Full reference: Matta, M., Mercer, S. H., & Keller-Margulis, M. A. (2023). Implications of bias in automated writing quality scores for fair and equitable assessment decisions. *School Psychology*, *38*(3), 173–181. https://doi.org/10.1037/spq0000517

Implications of Bias in Automated Writing Quality Scores for Fair and Equitable Assessment Decisions

Michael Matta[1], Sterett H. Mercer[2], and Milena A. Keller-Margulis[1]

[1] Department of Psychological, Health & Learning Sciences, University of Houston

[2] Department of Educational and Counselling Psychology, and Special Education, The University of British Columbia

Author Note

Michael Matta https://orcid.org/0000-0003-4266-0130

Sterett H. Mercer https://orcid.org/0000-0002-7940-4221

Milena A. Keller-Margulis https://orcid.org/0000-0001-7539-5375

The present study was supported by Society for the Study of School Psychology (Grant Early Career Research Award) to Michael Matta and Institute of Education Sciences (Grant R305A190100) to Milena A. Keller-Margulis. Correspondence concerning this article should be addressed to Michael Matta, University of Houston, 403 Farish Hall, 3657 Cullen Blvd., Houston, TX 77204. Email: mmatta@uh.edu

Abstract

Recent advances in automated writing evaluation have enabled educators to use automated writing quality scores to improve assessment feasibility. However, there has been limited investigation of bias for automated writing quality scores with students from diverse racial or ethnic backgrounds. The use of biased scores could contribute to implementing unfair practices with negative consequences on student learning. The goal of this study was to investigate score bias of writeAlizer, a free and open-source automated writing evaluation program. For 421 students in grades 4 and 7 who completed a state writing exam that included composition and multiple-choice revising and editing questions, writeAlizer was used to generate automated writing quality scores for the composition section. Then, we used multiple regression models to investigate whether writeAlizer scores demonstrated differential predictions of the composition and overall scores on the state-mandated writing exam for students from different racial or ethnic groups. No evidence of bias for automated scores was observed. However, after controlling for automated scores in grade 4, we found statistically significant group differences in regression models predicting overall state test scores three years later but not the essay composition scores. We hypothesize that the multiple-choice revising and editing sections, rather than the scoring approach used for the essay portion, introduced construct-irrelevant variance and might lead to differential performance among groups. Implications for assessment development and score use are discussed.

Keywords: Automated text evaluation; scoring bias; predictive bias; diagnostic bias; social consequences of validity.

Impact and implications statement

Although computer programs can improve the feasibility of writing assessment, they may also exacerbate pre-existing societal biases. Findings show that automated writing quality scores do not lead to differential predictions of student performance for different racial and ethnic groups. Rather, test sections assessing writing less authentically (i.e., multiple-choice revising and editing questions) might be responsible for group differences and unfair educational decisions.

Implications of Bias in Automated Writing Quality Scores for Fair and Equitable Assessment Decisions

The availability of sophisticated computer programs has improved the feasibility of psychoeducational assessment. One of the most ambitious innovations is the use of automated approaches to scoring writing. These approaches have been implemented in standardized assessments (e.g., Wechsler Individual Achievement Test [WIAT 4] Essay Composition; NCS Pearson, 2020), for instance, for the identification of school-aged students with writing difficulties as well as graduate school admission (e.g., writing essay of the Graduate Record Examination [GRE]) (O'Leary et al., 2018). This technology is attractive given that it reduces time and costs associated with personnel training and scoring procedures (Matta, Keller-Margulis, et al., 2022).

Additionally, some U.S. states have transitioned from hand-calculated to automated approaches in state-mandated assessments (Smith, 2018). More states might follow soon; for instance, the Texas Education Agency (TEA) has recommended using "automated scoring of writing as a way to ensure minimum validity and reliability in scoring, and also control for the costs of implementing a statewide, authentic writing assessment" (TEA, 2018, p. 5).

Although preliminary findings show that computer programs can adequately predict hand-rated writing quality scores for school-aged youth (Matta, Mercer, et al., 2022; Wilson & Rodrigues, 2020), little is known about potential biases and unintended consequences when using automated writing scores in educational settings. These issues can have serious implications for decision-making because the use of biased scores can lead to incorrect identification of students with writing difficulties, especially those with diverse backgrounds, resulting in adverse consequences for individuals and negative systemic effects. Therefore, the purpose of this study was to examine three types of bias potentially associated with automated writing quality scores and their implications for fair and equitable assessments.

