columns (3) and (4) use data only from non-Asian students. The full estimation sample includes students grades 3 to 8 from 2013–2019 taking the standard STAAR reading-comprehension test.

Table A8: Robustness of baseline race-based topic-relatability results on test performance with school fixed effects

	(1)	(2)
Race topic relatability	0.0152***	0.0142***
	(0.0045)	(0.0043)
School-Race FE	✓	
School-Race-Grade FE		✓
School-Passage-by-Exposure sum FE	✓	✓
N of topic-passages	4,920	4,920
N of student-passages	66,494,672	66,494,672

Notes: Each specification is a regression of the share of items answered correctly on a passage on race-based topic relatability. 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. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$. For columns (1) and (2), race-based topic relatability is calculated as described in section 4.1. The estimation sample includes students grades 3 to 8 from 2013–2019 taking the standard STAAR reading-comprehension test.

Table A9: Heterogeneity of race-based topic relatability effects along school diversity index

	(1)	(2)	(3)	(4)
Race topic relatability	0.0182***	0.0193***	0.0210***	0.0161*
	(0.0050)	(0.0055)	(0.0061)	(0.0087)
HHI Quartile	1	2	3	4
Race-Grade FE	✓	✓	✓	✓
Passage-by-Exposure sum FE	✓	✓	✓	✓
N of topic-passages	4,920	4,920	4,920	4,920
N of student-passages	16,623,668	16,623,668	16,623,668	16,623,668

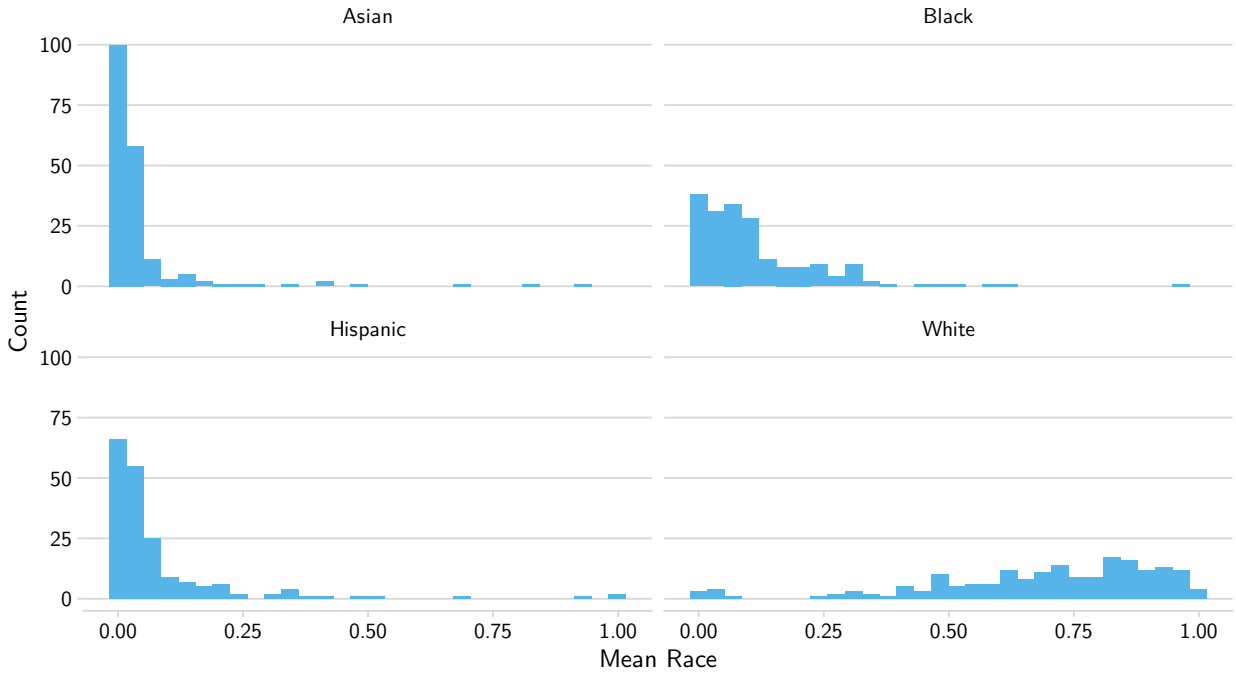
Notes: Each specification is a regression of the share of items answered correctly on a passage on race-based topic relatability on a subsample of the data. Subsamples are determined by splitting schools into quartiles of a school-integration index (HHI), and thus lower quartiles are more integrated. 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. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$. The estimation sample includes students grades 3 to 8 from 2013–2019 taking the standard STAAR reading-comprehension test.

Table A10: Impact of gender-based topic relatability on test performance by time-use sample

	(1)	(2)	(3)
Gender topic rel. (PSID)	0.00590		
	(0.00437)		
Gender topic rel. (ATUS)		-0.000148	
		(0.00554)	
ATUS-PSID average			0.00427
			(0.00726)
Gender-Grade FE	✓	✓	✓
Passage-by-Exposure sum FE	✓	✓	✓
N of topic-passages	4,920	4,920	4,920
N of student-passages	64,352,860	64,352,860	64,352,860

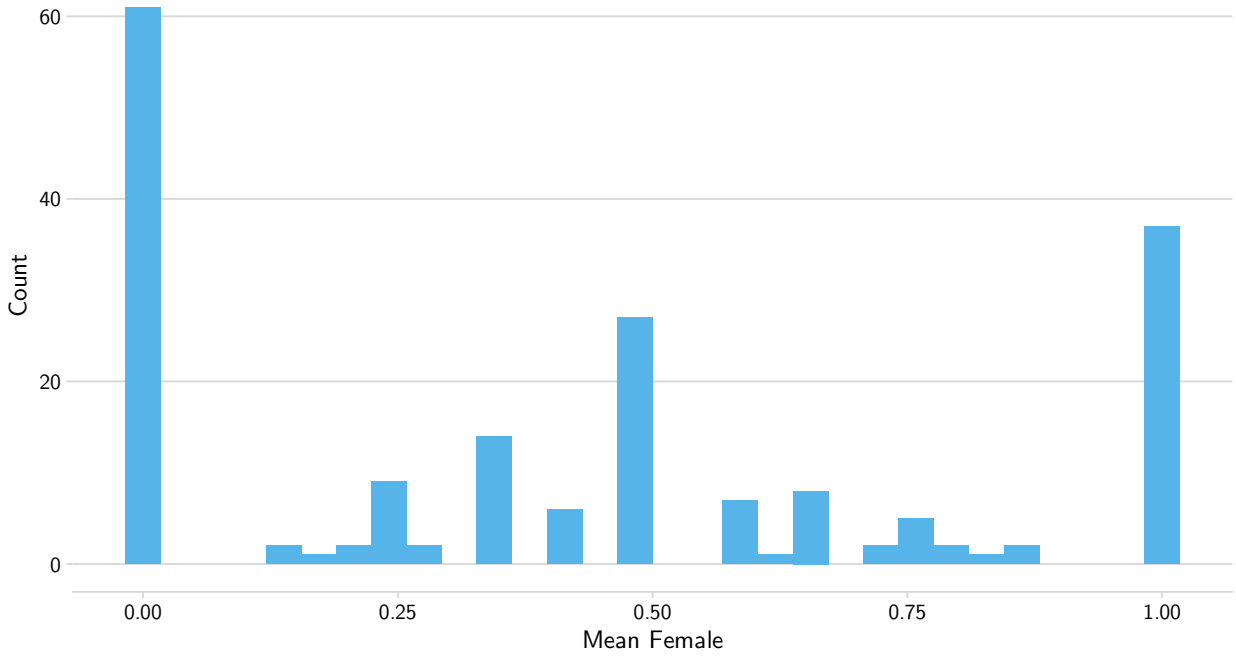
Notes: Each specification is a regression of the share of items answered correctly on a passage on gender-based topic relatability. Standard errors reported in parentheses below coefficient estimates are (a) obtained using a shock-level regression and (b) clustered by exam. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$. PSID gender-based topic relatability uses the Panel Study of Income Dynamics-Child Development Supplement to calculate gender-level topic exposures. ATUS-PSID average uses gender-level topic exposure calculated as an average of the ATUS-based topic-exposure measure and the PSID-based topic-exposure measure, after both topic-exposure measures are adjusted using an empirical Bayes shrinkage estimator which shrinks gender topic-exposure measures to the population-level topic-exposure means. The estimation sample includes students grades 3 to 8 from 2013–2019 taking the standard STAAR reading-comprehension test.

Figure A5: Histogram of race-based identity relatability



Notes: This histogram is constructed using passage level observations. Each observation is the mean predicted race for all characters in a passage, giving each character equal weight. Further details on the race-based identity relatability can be found in Section 5.1. The sample of passages include grade 3 to 8 STAAR reading-comprehension tests from 2013–2019.

Figure A6: Histogram of gender-based identity relatability



Notes: This histogram is constructed using passage level observations. Each observation is the mean predicted gender for all characters in a passage, giving each character equal weight. Further details on the gender-based identity relatability can be found in Section 5.1. The sample of passages include grade 3 to 8 STAAR reading-comprehension tests from 2013–2019.