The Writer(s)-Within-Community Model as a Framework for Automated Writing Assessment

According to the Writer(s)-Within-Community (WWC) model, writing is a social activity situated within a community and shaped by individual and environmental factors (Graham, 2018). In educational settings, teachers play an active role in creating writing communities of students by establishing shared goals and modeling writing products. Macro-level factors (such as cultural, political, and technological) also influence these teaching practices and the ways teachers develop writing communities, including the way students interact with each other and with teachers, the types of activities they engage in, and the classroom culture surrounding writing. For instance, high-stakes writing assessments designed by education agencies can affect classroom writing activities. Given the societal consequences associated with high-stakes test scores, teachers might focus their instruction on preparing students for these tests which may or may not predict future academic success. In other words, students might become proficient test takers, but not proficient writers.

Recent technological advances, such as automated scoring methods for writing, pose new threats to the evaluation of student writing proficiency on high-stakes assessments. If students are aware of the specific characteristics taken into account for the generation of automated scores, they could potentially write essays that artificially boost their score, rather than produce a sample reflective of their writing ability. For example, concerns regarding existing automated scoring approaches include an over-emphasis of the number of total words on the automated scores (Perelman, 2014). If automated approaches are not thoroughly examined, they might lead to potentially biased and invalid scores.

Distinguishing Bias from Fairness

Educators can use automated writing quality scores to assess student written expression across tasks and contexts and inform practice. In particular, they can use automated scores to draw inferences about expected student performance on a wide range of writing tests or over broader target domains (e.g., writing proficiency). Based on the proposed target domain, test scores might have different meanings and lead to intended or unintended consequences on student learning (Messick, 1989). The investigation of score bias is essential for the identification of flawed interpretations and unfair practices (Kane, 2013).

Although sometimes used interchangeably, bias and fairness refer to distinct concepts (Han et al., 2019). While bias is a property of test scores, fairness is associated with impacts of decisions on students. From a psychometric perspective, multiple regression models can reveal if predictions from test scores are biased for students from certain groups. For instance, a statistically significant interaction between automated writing scores and group membership for the prediction of student writing performance likely indicates biased scores (Warne et al., 2014). In this scenario, students with the same automated writing score are expected to obtain different outcomes as a function of their group. Given that prediction of student outcomes would differ between groups, the presence of test score bias might result in unfair, discriminatory decisions.

Types of Bias for Automated Writing Scores

Computer programs can generate writing quality scores with comparable technical adequacy to hand-calculated scores (Keller-Margulis et al., 2021). Yet, the validity of automated scores might vary among students from different linguistic or cultural backgrounds (Daoyk-Oyry

& Zeinoun, 2017). Given that automated scores are calculated through algorithms trained to predict human ratings, they might incorporate the contribution of sources extraneous to writing quality in the estimation of student performance. The set of linguistic indicators used to generate automated writing scores might only partially capture writing quality and exclude other important dimensions of the construct among students from certain backgrounds. The use of biased scores can amplify discrimination related to pre-existing societal biases and inequities against students from historically marginalized communities. If automated scores are biased, educators will likely draw inaccurate inferences about student writing performance. These inferences linking the use of biased automated scores to educational decisions (e.g., to identify students unable to meet grade-level standards) might lead to inaccurate predictions with adverse impacts on educational practices (Kane, 2013).

Empirical investigations have focused on three types of bias associated with automated writing scores (Baker & Hawn, 2021). First, *scoring bias* refers to the prediction of different human-rated writing scores for students obtaining the same automated scores for a given writing task. Data analyses for the examination of scoring bias generally involve linear regression models or discrepancy statistics (Mercer et al., 2021; Ramineni & Williamson, 2018). Second, *predictive bias* concerns differential predictions of future student outcomes as a function of group characteristics. In other words, students with the same automated writing quality scores are expected to obtain different writing outcomes at a later time; this type of bias can be formally assessed through multiple regression models (Matta, Mercer, et al., 2022). Third, *diagnostic bias* is analogous to predictive bias for binary outcomes. This is a common scenario in educational settings where cut-off scores are established, for instance, to identify students who do not meet the grade-level expectations on a given test. Diagnostic bias can be evaluated through logistic regression models or by comparing diagnostic accuracy metrics (e.g., Area Under the Curve [AUC]) between two groups. Although diagnostic bias is a special case of predictive bias, it is important to differentiate the two types of bias because diagnostic bias has more direct consequences on educational practices.

Societal Consequences of the Use of Automated Writing Scores in the Context of High-Stakes Assessments

Score bias might originate from the use of linguistic metrics on which minoritized communities differ from dominant groups (Gatlin et al., 2016; Johnson et al., 2022). The use of biased scores likely leads to negative consequences for students, teachers, and accountability programs.