Table A11: Additional robustness of impact of race-based identity relatability on test performance

	Texas Demographic Prior		Claude Sonnet 4		Levels		Above Race Median	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Race identity rel. (mean)	0.0103***		0.0091**					
	(0.0037)		(0.0043)					
Race identity rel. (w.mean)		0.0106***		0.0094**				
		(0.0033)		(0.0037)				
Race id. rel. (# Characters)					0.0006			
					(0.0008)			
Race id. rel. (# Mentions)						-0.0001		
						(0.0001)		
Id. rel. (mean) > Race median							0.0043***	
							(0.0011)	
Id. rel. (w.mean) > Race median								0.0040***
								(0.0011)
Reading Passage FE	✓	✓	✓	✓	✓	✓	✓	✓
Race-Grade FE	✓	✓	✓	✓	✓	✓	✓	✓
N of student–passages	59,305,495	59,305,495	58,940,948	58,940,948	59,305,495	59,305,495	59,305,495	59,305,495

Notes: Each specification is a regression of the share of items answered correctly on a passage on race-based identity relatability. Standard errors reported in parentheses below coefficient estimates are clustered by exam. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$. Identity relatability here is defined in levels rather than shares for columns (1) and (2): column (1) has number of characters matching the students race, whereas column (2) counts the total number of mentions of characters matching the (2). Note that since racial predictions are between 0 and 1, this more precisely corresponds to the "expected value" level, rather than true level. The racial prediction model is using last names from social security data and first names from six Southern states. It uses a Bayesian prior distribution matching national demographics in 2020. Columns (3) and (4) use the same identity-relatability share values as the main regression, but further dichotomize the score to be an indicator for above the median identity-relatability score within race. Thus, if the median Hispanic share of passages is 20%, passages predicted to have more than 20% of Hispanic characters are given a 1 for Hispanic students and others are given a 0. The estimation sample includes students grades 3 to 8 from 2013–2019 taking the standard STAAR reading-comprehension test.

Table A12: Additional robustness of impact of gender-based identity relatability on test performance

	Claude Sonnet 4		Levels		Above Gender Median	
	(1)	(2)	(3)	(4)	(5)	(6)
Gender identity rel. (mean)	0.0075***					
	(0.0020)					
Gender identity rel. (w.mean)		0.0064***				
		(0.0018)				
Gender id. rel. (# Characters)			0.0025***			
			(0.0006)			
Gender id. rel. (# Mentions)				0.0002*		
				(0.0001)		
Id. rel. (mean) > Gender median					0.0065***	
					(0.0014)	
Id. rel. (w.mean) > Gender median						0.0043***
						(0.0014)
Reading Passage FE	✓	✓	✓	✓	✓	✓
Gender-Grade FE	✓	✓	✓	✓	✓	✓
N of student–passages	58,940,948	58,940,948	59,305,495	59,305,495	59,305,495	59,305,495

Notes: Each specification is a regression of the share of items answered correctly on a passage on gender-based identity relatability. Standard errors reported in parentheses below coefficient estimates are clustered by exam. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$. Identity relatability is defined as share of characters matching the students gender. All columns calculate identity relatability using gender labels from a large language model, imputing from pronouns and other context clues. The estimation sample includes students grades 3 to 8 from 2013–2019 taking the standard STAAR reading-comprehension test.

Table A13: Heterogeneity in the impact of race-based identity relatability on test performance

	(1)	(2)	(3)	(4)	(5)
Race identity rel. (mean)	0.0115**	0.0120***	0.0126***	0.0119***	0.0104**
	(0.0044)	(0.0043)	(0.0043)	(0.0043)	(0.0043)
Race identity rel. (mean) × Historic	-0.0032				
	(0.0030)				
Race identity rel. (mean) × Celebrity		-0.0047			
		(0.0052)			
Race identity rel. (mean) × Famous			-0.0052		
			(0.0037)		
Race identity rel. (mean) × Any children				-0.0050*	
				(0.0028)	
Race identity rel. (mean) × Memoir					0.0011
					(0.0026)
Reading Passage FE	✓	✓	✓	✓	✓
Race-Grade FE	✓	✓	✓	✓	✓
N of student–passages	59,305,495	59,305,495	59,305,495	59,305,495	59,305,495

Notes: Each specification is a regression of the share of items answered correctly on a passage on race-based identity relatability. Details of additional variables are described in Section B.6. Standard errors reported in parentheses below coefficient estimates are clustered by exam. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$. Identity relatability is defined as share of characters matching the students race. The racial prediction model is using last names from social security data and first names from six Southern states. It uses a Bayesian prior distribution matching national demographics in 2020. The estimation sample includes students grades 3 to 8 from 2013–2019 taking the standard STAAR reading-comprehension test.

Table A14: Heterogeneity in the impact of gender-based identity relatability on test performance

	(1)	(2)	(3)	(4)	(5)
Gender identity rel. (mean)	0.0041**	0.0060***	0.0039*	0.0076***	0.0108***
	(0.0018)	(0.0018)	(0.0019)	(0.0023)	(0.0025)
Gender identity rel. (mean) × Historic	0.0133***				
	(0.0044)				
Gender identity rel. (mean) × Celebrity		0.0075*			
		(0.0039)			
Gender identity rel. (mean) × Famous			0.0114***		
			(0.0039)		
Gender identity rel. (mean) × Any children				-0.0001	
				(0.0034)	
Gender identity rel. (mean) × Memoir					-0.0086**
					(0.0040)
Reading Passage FE	✓	✓	✓	✓	✓
Gender-Grade FE	✓	✓	✓	✓	✓
N of student–passages	59,305,495	59,305,495	59,305,495	59,305,495	59,305,495

Notes: Each specification is a regression of the share of items answered correctly on a passage on gender-based identity relatability. Details of additional variables are described in Section B.6. Standard errors reported in parentheses below coefficient estimates are clustered by exam. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$. Identity relatability is defined as share of characters matching the students gender. All columns calculate identity relatability using gender labels from a large language model, imputing from pronouns and other context clues. The estimation sample includes students grades 3 to 8 from 2013–2019 taking the standard STAAR reading-comprehension test.

Table A15: Impact of survey relatability on test performance

	(1)	(2)	(3)
Race survey relatability	0.0119***		0.0119***
	(0.0041)		(0.0042)
Gender survey relatability		0.0139***	0.0139***
		(0.0021)	(0.0021)
Black-white rel. diff.	-0.0038		-0.0038
	(0.0013)		(0.0013)
Hispanic-white rel. diff.	-0.0037		-0.0037
	(0.0013)		(0.0013)
Male-female rel. diff.		0.0025	0.0025
		(0.0004)	(0.0004)
Race-Grade FE	✓		
Gender-Grade FE		✓	
Race-Gender-Grade FE			✓
Passage FE	✓	✓	✓
N of student-passages	64,352,860	64,352,860	64,352,860

Notes: Each specification is a regression of the share of items answered correctly on a passage on survey relatability. Standard errors reported in parentheses below coefficient estimates are clustered by exam. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$. Race survey relatability and gender survey relatability are created using separate survey questions eliciting assessments of relatability for a passage. Coefficient estimates (and standard errors) alternatively scaled by racial and gender differences in survey relatability are displayed in the bottom rows. The full estimation sample includes students grades 3 to 8 from 2013–2019 taking the standard STAAR reading-comprehension test.

Table A16: Correlation across race relatability measures

		Topic		Identity		Survey
	(1)	(2)	(3)	(4)	(5)	(6)
Topic relatability					-0.433	0.501
					(0.339)	(0.340)
Identity relatability	-0.0151		-0.0206			0.435***
	(0.0174)		(0.0170)			(0.106)
Survey relatability		0.00546	0.0129	0.232***	0.236***	
		(0.00923)	(0.00873)	(0.0575)	(0.0569)	
Within FE R^2	0.005	0.001	0.012	0.099	0.107	0.105

Notes: Each specification is a regression of one race relatability measure on other race relatability measures. Each specification includes race-grade fixed effects and passage fixed effects. Standard errors reported in parentheses below coefficient estimates are (a) clustered by exam. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$. Within fixed effect R^2 are displayed for each specification. The estimation sample includes students grades 3 to 8 from 2013–2019 taking the standard STAAR reading-comprehension test which contains characters.

Table A17: Correlation across gender relatability measures

	Topic			Identity		Survey
	(1)	(2)	(3)	(4)	(5)	(6)
Topic relatability					0.281	0.236
					(0.445)	(0.341)
Identity relatability	0.0133		0.00903			0.563***
	(0.0142)		(0.0141)			(0.0571)
Survey relatability		0.0153	0.00749	0.559***	0.556***	
		(0.0121)	(0.0123)	(0.0421)	(0.0434)	
Within FE R^2	0.008	0.010	0.010	0.317	0.319	0.318

Notes: Each specification is a regression of one gender relatability measure on other gender relatability measures. Each specification includes race-grade fixed effects and passage fixed effects. Standard errors reported in parentheses below coefficient estimates are (a) clustered by exam. * $p < .10$, ** $p < .05$, *** $p < .01$. Within fixed effect R^2 are displayed for each specification. The estimation sample includes students grades 3 to 8 from 2013–2019 taking the standard STAAR reading-comprehension test which contains characters.