First, biased scores likely perpetuate the achievement gaps between students from marginalized groups and their peers. Given that scores on criterion-referenced writing tests tend to be lower for students from minority communities (e.g., Persky et al., 2003), biased automated scores can hinder accurate identification of struggling writers and perpetuate discrimination. Research has shown that the combined effects of biased scores and pre-existing societal inequities increase racial and ethnic disparities in education (Ferri & Connor, 2005). Biased automated scores might prevent students with writing difficulties from receiving appropriate academic support and influence placement in special education programs (Skiba et al., 2008).

Second, biased scores may have negative consequences on teacher performance evaluation (Rentner et al., 2016). When school districts include student performance in teacher evaluations, biased scores might provide administrators with inaccurate information about educators’ abilities and unfairly penalize teachers in schools with more students from historically marginalized communities or low-income families. Biased scores are also less likely to result in useful feedback on writing teaching practices.

Third, biased scores can have adverse consequences on decision-making at the institutional level. In an educational system where the results of criterion-referenced assessments are used for accountability, teachers and administrators might feel obligated to center classroom activities on improving test scores and thereby narrow the focus of the school curriculum (Perna & Thomas, 2009). Therefore, biased scores can lead to instructional decisions that do not promote student long-term academic success.

Purpose of the Study

This study examines three types of bias potentially associated with automated writing quality scores generated via writeAlizer (Mercer, 2020), a free and open-source automated method for scoring writing. Evaluating score bias is important because this evidence can be used to improve ways in which automated scores are generated and used in applied settings, hence reducing biases against historically marginalized groups and promoting fair and equitable assessment decisions. Specifically, we investigate whether automated scores show differential validity and diagnostic accuracy for students from four racial and ethnic groups. First, we examined the validity of writeAlizer scores calculated for untimed, expository writing samples; this was a necessary preliminary step given that writeAlizer scores have only been used for timed, narrative writing samples in previous studies. Second, after showing validity coefficients comparable to hand-rated scores, we addressed the two main research questions:

1. Do automated writing quality scores demonstrate scoring bias? Specifically, do students with the same automated scores obtain different human-rated writing quality scores as a function of their racial and ethnic group?

2. Do automated writing quality scores demonstrate predictive and diagnostic bias? In other words, are students with the same automated scores expected to perform differently on the state-mandated writing test three years later as a function of their group membership? Are the results similar to human-rated writing quality scores?

Method

Participants

Four-hundred twenty-one students (boys *n* = 220, 52.26%) from two campuses located in the Southwestern U.S. participated in the study. The majority of the students were White (33.3%) and Hispanic (28%), followed by Black (17.3%), Asian (16.4%), and biracial (5%). Approximately 27% of the students were eligible for reduced-price or free lunch at school, and 10% did not have English as their primary language. Table 1 includes all demographic information.

Measures

STAAR writing test

The State of Texas Assessments of Academic Readiness (STAAR) writing test is a criterion-referenced writing assessment for students in grades 4 and 7. The test requires students to write an expository essay in response to a written prompt and answer multiple-choice questions about revising and editing. The essay prompts were different across grade levels. In grade 4, students were asked to think about the contexts where they do hard work and write about one of them. In grade 7, students wrote about why it is important for people to show appreciation for each other. In both grades, students also read four passages and responded to questions about ways to make revisions and corrections. Students could obtain up to 8 points in grade 4 and 16 points in grade 7 on the composition portion of the exam and up to 24 points in grade 4 and 30 points in grade 7 for the multiple-choice sections. This means that the composition portion accounted for 25% of the STAAR total score in grade 4 (i.e., 8 out of 32 points) and 35% of the total score in grade 7 (i.e., 16 out of 46 points).

TEA set performance standards through a multi-year development process and a series of linking studies to establish reliability and validity of the test scores. STAAR writing scores were strongly correlated with ReadiStep (*r* = .63) and EXPLORE (*r* = .66), two tests of academic achievement that are linked to the Scholastic Aptitude Test (SAT) and the American College Testing (ACT). There were also strong correlations between the STAAR writing scores obtained by students in grades 4 and 7 (*r* = .62).

writeAlizer

writeAlizer (Mercer, 2020) is an R package that combines the outputs of freely available automated text evaluation programs to generate automated scores of overall writing quality. writeAlizer scores are calculated through the application of scoring models developed on an independent set of narrative writing samples completed in 7 min by upper elementary students in a previous study (Keller-Margulis et al., 2021). The scoring models used to generate automated writing scores were trained to predict human scores of overall quality defined by idea development and idea organization. Full documentation on the scoring models is available on GitHub.

In this study, writeAlizer was based on ReaderBench (Dascălu, 2014), an open-source program for evaluating textual complexity and cohesion. ReaderBench uses advanced natural language processing to measure hundreds of indices ranging from simple (e.g., word and sentence length, and unique words used) to sophisticated measures of text characteristics (e.g., lexical chains and discourse connectives). writeAlizer scores generated from ReaderBench output showed validity and reliability coefficients comparable to human ratings of writing quality for timed, narrative samples (Matta et al., 2022).

Study Design and Procedures

Classroom teachers collected data as part of state testing in spring 2018 and 2021. Students completed the STAAR writing test in grades 4 and 7 in one 4-hour session. Students were free to choose the order and the amount of time to complete each subtest. Two independent trained raters scored each writing sample using a 4-point holistic rubric defining overall writing quality as text organization, idea development, and language and convention use.

The university Institutional Review Board approved use of deidentified student data for the study. Graduate students transcribed the STAAR composition samples into a digital format. Then, the text files were processed through ReaderBench, and the resulting output was imported into writeAlizer. For each sample, we generated one automated writing quality score. For comparison, we extracted the number of total words written (TWW) from the output to provide a lower limit for the interpretation of the results.

Data Analysis

Analyses were conducted in RStudio (RStudio Team, 2020). Of the 421 students enrolled in the two campuses, we received copies of the STAAR writing samples for all but two students in grade 4 (99.52%) and for 332 students in grade 7 (78.86%). One student was removed from the original dataset because no test scores were available at either grade. Under the assumption that data were missing at random, we used multiple imputations to generate 1,000 complete datasets via the mice R package (van Buuren & Groothuis-Oudshoorn, 2011). Analyses were conducted on each imputed dataset separately and pooled together using Rubin’s rule via the mice and psfmi (Heymans & Eekhout, 2021) R packages.

Initially, we examined validity and diagnostic accuracy of writeAlizer scores using the entire sample. First, scoring validity was established by calculating Pearson's *r* coefficients between writeAlizer scores and STAAR composition scores based on the same samples in grades 4 and 7. Second, predictive validity was examined by calculating the correlations between writeAlizer scores in grade 4 and STAAR composition and total scores in grade 7, respectively. Third, diagnostic accuracy was calculated through Area Under the Curve (AUC) coefficients. writeAlizer scores for students in grade 4 were used to predict the probability of students not meeting grade-level expectations (below the 35th percentile) on the STAAR test in grade 7.

To assess scoring bias of writeAlizer (Research Question #1), we used multiple regression models with a multicategory moderator to examine the extent to which writeAlizer scores and race/ethnicity (with White students as the reference group) would predict STAAR composition scores in grades 4 and 7, respectively. A significant interaction between writeAlizer scores and race/ethnicity would show evidence of scoring bias. To facilitate interpretation of results, we also estimated the same regression model with TWW and STAAR composition score as expected lower and upper bounds for interpretation of the regression coefficients, respectively.

To assess predictive and diagnostic bias (Research Question #2), we conducted multiple regression models with a multicategory moderator using writeAlizer scores, race/ethnicity, and their interaction for the prediction of STAAR scores in grade 7. First, we estimated these predictors in relation to STAAR composition and total scores to investigate predictive bias. Given that the continuous variables were standardized, we interpreted the beta coefficients for the main effect of race/ethnicity in line with Cohen's considerations for standardized mean difference effect sizes (Cohen, 1988); coefficients lower than 0.20 indicated a negligible difference, between 0.20 and 0.50 small, between 0.50 and 0.80 moderate, and greater than 0.80

large. Second, we used the same variables to predict not reaching proficiency on the STAAR writing test to examine diagnostic bias. A significant interaction between writeAlizer scores and race would show evidence of predictive or diagnostic bias. Consistent with the previous set of data analyses, we also estimated the same models for TWW and STAAR composition scores to facilitate interpretation of the magnitude of the regression coefficients.

Results

Means and standard deviations for writeAlizer and STAAR scores (composition, total, and proficiency) are presented in Table 2.

Validity and Diagnostic Accuracy of writeAlizer Scores

Validity and diagnostic accuracy coefficients for writeAlizer are reported in Supplemental Tables S1 and S2. Overall, a clear pattern emerged. Validity and accuracy coefficients of writeAlizer scores consistently outperformed TWW, were comparable to STAAR composition scores, and underperformed the STAAR total score. For instance, writeAlizer scores in grade 4 demonstrated stronger scoring validity (r = .638, 95% CI [.576, .693]) than TWW (r = .508, 95% CI [.431, .577]) and displayed comparable predictive validity (r = .478, 95% CI [.394, .555]) to STAAR composition scores in grade 4 (r = .468, 95% CI [.382, .546]) for the prediction of STAAR composition scores three years later. Similar results were observed for diagnostic accuracy.