Table A18: Effect of balancing relatability across tests on share of Black students meeting performance standards

	Grade 3	4	5	6	7	8
Below Approaches (pp)	-0.004	-0.003	-0.002	-0.000	-0.003	-0.001
Below Meets	-0.005	-0.003	-0.003	-0.000	-0.004	-0.001
Below Masters	-0.005	-0.003	-0.003	-0.000	-0.004	-0.001

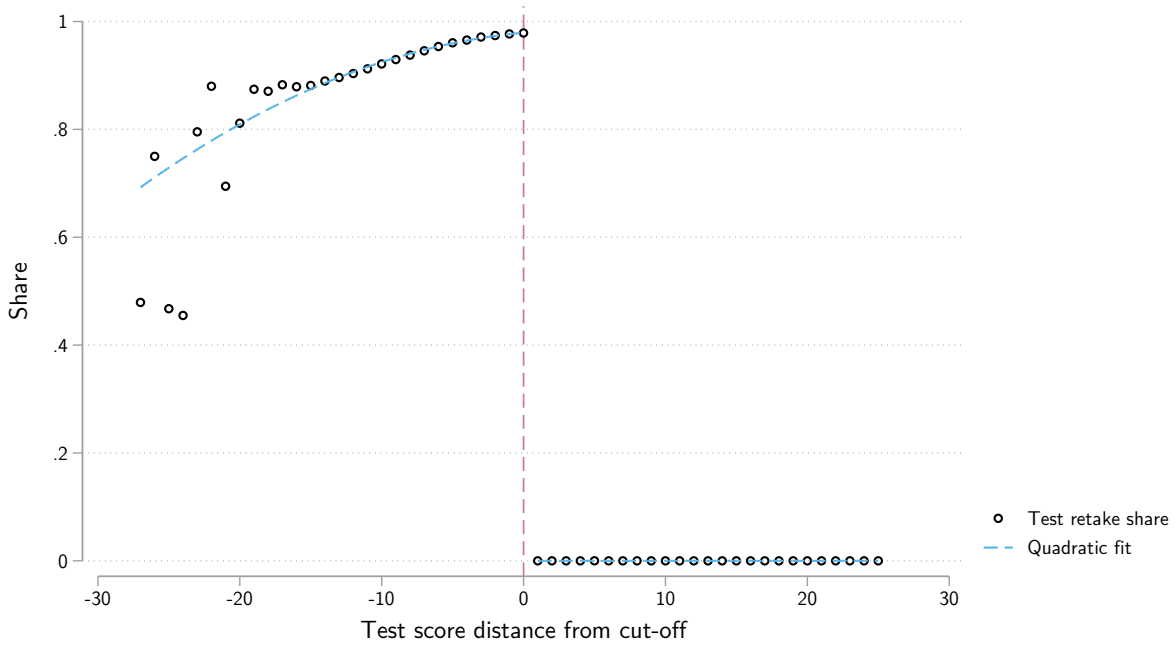
Notes: Numbers show the change in the share of students below a given performance standards category by rescaling relatability to a benchmark test within the grade. Observations are at the test-level. The sample includes all students grades 3 to 8 from 2013–2019 taking the standard STAAR reading-comprehension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

Table A19: Effect of balancing relatability across tests on share of Hispanic students meeting performance standards

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

Notes: Numbers show the change in the share of students below a given performance standards category by rescaling relatability to a benchmark test within the grade. Observations are at the test-level. The sample includes 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.

Figure A7: 5th and 8th test retake rates by distance from cut-off



Notes: This graph plots the share of 5th and 8th grade students within each score that retook the reading-comprehension exam. A student is considered to retake an exam if they take the initial reading-comprehension exam and subsequently takes the next available reading-comprehension exam in the same school year. The score is defined as the number of questions the student got correct on the first exam minus the TEA-designated cut-off for satisfactory performance on the exam, such that a positive score means the student performed satisfactorily on the exam and a negative score means the student did not perform satisfactorily on the exam. The dashed line plots a quadratic fit on either side of the cut-off. The sample includes all students grades 5 and 8 from 2013–2019 taking the standard STAAR reading-comprehension test, excluding years and grades where test retakes were not offered.

B Construction of variables

B.1 Passage example of topic- and identity-relatability inputs

We provide a concrete illustration of how we generate the topic salience and character demographic predictions from an actual STAAR reading passage. There is a passage titled "Sink or Sail" in the 2014 exam for grade 7, which originally appeared in the June 2010 edition of *OWL Magazine*.[24] This article presents a debate between two student readers about whether 14-year-old Laura Dekker from the Netherlands should be allowed to sail around the world solo. Arguments are made by Luka, age 12, who supports allowing Laura to sail given her extensive experience, and Maya, age 13, who believes that teenagers should wait until age 18 to attempt such voyages. The passage also contains exhibits detailing Laura's sailing experience, her boat specifications, and her planned two-year voyage across multiple oceans with a support team shadowing her for safety.

Test takers are asked to answer 7 questions related to the content in this passage. These questions range in the skills they seek to assess. Given that the passage provides opposing viewpoints from two sides, some questions ask students to assess why Luka or why Maya included a certain element in their respective arguments. Another question asks students to distill the common point of agreement between the two presented arguments. Still others ask students about specific concepts such as: (a) identifying an example of "stereotyping" from Maya's argument and (b) discerning the meaning of "shadowing" as it is used in the context of the passage.

In this passage, the *water sports* topic receives the highest topic salience score ($m_{t,p}$=0.084), reflecting terms such as "boat," "sail," "sailing," "ocean," and "wave" that appear throughout the text. The *traveling* topic receives a topic salience score of 0.018, triggered by words like "voyage" and "journey" that describe Laura's planned circumnavigation. Several other topics appear with very low salience scores (between 0 and 0.01), some correctly and others

[24] The PDF of this text may be found online at https://web.archive.org/web/20220703062658/https://tea.texas.gov/student assessment/testing/staar/staar-released-test-questions.

incorrectly. [25]

The passage features five named characters. Laura Dekker is predicted to be 94% White, 3% Hispanic, and less than 1% Black or Asian, placing her name at the 87th percentile of the White name distribution. Maya, one of the student debaters, provides an example of a name with relatively broad demographic appeal: she is predicted to be 29% Black, 22% White, 18% Asian, and 12% Hispanic, with her name appearing in the upper portions of multiple racial groups' name distributions rather than being concentrated in a single group. Luka, the other student debater, is predicted to be 63% White and 10% Asian. Jesse Martin and Michael Perham (previous record holders in youth sailing mentioned for comparison) are predicted to be 79% and 94% White, respectively.

B.2 Additional race topic-exposure measures

In addition to our standard measure of race-based topic exposure, we construct additional measures through various changes to the process described in section 3.2.2. We first make various restrictions to the ATUS sample. We then aggregate data from time-use diaries differently.

We create subsamples of the ATUS sample across five dimensions: (a) weekend/weekday survey response, (b) U.S. Census region, (c) state, (d) age, and (e) 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

[25] For instance, *computer* is correctly triggered by mentions of "internet" in the description of Laura's planned distance education program. However, *basketball* is incorrectly triggered because the term "court" appears when discussing Dutch judicial courts considering whether Laura should be allowed to travel solo.

in the "South" U.S. Census Region which includes Texas.[26] We also create a Texans-only sample which has low respondent counts but potentially even higher alignment to Texans interest and familiarity. Third, there may be differences across age in activity participation. We create the following four subsamples by age: 15-34, 35-49, 50-64, and 65-85. The four subsamples are mutually exclusive and 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 topic-relatability measure, r. We ultimately use 5 alternative relatability measures: $r^{weekend}$, r^{south}, r^{texas}, r^{parent}, $r^{age15-34}$.

We next create new measures of race-based topic exposure aggregating time-use data differently. Previously, we calcualted exposure e_{dt} 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 can think of this as an extensive margin measure of topic exposure. Alternatively, we can make use of the intensity of the time spent related to a particular topic.

We create four alternative "intensive margin" measures of topic exposure. Our first measure uses an average of minutes spent in a particular topic for a demographic group

$$e_{dt}^{int_simp} = \frac{1}{W_d} \sum_{j \in d} W_j time_{jt},$$

where W_j is the survey weight for respondent j and W_d is the total survey weight for respondents in group d. Our next two measures transform minutes spent using an inverse

[26] 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.

hyperbolic sine function and a square root function prior to computing group-level averages:

$$e_{dt}^{int_sine} = \frac{1}{W_d} \sum_{j \in d} W_j \operatorname{arcsinh}(time_{jt})$$
$$e_{dt}^{int_sqrt} = \frac{1}{W_d} \sum_{j \in d} W_j \sqrt{time_{jt}}$$

Finally, we construct a measure based on the average percentile of minutes spent in a particular topic category:

$$e_{dt}^{int_tile} = \frac{1}{W_d} \sum_{j \in d} W_j PercentileRank_t(time_{jt})$$

where $PercentileRank_t(time_{jt})$ is the population-level percentile rank of time spent in topic t for respondent j.

This process yields four additional race-based topic-relatability measures when combined with our baseline topic-salience measures.

B.3 Measures of topic salience

Formally, our baseline measure is calculated as follows. 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:

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

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," using only nouns, removing words from the dictionary, and expanding the dictionaries to include all one-word synonyms of the original words in a given topic dictionary.[27] 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.4 PSID topic relatability

We use the Child Development Supplement (CDS) of the Panel Study of Income Dynamics (PSID) to create an alternative measure of topic relatability. The PSID is a longitudinal survey of U.S. families. Families were selected and interviewed in 1968, and every subsequent wave has mostly included individuals from these families and their descendents. The CDS component of the PSID, which focuses on information regarding children, was launched in its initial form in 1997 and updated in 2014. The CDS includes children's daily time diaries from one weekday and one weekend, filled out by the caregiver or the child, depending on the age of the child.

We collect data from the 2014 and 2019 waves of the PSID-CDS. As our measure of a child's race/ethnicity, we primarily use child's reported race from the parent(s). If both parents' reports are aligned, the child is assigned that race/ethnicity. If there is a report from only one parent, the child is assigned the race/ethnicity that is reported by that parent.

[27] 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 *running*, *runs*, and *ran* to the stem *run*.

If parents disagree, they are coded as mixed race. We use the child's report of their race only for children that are missing race/ethnicity characterizations. We define race as we have throughout this paper, with Hispanic as its own "race" and all other racial groups being non-Hispanic. We restrict our analysis to children between the ages of 8 and 17.

We construct the topic exposure much in the same as we do for the ATUS data. We assign each PSID activity code to a topic category t. We observe a single weekday diary and a single weekend diary for the vast majority of children in the CDS. We opt to combine these minutes for each activity. Then, we calculate PSID topic exposure as the fraction of children of demographic group who report any activity related to topic $t \in \mathcal{T}$, using sampling weights provided by the PSID-CDS.

Table B1 provides summary statistics for the PSID sample. Comparing the PSID sample to the ATUS sample summary (displayed in Table 2), we find that children on average do a higher number of activities than adults do. This may reflect the difference between the nature of children's time use and adult's time use, or this may reflect differences in activity coding across the PSID and the ATUS. Children engage in more leisure activity than adults; on average, children in the sample are exposed to one additional topic compared to adults and spend 30% of their time in these topics compared with 20% for adults. We find under-sampling of White and female respondents compared to the ATUS sample.

The demographic distribution in the weighted PSID sample masks significant imbalance across racial groups in the underlying observations. The weighted ATUS sample and the unweighted ATUS sample are fairly similar in its racial composition, with deviations fewer than 2 percentage points for all racial groups. By contrast, the unweighted PSID sample deviates significantly from the weighted PSID sample. Only 15 Asian children are included in the PSID sample, making up 1% of the observations. Hispanic children are 9% of the raw sample, totaling 134 observations. By contrast, Black children make up 42% of the sample. These differences reflect the structure of the PSID, which uses descendants of the 1968 initial cohort as its main study sample. While recent waves of the PSID have added

Table B1: Summary statistics for PSID-CDS sample

	Mean	SD	5%	95%
Activities				
Number of activities	17.33	4.73	9	25
Number of min spent per done activity	88.74	30.38	55	144
Number of topics	3.42	1.59	1	6
Share of time spent in topics	0.30	0.16	0.08	0.60
Demographics				
Asian	0.02			
Black	0.14			
Hispanic	0.16			
White	0.58			
Female	0.47			
N	1,555			

Notes: The table displays sample means, standard deviations, and the 5th/95th percentile value for each category. Respondent observations are weighted by PSID-CDS sampling weights provided by the Panel Study of Income Dynamics. The sample includes PSID-CDS respondents from 2014 and 2019 aged between 8 and 17.

additional racial groups, they still do not fully reflect the experiences of recently immigrated populations to the U.S.

B.5 Leveraging non-race- and non-gender-based topic exposure

The ATUS data and the student data have additional measures of individual demographic characteristics beyond race and gender such as economic disadvantage. We outline below a process which allows us to leverage socioeconomic status variation to construct new measures of exposure and relatability.

We start by bringing SES definitions in the ATUS in close concordance to the level 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 low SES. 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 this ambiguity. This may result in sharper topic-exposure differences across SES than may be true in the actual population.

In our analysis, we use a measure of race-SES topic relatability. The topic-exposure basis for this variable is generated in a procedure much like before. We categorize respondents into groups $c \in \mathcal{C}$ which is the race-SES status of each respondent. Then e_{ct} is simply the share of respondents in group c which reported any number of minutes participating in activities related to topic 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.

B.6 Measures of identity relatability

Named entity recognition and character extraction. We use OpenAI's GPT-5-mini (2025-08-07 release) to extract named characters from each passage through structured prompts that request specific attributes for each person mentioned by name. We verify robustness using Claude Sonnet 4 (20250514 release from Anthropic). The LLM identifies all human persons referenced with either a first or last name, excluding non-human entities, individuals referenced only by role or title (e.g., "Mom," "the teacher"), and names appearing solely in copyright attributions or photo credits. For each character retained, we record their first name, last name, any title or honorific, the number of times mentioned by name,

and the specific text snippets where they appear by name. We also ask the LLM to classify the character's gender based on context clues (for example, pronouns or gendered titles such as "Mr." or "Mrs.").

We ask the LLM for a few additional attributes of each character, which we leverage for the heterogeneity analyses in Tables A13 & A14. We ask the LLM whether the character is a historical figure, whether they might be considered a celebrity to a grade-school student, whether the character is a child or adult, and whether the character is an author. Passages are then classified as Historic or Celebrity if they contain *any* character flagged with that attribute, and Famous is the union of the two. A passage is labeled to have "Any Children" if at least one character is classified as a child. The Memoir and No author labels are for passages that either have an author that is mentioned by name more than once or do not have any author detected by the LLM at all.

Racial prediction. For characters with either a first or last names, we predict race using the Bayesian Improved Surname Geocoding (BISG) method available via the `wru` R package. This approach combines surname data released by the US Census Bureau with first-name data from voter registration files covering approximately 37.8 million registered voters across six Southern states (Alabama, Florida, Georgia, Louisiana, North Carolina, and South Carolina). These voter files are aggregated by L2 Inc. and sourced between October 2020 and February 2021 (Imai et al. 2022). The method produces probabilistic predictions across racial categories (Asian, Black, Hispanic, White). We apply two different Bayesian priors in our robustness checks, both derived from the 2020 Census: one based on population shares for the entire United States and another based on Texas-specific population shares (see Table A11).

C Validation of topic salience

We validate our topic-salience measure 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.[28] 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.

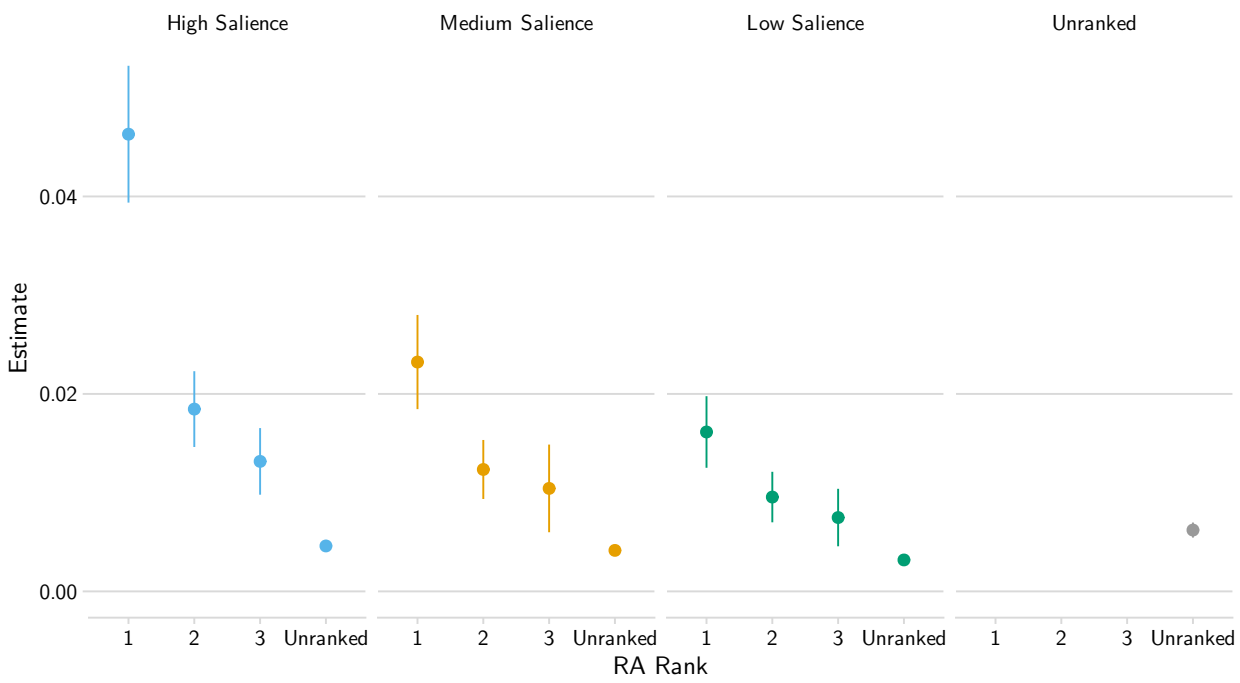
Now, consider how our topic-salience measure m_{tp} should relate to the human labeling results. We would expect that when a research assistant ranks topics t_1, t_2, t_3, respectively, as the 1st, 2nd and 3rd most salient topics in a passage, that $m_{t_1p} \geq m_{t_2p} \geq m_{t_3p}$. Further, if three passages p_a, p_b, p_c are labeled as a high, medium and low salience passage with respect to top-ranked topic t, we would expect that $m_{tp_a} \geq m_{tp_b} \geq m_{tp_c}$.

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

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

[28] 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.[29] 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}:

$$\bar{v}_{tp}^{\perp} = \frac{\sum_d w_{dp} e_{dt} v_{dp}^{\perp}}{\sum_d w_{dp} e_{dt}} \tag{9}$$

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

[29] 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

$$\bar{Y}_{tp}^{\perp} = \varepsilon + \beta \bar{r}_{tp}^{\perp} + q'_{tp}\psi + \zeta_{tp} \tag{10}$$

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. Crucially, while we have already orthogonalized Y_{dp} and r_{dp} with race-grade fixed effects and passage-by-exposure sum fixed effects, we require topic-grade fixed effects and passage fixed effects to purge this variation in m_{tp}.