Additionally, validity and diagnostic accuracy were calculated for students across four racial and ethnic groups separately (Supplemental Tables S3 and S4). The pattern of the results replicated the findings for the entire sample.

Research Question #1: Scoring Bias of writeAlizer Scores

Table 3 includes the results of multiple regression models examining scoring bias of writeAlizer scores within grade levels. writeAlizer scores were similarly predictive of STAAR composition scores in grade 4 (β = .64, *t*(383.35) = 8.86, *p* <.001) and grade 7 (β = .65, *t*(327.35) = 8.82, *p* <.001). For race/ethnicity, we found a statistically significant difference between White and Black students in grade 4 with the latter group showing lower scores on STAAR composition (β = -0.28, *t*(384.98) = -2.48, *p* = .01). This difference did not replicate in grade 7. No interaction effect was statistically significant in grades 4 or 7. The same pattern was observed in the regression models with TWW as predictors (see Supplemental Table S5).

Research Question #2: Predictive and Diagnostic Bias of writeAlizer Scores

Table 4 reports the multiple regression coefficients for predictive bias of writeAlizer scores. Results differed by outcome in the model. When used to predict STAAR composition scores in grade 7, there were no statistically significant effects of race/ethnicity or interactions with writeAlizer scores. By contrast, when used to predict STAAR total scores, there were small and statistically significant differences by race/ethnicity. After controlling for writeAlizer scores in grade 4, White students on average had higher total scores than Hispanic (β = -0.32, *t*(356.71) = -2.89, *p* = .004) and Black students (β = -0.46, *t*(342.59) = -3.49, *p* = .001) in grade 7. We found the same pattern of results from regression models examining the predictive bias of TWW and STAAR composition scores in grade 4 (see Supplemental Tables S6 and S7).

Lastly, we investigated diagnostic bias (Table 5). As expected, writeAlizer scores were statistically significant predictors of STAAR non-proficiency three years later (β = -1.27, *t*(340.47) = -3.24, *p* = .001). After controlling for this effect, there was a significant effect of race/ethnicity with White students more likely to reach proficiency than Hispanic students (β = 1.20, *t*(349.93) = -2.28, *p* = .02, *OR* = 3.31 (95% CI [1.18, 9.34]). Findings were similar when TWW (Supplemental Table S8) and STAAR composition scores were entered as predictors for the regression model (Supplemental Table S9).

Discussion

This study investigated the use of automated writing scores and the implications of bias for students from historically marginalized groups. In previous work, writeAlizer scores showed good validity and diagnostic accuracy for narrative tasks completed in 3 and 7 min by students in grades 2 to 5 (Matta, Mercer, et al., 2022; Mercer et al., 2019). In the current study, we evaluated the use of writeAlizer for untimed, expository writing samples and demonstrated that writeAlizer scores have good validity and reasonable diagnostic accuracy when the genre of the writing samples differed from the genre of samples used to train the models. Consistent with prior results for narrative samples, writeAlizer scores on expository samples show stronger validity and diagnostic accuracy than the simple count of words and are comparable to human-rated scores.

This examination of automated score bias expands the literature in two directions. First, no scoring bias was observed against students from historically marginalized groups on the expository essay portion of the state exam. In other words, students with the same automated scores were not expected to receive different scores for the same writing sample as a function of their racial or ethnic backgrounds. The lack of evidence for scoring bias, along with moderate to strong correlations between automated and hand-rated scores for the same set of writing samples, supports accuracy and appropriateness of automated scores (Yang et al., 2002). The results of the regression model revealed, however, that there was a significant main effect of group with Black students in grade 4 expected to obtain lower human-rated writing scores than White students. Although group differences are not considered evidence of test bias (Warne et al., 2014), this finding has implications for student learning, especially when scores are used for decision-making. For example, the use of the same regression line to predict human-rated scores from automated scores can lead to inaccurate inferences for all students. When scores misrepresent writing performance in either direction (i.e., underestimate or overestimate), educators might fail to identify students at risk for poor outcomes and provide adequate academic support. Given that the effect does not replicate when the regression model was estimated for the same students in grade 7, differences in human-rated writing scores might depend on the characteristics of the sample or resolve as students develop more sophisticated writing abilities.

Second, automated scores did not show predictive and diagnostic bias in relation to the STAAR scores three years later. Our findings do not provide evidence that automated scores display different predictions between students from different racial or ethnic groups. However, we observed overall group differences depending on the outcome used in the regression model. In relation to predictive bias, after controlling for scores on the essay portion of the state exam in grade 4 (generated through either an automated or human-rated approach), there were no group differences in composition scores in grade 7; Black and Hispanic students were not expected to receive lower writing quality scores. By contrast, both groups were more likely to exhibit lower total scores, which combine human-rated writing quality with revising and editing, on the state-mandated writing test. In relation to diagnostic bias, there were statistically significant differences between Hispanic and White students, with the former group less likely to meet the achievement standards. After controlling for the effect of writing quality, the odds of failing to reach proficiency levels for Hispanic students were 3.31 times that of White students on the state-mandated writing test three years later.