E Robustness

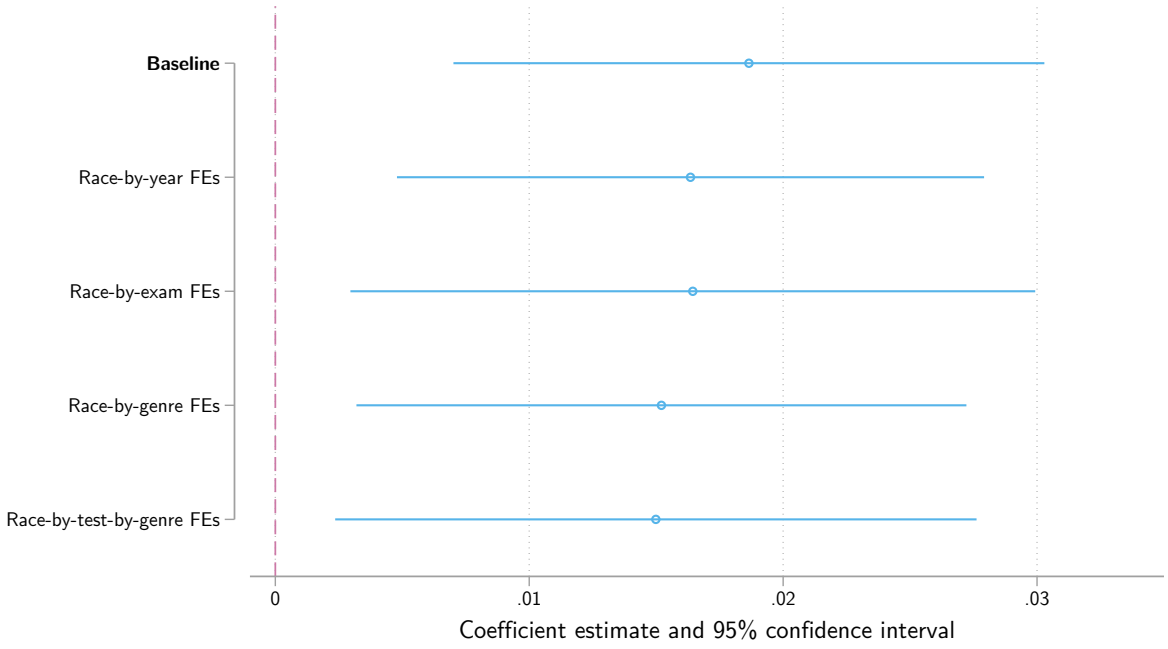
Our estimates are downstream of modeling and data choices that are baked into our race-based relatability measure and model specification. First, we show that the estimates are qualitatively the same with alternative fixed effect specifications. Second, we show that our results are not sensitive to specific 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.

Fixed effects 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. Using race-by-genre fixed effects, which is conditioning on passage category ("literary", "informational", or "mixed"), also results in a qualitatively similar coefficient, while making the comparison set less restrictive.

Leave-one-out Examining the relatability measure, we 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 outliers that swing down our estimates the most (relative to the baseline) are a maximum of a 30% decrease in the point estimates, and these differences are not statistically significant.

Topic salience 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.3 for details). Further, we consider a variety of changes to the NLP data processing steps such as leaving the words unstemmed,

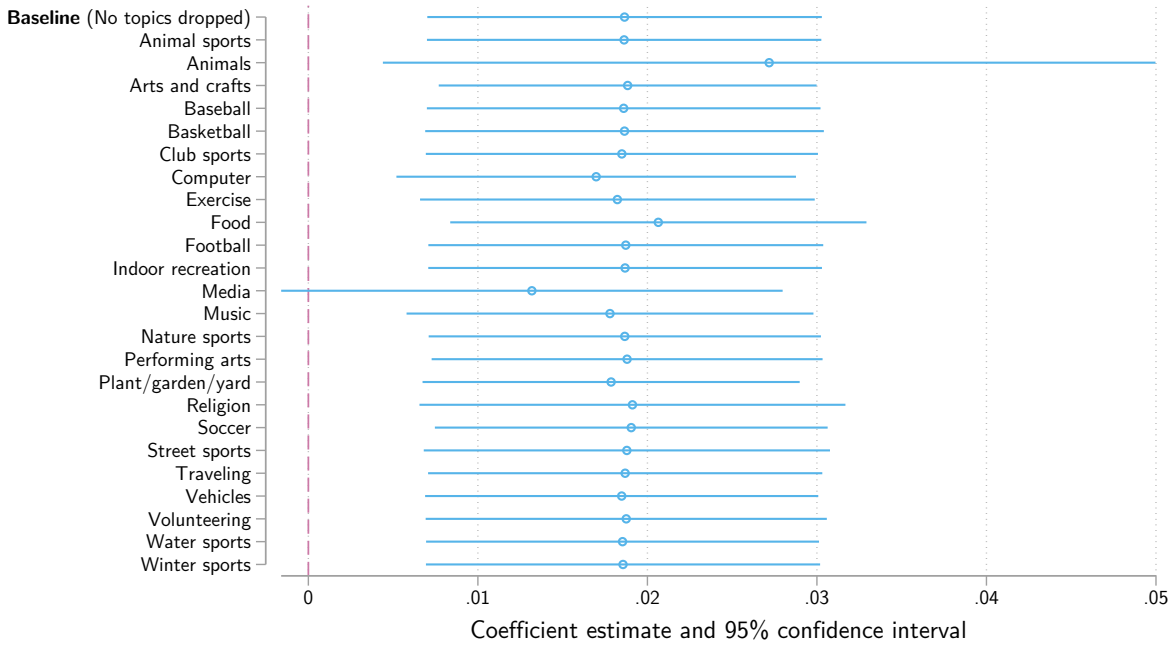
Figure E1: Robustness of baseline effect to different levels of saturation in the specifications of fixed effects



Notes: Each coefficient estimate is from a separate specification which regresses the share of items answered correctly on a passage on race-based topic relatability. Unreported controls include: (1) "unit" fixed effects at the level indicated in the legend and (2) exposure-sum-by-passage fixed effects. The baseline specification includes unit fixed effects at the race-by-grade level. Exposure-robust standard errors clustered by exam are used to construct the 95% confidence intervals. Observations are at the race-passage level and are weighted by the number of student-items. The estimation sample includes all students grades 3 to 8 from 2013–2019 taking the standard STAAR reading-comprehension test.

using only the nouns, expanding the dictionaries to include all one-word synonyms of original words for each topic, and discretizing the shocks. We see in Figure E3 that the results are qualitatively similar, with all significant except for the discretized shocks which is relying only on the extensive margin variation and ignoring the intensive margin variation that the NLP methods 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.

Figure E2: Leave-one-out topic-relatability estimates

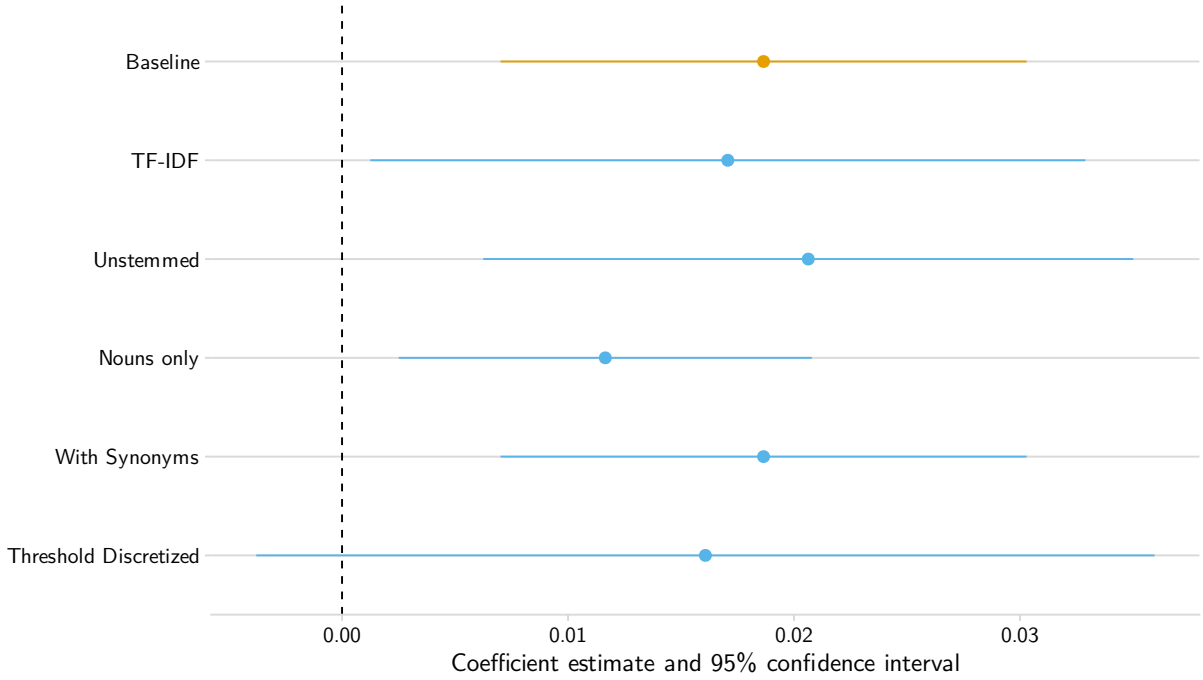


Notes: Each coefficient estimate is from a separate specification which regresses the share of items answered correctly on a passage on race-based topic relatability. Unreported controls include: (1) race-grade fixed effects and (2) exposure-sum-by-passage fixed effects. Each estimate, aside from the baseline estimate, uses a measure of topic relatability which excludes the indicated topic in the calculation process. Exposure-robust standard errors clustered by exam are used to construct the 95% confidence intervals. Observations are at the race-passage level and are weighted by the number of student-items. The estimation sample includes all students grades 3 to 8 from 2013–2019 taking the standard STAAR reading-comprehension test.

Topic exposure Lastly, we consider how we construct the race-based exposure measure. 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.2. We see in Figure E5 that using the youngest age group and respondents with children provides a more predictive point estimate.[30] In particular, we find the difference in coefficients between our baseline estimate and the estimate from respondents with children is statistically significant. Further, restricting to Southerners or Texans seems to slightly attenuate the estimate, and only using weekend responses has close to no difference. Nonetheless, these all these differences are relatively minor and collectively

[30]Results for other age groups are qualitative similar to the baseline estimate.

Figure E3: Robustness to different NLP measures for topic salience

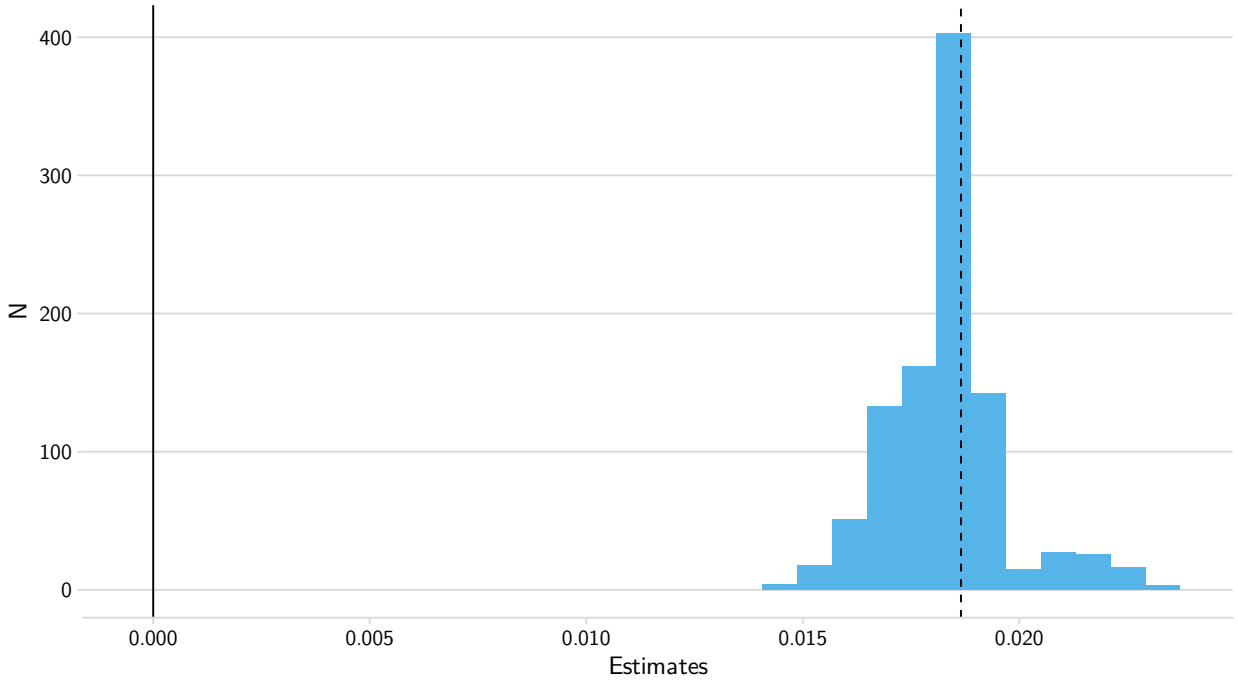


Notes: Each coefficient estimate is from a separate specification which regresses the share of items answered correctly on a passage on race-based topic relatability. Unreported controls include: (1) race-grade fixed effects and (2) exposure-sum-by-passage fixed effects. The "TF-IDF" specification uses inverse document frequency weights to calculate the topic salience. The "unstemmed" specification uses unstemmed words when matching between the dictionaries and passages. The "Nouns only" specification filters down to nouns when matching between the dictionaries and passages. The "With Synonyms" expands each dictionary to include all one-word entries in a thesaurus. Lastly, the "Threshold Discretized" specification sets the top 1/24th topic-salience values to 1 and the rest to zero, within grade-level. Further details on variable construction can be found in Appendix Section B.3. Exposure-robust standard errors clustered by exam are used to construct the 95% confidence intervals. Observations are at the race-passage level and are weighted by the number of student-items. The estimation sample includes all students grades 3 to 8 from 2013–2019 taking the standard STAAR reading-comprehension test.

point to affirming the baseline estimates that just use the ATUS data as-is.

We also consider using "intensive margin" topic-exposure measures when constructing relatability to make full use of the entire distribution of time-use data available in the ATUS data. We exercise appropriate caution in creating these measures, as we believe these measures capture a different attribute of individual relationships to topics. First, while more time spent in an activity is positively correlated with familiarity or interest in that activity, the precise relationship between those objects is unlikely to be linear or follow a standard

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

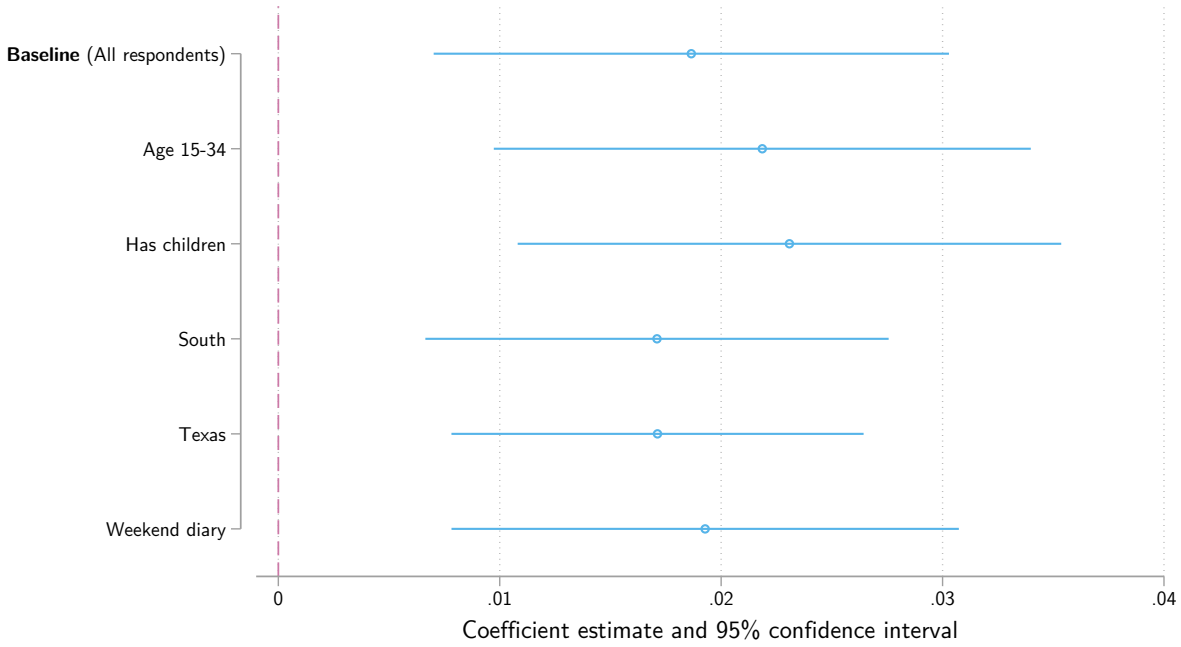


Notes: Each of the 1000 observations is a separate regression of the share of items answered correctly on a passage on race-based topic relatability. Unreported controls include: (1) race-grade fixed effects and (2) exposure-sum-by-passage fixed effects. Each regression uses a different estimate of topic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

functional form. For example, a basic level of interest and familiarity may be associated with a wide range of minutes spent, with differentiation only occurring at the tail ends of minutes spent. Second, the time spent distribution will vary widely across topics. An activity such as golf will be characterized by extreme values of time spent (0 minutes and 180+ minutes, for example) while an activity such as TV viewing will follow a smoother distribution. This raises scaling concerns across activities.