Implications for Research on Test Bias

This pattern of results may hint at two broader factors often neglected in the test bias literature: a) the presence of statistically different intercepts between groups, and b) the nature of the construct assessed by the criterion measure and its response format (Zieky, 2016). Our findings reveal that, whenever a statistically significant group effect occurred, students from historically marginalized communities were expected to obtain lower outcome scores compared to White students. One of the factors contributing to this performance gap is likely the disproportionate number of students with diverse linguistic backgrounds from marginalized groups. For instance, in the 2019-2020 school year, 87% of students in Texas identified as emergent bilingual or English learners were Hispanic (TEA, 2020). This issue is complex, as the interaction between race/ethnicity and language creates opportunities for differences in test scores and the adoption of unfair practices. Children growing up in urban communities, especially in the U.S. South, Southwest, and California, might learn to use expressions typical of Latino or Black English, where grammar and syntax are partially different from Standard American English, with negative consequences on their state test scores (Baker-Bell, 2020; Horton-Ikard & Pittman, 2010). Analogous issues can affect the assessment of students who are English learners. The application of scoring procedures designed for native speakers might lead to lower test scores for language learners (Sireci & Faulkner-Bond, 2015).

Additionally, the disproportionate representation of Black and Hispanic students in low score ranges is the consequence of historic, systematic biases against marginalized groups in U.S. school systems. Since the second half of the nineteenth century, overt and covert eugenic movements have fought to create and preserve school segregation and exploit applied quantitative methods to legitimize human hierarchies through scientific racism (American Psychological Association [APA], 2021; Sullivan et al., 2019). Biased inferences from empirical evidence of achievement gaps among racial and ethnic groups have been used to allocate fewer resources to schools and districts, which in turn have systematic effects on the support available for historically marginalized students (Darden & Cavendish, 2011).

Therefore, we argue that group differences should be considered for their contribution to incorrect predictions and unfair decisions. Although the presence of mean differences does not show that test scores are biased against lower-performing groups, the differences may expose construct-irrelevant sources of variance in the outcome measure. Given that we found no group differences when the regression model predicted writing quality on the essay portion of the state exam but found group differences in overall state exam scores, the construct-irrelevant variance might be in the revising and editing sections of the state-mandated writing test. Several factors might explain this pattern, such as the multiple-choice response format of the revising and editing sections or the writing construct measured (e.g., writing quality vs. editing/revising) (Heck & Crislip, 2001). For instance, the composition task can be conceptualized as a writing task (i.e., students generate their own text) whereas the multiple-choice test is a reading and writing task that taps into different content knowledge (e.g., reading and interpreting context, then responding). Therefore, results not only speak to differences in testing tasks but further highlight the shared but unique knowledge associated with reading and writing development. Although the current study design cannot determine the extent to which these factors contribute to group differences, the results raise concerns about the fairness of writing assessment methods. The evidence is particularly alarming given that students from historically marginalized communities are expected to obtain overall lower scores on the state-mandated writing test even when they show no statistical differences in scores on the more authentic expository essay section of the state exam.

Lastly, this study also illustrates a methodology that researchers and software developers should employ when validating automated approaches to scoring writing. Although we found no evidence of bias associated with writeAlizer scores against students from historically marginalized groups, the examination of other automated approaches might lead to different conclusions. Therefore, this manuscript should serve as a roadmap to use with existing automated methods to detect bias and evaluate the extent to which they have similar validity to human ratings of writing quality. By doing so, researchers could expose additional negative societal consequences that would arise from using automated scoring of high-stakes summative state testing. Additional work is necessary to reveal whether similar consequences (e.g., effects on teacher performance evaluation) emerge when using automated scoring with formative, classroom-based writing samples.

Constraints on Generality

This study has three limitations. First, there is evidence that writeAlizer scores might underperform with untimed, expository essays. The scoring validity for the simple count of words and writeAlizer scores had nearly identical magnitude for students in grade 7. This finding might depend on the use of statistical models developed for 7-min, narrative samples for the scoring of expository essays with virtually no time constraint. Future studies should refine the scoring models to generate writeAlizer scores for compositions of the same writing genre and completed with similar procedures. Another explanation of this finding might have to do with a ceiling effect. In fact, the magnitude of scoring validity coefficients of automated scores rarely exceeds .80 (Cotos, 2014). This phenomenon might depend on the limited range of writing scores or the intrinsically subjective nature of writing assessment. More research on the psychometric properties of automated writing scores is needed.