With these points in mind, we construct four alternative measures of topic exposure, which is detailed in section B.2. Figure E6 shows the coefficient estimates using these new measures. We see that all intensive margin measures show smaller coefficients than the main relatability measures. However, the effect sizes are dependent on the functional form which

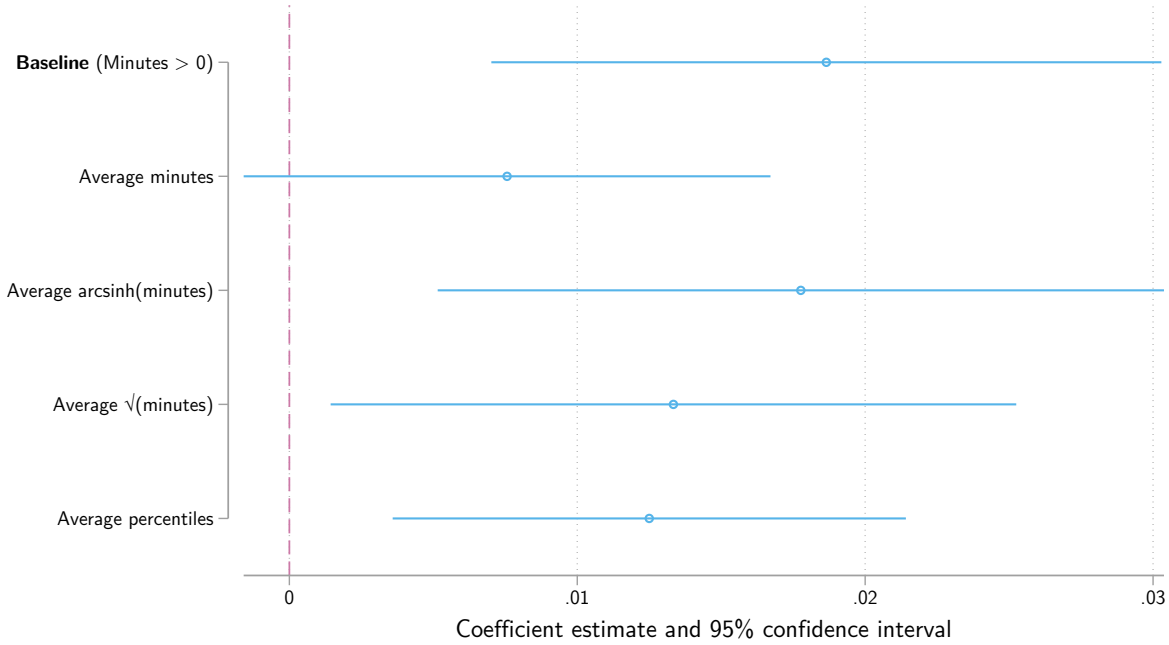
Figure E5: Robustness to different ATUS samples for calculating topic exposure



Notes: Each coefficient estimate is from a separate specification which regresses the share of items answered correctly on a passage on race-based topic relatability. Unreported controls include: (1) race-grade fixed effects and (2) exposure-sum-by-passage fixed effects. Each estimate uses a different ATUS sample in calculating topic exposure. The baseline measure makes no restriction to the ATUS sample. "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 Region. "Texas" corresponds to respondents in Texas. "Weekend diary" corresponds to using only the ATUS sample by which respondents' diary day is on the weekend. Exposure-robust standard errors clustered by exam are used to construct the 95% confidence intervals. Observations are at the race-passage level and are weighted by the number of student-items. The estimation sample includes all students grades 3 to 8 from 2013–2019 taking the standard STAAR reading-comprehension test.

converts minutes spent into topic exposure. As suspected, a similar linear aggregation of minutes spent introduces noise in our topic-exposure estimates and downwardly biases our coefficient estimates. Utilizing a more concave functional form results in estimates that are much more similar in magnitude to our main results.

Figure E6: Robustness to different ways of aggregating ATUS time diaries



Notes: Each coefficient estimate is from a separate specification which regresses the share of items answered correctly on a passage on race-based topic relatability. Unreported controls include: (1) race-grade fixed effects and (2) exposure-sum-by-passage fixed effects. Each estimate uses a different methodology to aggregate ATUS when calculating race-based topic relatability. The baseline estimate uses the share of a racial group having non-zero minutes. "Average minutes" uses an average of minutes within a racial group. "Average arcsinh(minutes)" uses an average of the inverse hyperbolic sine of minutes within a racial group. "Average $\sqrt{}$(minutes)" uses an average of the square root of minutes within a racial group. "Average percentiles" first calculates the sample-level percentile of topic minutes and computes the average of the percentile within a racial group. Exposure-robust standard errors clustered by exam are used to construct the 95% confidence intervals. Observations are at the race-passage level and are weighted by the number of student-items. The estimation sample includes all students grades 3 to 8 from 2013–2019 taking the standard STAAR reading-comprehension test.

F Details on survey relatability

This appendix section contains detailed information on participant recruitment, participant characteristics, and prcoedures for the survey which generates our survey relatability measure.

F.1 Recruitment

Respondents were recruited on the online study platform Prolific. The study was titled "Relatability of different text," with a listed base payment rate of 3 U.S. dollars for approximately 20 minutes of their time. The participants saw the following recruitment text:

> The survey consists of reading text and answering a handful questions. We estimate it takes no longer than 20 minutes to complete.
>
> Participants can earn up to $4 by completing the survey and correctly answering simple comprehension questions.
>
> There are no requirements for taking part in this study, simply answer the questions as honestly as you can.
>
> Thank you for your interest.

Once they agreed to participate in the study, participants were informed of all payment and participation rules at the beginning of the survey. In particular, participants were told they could voluntarily leave the survey at any point. Ultimately, 1,066 respondents provided partial or full responses to our survey.

The respondent pool is selected to be representative of the United States adult population. We rely on the Prolific platform's built in representative sample screener. Prolific uses age by sex by race proportions calculated from U.S. Census Bureau estimates from 2021. The age groups are: 18–24, 25–34, 35–44, 45-54, and 55–100. The races are: Asian, Black, Mixed, Other, and White. The platform did not allow screening for Hispanic respondents.

F.2 Survey procedures

Respondents were asked a series of demographic questions at the beginning of the survey. Respondents were asked to provide basic information on their gender, race, ethnicity, and age. We also asked respondents to provide their state of residence and whether they received free- or reduced-price lunch while in school. Finally, respondents were asked about their interactions with children in their daily lives. This included questions about whether they are a parent or guardian, the frequency with which they interact with children other than their own direct children, and the demographic characteristics of the children with which they interact.

Respondents were asked to read and answer questions regarding 10 pieces of text. Each piece of text corresponded to a summarized passage in our test passage sample. Passages were summarized using Claude 3.5 Sonnet, an LLM developed by Anthropic. The LLM was told to act as an "an expert literary analyst tasked with creating informative summaries of various passages." For each passage, the LLM was prompted to carefully read the passage, analyze the passage, and create a summary of the passage. Summaries were required to meet the following criteria:

- Be approximately 220 words long
- "Capture the essence of the passage without revealing every detail"
- Do not include commentary on the passage
- Retain the style and tone of the original passage
- Write at an 8th-grade reading level

Each summarization criteria was developed through testing the model with a variety of prompts and sample passages. We set a word count target to make sure that passage length is not a significant contributor to respondents' assessments of relatability. However, we are not prescriptive about this target as (a) we wanted to allow for word count flexibility in

creating a high-quality summary, and (b) we found that our LLM was not generally well-equipped to hit a hard word count target. We also found that sometimes the LLM would over-emphasize certain details in a passage over others. While we observed no overall pattern in this emphasis, our final prompt prevents the model from becoming overally detailed given the short summary length. In piloting, the LLM had a tendency to include commentary on the passage itself, commenting on the morals or lessons that could be gleaned from the story. Our prompt focuses the LLM to provide only direct summarization gleaned from the text. Finally, we give the LLM a target, uniform reading level so that the perceived difficulty of the passage also does not sway respondents when assessing relatability.

After reading a summary, respondents were asked to assess the relatability of the passage for racial groups and gender separately. To focus respondents' attention, we provided a basic relatability definition at the beginning of the survey. The definition reads as follows:

> We will primarily ask you to assess these summaries based on how ***relatable*** *you think they are to children.* A summary or story is considered relatable to a child if he/she (a) is familiar with topics in the text, (b) has interest in the text content, and/or (c) is represented by characters in the text.

The respondents were also told that "when answering these questions, feel free to consider children you know, children that you grew up with, or your own experiences as a child." Respondents were asked two separate questions, one for race and one for gender. For race, respondents were asked:

> Please rank the following groups by how relatable this summarized story would be to members of each group. Assign a unique rank from 1(most relatable) to 4 (least relatable).

The four options provided were "Asian children", "Black children", "Hispanic children", "White children." For gender, respondents were asked:

> Which group would find this summarized story more relatable: boys or girls?

Repsondents also were asked to provide a personal rating of relatability to "children with similar childhood experiences as mine."

With surveys, one may be concerned about low quality responses due to respondent fatigue or low effort. These concerns are especially true in our current survey structure where respondents are asked to perform the same task multiple times. We address these concerns in several ways. First, we randomized both the set of passages given to a respondent and the order in which these passages are presented. We also randomize the order in which the racial groups are presented in the race question. Together, these question randomization adjustments ensure that if respondent quality declines throughout the survey or if respondents provide low effort responses, they are distributed randomly throughout passages and demographic group assessments. Second, we include an incentive for respondents based on successfully answering a reading-comprehension question related to the summary. Respondents answering this question correctly receives an additional \$0.10, which means respondents can earn up to \$1 if they successfully answer all 10 comprehension questions. reading-comprehension questions are created using the same LLM and manually validated for any errors.

F.3 Respondent and response characteristics

Table F1 summarizes the respondent sample. In comparison to U.S. Census Bureau estimates, Hispanic adults are slightly underrepresented in our sample (13% vs. $\sim$ 16%). The non-Hispanic sample is largely representative of the adult population. This is not surprising as the Prolific representative sample screener does not account for ethnicity. We find that the majority of our adult sample is between the ages of 35–64. This results in a over-representation for this age group, at the expense of individuals aged 65 or above, who are under-represented. This may also reflect differences between the age ranges collected and the Prolific screener: the oldest category in the platform's screen was age 55 and above. The Texas resident share and eligibility for free- and reduced-price lunch as a child are both roughly in line with national statistics. We also find that a large proportion of our adult respondents interact regularly with White children (71%). Proportions are significantly lower

Table F1: Survey respondents summary statistics

	Mean	SD
Race and gender		
Asian	0.06	
Black	0.12	
Hispanic	0.13	
White	0.62	
Female	0.50	
Age		
18–34	0.30	
35–64	0.58	
65+	0.12	
Other characteristics		
Texas resident	0.10	
Free- and reduced-price lunch as child	0.45	
Has children	0.49	
Demographics of children interacted with		
Asian	0.17	
Black	0.32	
Hispanic	0.28	
White	0.71	
Boys	0.76	
Girls	0.73	
Survey response statistics		
Number of passages answered	9.69	1.31
10/10 passages answered	0.90	
Minutes	27.01	13.17
Comprehension score	8.52	2.04
N	1,066	

Notes: The table displays sample means and standard deviations of respondent characteristics in the survey sample. Race includes individuals who identify as Hispanic. Asian, Black, and White exclude individuals who identify as Hispanic. "Free- and reduced-price lunch as child" includes all individuals who reported being eligible for free- and reduced-price lunch when they were children. A passage is considered as "answered" if a respondent answers all relatability-related questions related to that passage. Survey minutes are winsorized at the 1% and 99% levels, as a small portion of respondents did not formally quit our survey and allowed the survey to time out. The sample includes all respondents who answered at least one relatability-related question.

for all other racial groups.

We also show basic summary statistics on the survey responses elicited. On average, 90% of respondents answered questions for every passage and more than 95% of respondents

answered questions for at least 9 passages. Respondents on average took slightly longer than expected to finish the survey (27 minutes). The average comprehension score was 8.52 out of 10. The median score was a 9, with over two-thirds of the respondent sample receiving this score or higher.

We take the survey responses to create survey relatability measures. Our preferred measure comes from the predicted probabilties after fitting an rank-ordered logit model for each passage. Suppose we observe a set of rankings $d_{ip1} \succ d_{ip2} \succ \ldots d_{ip4}$ of racial groups for individual i and passage p. We assume latent relatability assessment is given by: $R_{ipd} = V_{ipd} + \varepsilon_{ipd}$, where ε_{ipd} is assumed to be Type I extreme value. We assume $V_{ipd} = \delta_{pd}$ where δ_{pd} are relatability assessment levels for group $d \in D$ for passage p. Then the probability of observing the ranking is

$$\Pr\left(d_{ip1} \succ d_{ip2} \succ \ldots d_{ip4} \mid \delta_p\right) = \prod_{m=1}^{3} \frac{\exp\left(\delta_{pd_{ipm}}\right)}{\sum_{h \in D \setminus \{d_{ip1}, \ldots, d_{ip,m-1}\}} \exp\left(\delta_{ph}\right)},$$

that is, the probability of observing the ranking is the probability of "preferring" d_{ip1} over all alternatives, preferring d_{ip2} for all alternatives excluding d_{ip1}, etc. We estimate δ separately for each passage, meaning the log-likelihood estimator maximizes for a given p

$$\mathcal{L}_p(\delta_p) = \sum_i \sum_{m=1}^{3} \left[\delta_{pd_{ipm}} - \log \left(\sum_{h \in D \setminus \{d_{ip1}, \ldots, d_{ip,m-1}\}} \exp\left(\delta_{ph}\right) \right) \right]. \tag{F1}$$

We estimate the δ_p that maximizes $\mathcal{L}_p$ given the rankings we observe from respondents. For identification, we impose one constraint per passage, choosing a reference group d^* and setting $\delta_{pd^*} = 0$. We note that δ_p for gender is trivially obtained since there are only two alternatives. After obtaining δ_p from maximizing equation (F1), we calculate

$$\hat{\Pr}(d \text{ is rank } 1 \mid p) = \frac{\exp(\hat{\delta}_{pd})}{\sum_{h \in D} \exp(\hat{\delta}_{ph})}, \tag{F2}$$

which is the predicted probability that racial group d is assessed to relate most to passage

Table F2: Survey relatability summary

	Asian	Black	Hispanic	White	Female	Male
Rank	2.771	2.727	2.731	1.771	1.587	1.413
	(0.495)	(0.385)	(0.384)	(0.473)	(0.312)	(0.312)
Rank = 1	0.156	0.135	0.115	0.594	0.413	0.587
	(0.169)	(0.139)	(0.143)	(0.219)	(0.312)	(0.312)
$\hat{\Pr}(d \text{ is rank } 1 \mid p)$	0.168	0.165	0.170	0.498	0.413	0.587
	(0.138)	(0.111)	(0.110)	(0.206)	(0.312)	(0.312)
N passages	205					
N respondents	640					

Notes: The table displays means of relatability rankings provided by survey respondents for each demographic group. Standard deviations are reported in parentheses. Means are calculated at the passage-level after responses are averaged within passage. For example, the mean rank for Asian is calculated after first calculating the mean rank for Asian children for each passage and then calculating the mean of this mean rank across passages. "Rank" is the raw rank provided by respondents, where 1 indicates most relatable and 4 indicates least relatable. "Rank = 1" is the share that said a passage was most relatable for a particular race. "$\hat{\Pr}(d \text{ is rank } 1 \mid p)$" is the predicted probability that racial group$d$ is assessed to relate most to passage p from maximizing the rank-ordered logit estimator specified in equation (F1). The sample includes all respondents who answered at least one relatability-related question.

p. We define $r_{dp}^{survey} = \hat{\Pr}(d \text{ is rank } 1 \mid p)$.

Table F2 provides a summary of relatability responses across passages. Respondents on average rate passages to be a full one relatability rank higher than other racial groups. Passages are four times more likely to be rated as most relatable to White children than for other racial groups. However, the aggregate differences across racial groups are more muted for the predicted probability measure. For gender, we find that passages are generally rated as being more relatable to boys than to girls.

G Details on STAAR standards analysis

Every year, the TEA sets STAAR performance standards which are meant to link STAAR results to state-mandated curriculum standards. Students are them assigned to standards categories based on their results on the STAAR tests. These categories allow educators, parents, and students to anchor better make sense of their test performance. The categories are also used to direct extra resources to students and in certain cases to prevent them from being promoted to the next grade. Figure A7 demonstrates one such outcome based on the standards categories. In Figure A7, we plot the shares of 5th grade and 8th grade reading-comprehension test takers who have to retake the test. We organize the horizontal axis by initial test score, normalizing across tests based on the prevailing "satisfactory" performance cut-off for that test. We observe a stark discontinuity around this cut-off, with 98% of students just below the cut-off retaking the exam and virtually 0% of students just above the cut-off retaking the exam.

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 academic 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. Common across these regimes is the conversion of raw test scores to scale scores to the assignment of performance standards categories.

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.

We now consider the extent to which content relatability affects the misclassification of students to standards categories. 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 topic relatability and average identity relatability for each test by race. Then, we use $\hat{\beta}$ from estimating equation (4) and $\hat{\beta}^{identity}$ from estimating equation (6) to estimate the "content relatability benefit" of each test

$$rel_benefit_{ghd} = \hat{\beta} \cdot \bar{r}_{ghd} + \hat{\beta}^{identity} \cdot \bar{r}^{identity}_{ghd}$$

where h indexes exam years and $\bar{r}_{ghd}$ is the average topic relatability for each test gh and race d. We identify a "benchmark" test within each grade with (a) the lowest Black-White difference in $rel_benefit$ and (b) the lowest Hispanic-White difference in $rel_benefit$. Let $\bar{r}_{gh^*d}$ and $\bar{r}^{identity}_{gh^*d}$ represent the average relatabilities from this benchmark exam for grade g and group d.

For each student-exam on a non-benchmark test, we calculate the corresponding counterfactual score, $\tilde{score}_i$, if the student had instead took this "least-biased" benchmark exam. To do so, we calculate the difference between content relatability for that test compared to

this benchmark test and adjust overall scores based on $\hat{\beta}$ and $\hat{\beta}^{identity}$. Formally, we calculate

$$\tilde{score}_i = score_i + N_{g(i)h(i)} \left[\hat{\beta} \left(\tilde{r}_{g(i),d(i)} - \bar{r}_i \right) + \hat{\beta}^{identity} \left(\tilde{r}^{identity}_{g(i),d(i)} - \bar{r}^{identity}_i \right) \right] \tag{G1}$$

where i indexes student-exams, $g(i)$ indicates the grade level, $d(i)$ is the student's race, and $h(i)$ indicates the calendar year of the exam. $score_i$ is the score for the student on the exam, $N_{g(i)h(i)}$ is the number of questions on the exam, and $\hat{\beta}$ is the estimated coefficient from equation 4. $\bar{r}_i$ is the average topic relatability for the student's exam and $\tilde{r}_{g(i),d(i)}$ is the adjusted average topic relatability for $d(i)$ on the benchmark exam for grade $g(i)$. We obtain $\tilde{r}_{g(i),d(i)}$ from $\bar{r}_{gh^*d}$ by applying an adjustment such that the average cross-demographic group relatability is the same between the benchmark test and the non-benchmark test. This is essential so that the number of students receiving each score remains relatively stable with the score adjustment.

We assign students to performance standards categories based on this new, relatability-adjusted score, $\tilde{score}_i$. Since raw scores can only be a whole number, students with non-integer $\tilde{score}_i$ are partially attributed to the two nearest integers. Specifically, we assign $(\lceil \tilde{score}_i \rceil - \tilde{score}_i)\%$ of a student to score $\lfloor \tilde{score}_i \rfloor$ and the rest to $\lceil \tilde{score}_i \rceil$. 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 re-assign student i to one of four performance standards categories for exam $g(i)m(i)$.