Second, the investigation of score bias was limited by the use of discrete categories that do not represent homogeneous groups. Race and ethnicity are multifaceted, social constructs that interact with other variables, such as socio-economic status and linguistic backgrounds (Han et al., 2019). The inclusion of individuals with different backgrounds in the same group likely increases the standard error of regression coefficients which, in turn, might affect the estimation of statistically significant effects. Relatedly, we were unable to identify the sources of bias responsible for the results. Future studies should consider the intersectionality of racial and ethnic groups with other demographic characteristics, identify the optimal balance with representativeness and study feasibility, and examine potential factors underlying the significant group differences observed in the study.

Lastly, several key aspects of writeAlizer have yet to be examined given that its validation process is only in the early stages. Although it has shown promising results for diagnostic accuracy, we are unable to use writeAlizer scores to make inferences about student risk status. In addition, writeAlizer produces a continuous variable of writing quality. Given that the scaling of writeAlizer scores is different from other scoring systems, educators are not able to calculate discrepancies with hand-rated scores. The availability of the same scale would allow for the integration of automated and hand-rated scores. For example, the writing essay of the GRE is assessed by the e-rater® Automated Scoring Engine and a human rater; when the two scores are within a negligible difference, samples are not reviewed by a second rater. Future studies should develop norms to identify at-risk students and establish ways in which educators can use writeAlizer scores.

Implications for Practice

In line with the WWC model, this study shows the effect of macro-level factors on the student scores in the context of high-stakes writing assessments. The inclusion of tasks with differential construct manifestation and underrepresentation might introduce possible sources of bias into test scores against certain groups of students (Daoyk-Oyry & Zeinoun, 2017). Taken together, our results show that writing scores might lead to unfair educational decisions not because of a particular approach to scoring written compositions (automated vs hand-rated) but rather because of the inclusion of other writing components in the test that are only partially associated with the ability to write a cohesive and well-organized text, namely, revising and editing skills. For instance, in the revising section of the state writing test, students in grade 7 were required to read a two-page essay about butterfly wings and select one sentence from four alternatives serving as the controlling idea for the paragraph; among other things, this question artificially isolates revising skills, favors those with prior knowledge of the topic, and assumes that students are familiar with the concept of *controlling ideas*. Moreover, given that classroom instruction is generally directed toward authentic writing, educators are forced to teach to the test, a method primarily intended to improve student performance on high-stakes examinations, rather than master writing skills. While this may lead to better test scores, it does not necessarily assist students to develop real-world skills. In a more authentic assessment of writing skills, students would be asked to write all they know about butterflies or, more generally, their favorite animal so that improved interest in the topic and motivation could enhance their writing performance.

References

American Psychological Association (2022, February). *Historical chronology. Examining psychology's contributions to the belief in racial hierarchy and perpetuation of inequality for people of color in U.S.* https://www.apa.org/about/apa/addressing-racism/historical-chronology

Baker, R.S., & Hawn, A. (2021). Algorithmic Bias in Education. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-021-00285-9

Baker-Bell, A. (2020). *Linguistic justice: Black language, literacy, identity, and pedagogy*. Routledge.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge Academic.

Cotos, E. (2014). Automated Writing Evaluation. In E. Cotos (Ed.), *Genre-Based Automated Writing Evaluation for L2 Research Writing* (pp. 40–64). Palgrave Macmillan UK. https://doi.org/10.1057/9781137333377_3

Daouk-Öyry, L., & Zeinoun, P. (2017). Testing across cultures: Translation, adaptation and indigenous test development. In B. Cripps (Ed.), Psychometric Testing: Critical Perspectives (pp. 221–233). John Wiley & Sons. https://doi.org/10.1002/9781119183020.ch16

Darden, E.C., & Cavendish, E. (2012). Achieving resource equity within a single school district: Erasing the opportunity gap by examining school board decisions. *Education and Urban Society*, *44*(1), 61–82. https://doi.org/10.1177/0013124510380912

Dascălu, M. (2014). *Analyzing Discourse and Text Complexity for Learning and Collaborating* (Vol. 534). Springer International Publishing. https://doi.org/10.1007/978-3-319-03419-5

Ferri, B., & Connor, D. (2005). Tools of Exclusion: Race, Disability, and (Re)segregated Education. *Teachers College Record*, *107*, 453–474. https://doi.org/10.1111/j.1467-9620.2005.00483.x

Gatlin, B., Wanzek, J., & Al Otaiba, S. (2016). An Examination of Kindergarten Oral Language for Black Students: Are There Meaningful Differences in Comparison to Peers? *Reading & Writing Quarterly*, *32*(5), 477–498. https://doi.org/10.1080/10573569.2015.1039737

Graham, S. (2018). A revised writer(s)-within-community model of writing. *Educational Psychology*, *53*(4), 258–279. https://doi.org/10.1080/00461520.2018.1481406

Han, K., Colarelli, S.M., & Weed, N.C. (2019). Methodological and statistical advances in the consideration of cultural diversity in assessment: A critical review of group classification and measurement invariance testing. *Psychological Assessment*, *31*(12), 1481–1496. https://doi.org/10.1037/pas0000731

Heck, R. H., & Crislip, M. (2001). Direct and indirect writing assessments: Examining issues of equity and utility. *Educational evaluation and policy analysis*, *23*(3), 275-292. https://doi.org/10.3102/01623737023003275

Heymans, M., & Eekhout, I. (2021). *psfmi. Prediction Model Selection and Performance Evaluation in Multiple Imputed Datasets*. https://mwheymans.github.io/psfmi/

Horton-Ikard, R., & Pittman, R.T. (2010). Examining the Writing of Adolescent Black English Speakers: Suggestions for Assessment and Intervention. *Topics in Language Disorders*, *30*(3), 189–204. https://doi.org/10.1097/TLD.0b013e3181efc3bd

Johnson, M. S., Liu, X., & McCaffrey, D. F. (2022). Psychometric methods to evaluate measurement and algorithmic bias in automated scoring. Journal of Educational Measurement. https://doi.org/10.1111/jedm.12335

Kane, M. (2013). Validating the Interpretations and Uses of Test Scores: Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Keller-Margulis, M.A., Mercer, S.H., & Matta, M. (2021). Validity of automated text evaluation tools for written-expression curriculum-based measurement: A comparison study. *Reading and Writing*, *34*, 2461–2480. https://doi.org/10.1007/s11145-021-10153-6

Matta, M., Keller-Margulis, M.A., & Mercer, S.H. (2022). Cost analysis and cost-effectiveness of hand-scored and automated approaches to writing screening. *Journal of School Psychology*, *92*, 80–95. https://doi.org/10.1016/j.jsp.2022.03.003

Matta, M., Mercer, S.H., & Keller-Margulis, M.A. (2022). Evaluating validity and bias for hand-calculated and automated written expression curriculum-based measurement scores. *Assessment in Education: Principles, Policy & Practice.* https://doi.org/10.1080/0969594X.2022.2043240

Mercer, S.H. (2020). *writeAlizer: Generate predicted writing quality and written expression CBM scores* (1.2.0) [Computer software]. https://github.com/shmercer/writeAlizer/.

Mercer, S.H., Cannon, J.E., Squires, B., Guo, Y., & Pinco, E. (2021). Accuracy of Automated Written Expression Curriculum-Based Measurement Scoring. *Canadian Journal of School Psychology*, *36*(4), 304–317. https://doi.org/10.1177/0829573520987753

Mercer, S.H., Keller-Margulis, M.A., Faith, E.L., Reid, E.K., & Ochs, S. (2019). The Potential for Automated Text Evaluation to Improve the Technical Adequacy of Written Expression Curriculum-Based Measurement. *Learning Disability Quarterly*, *42*(2), 117–128. https://doi.org/10.1177/0731948718803296

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). American Council on education and Macmillan.

NCS Pearson. (2020). *Wechsler Individual Achievement Test* (4th ed.).

O'Leary, M., Scully, D., Karakolidis, A., & Pitsia, V. (2018). The state-of-the-art in digital technology-based assessment. *European Journal of Education*, *53*(2), 160–175. https://doi.org/10.1111/ejed.12271

Perelman, L. (2014). When "the state of the art" is counting words. *Assessing Writing*, *21*, 104–111. https://doi.org/10.1016/j.asw.2014.05.001

Perna, L.W., & Thomas, S.L. (2009). Barriers to College Opportunity: The Unintended Consequences of State-Mandated Testing. *Educational Policy*, *23*(3), 451–479. https://doi.org/10.1177/0895904807312470

Persky, H.R., Daane, M.C., & Jin, Y. (2003). *The Nation's Report Card: Writing 2002* (No. NCES2003–529). U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. https://nces.ed.gov/nationsreportcard/pdf/main2002/2003529.pdf

Ramineni, C., & Williamson, D. (2018). *Understanding Mean Score Differences Between the e-rater Automated Scoring* (ETS Research Report Series).

Rentner, D.S., Kober, N., Frizzell, M., & Ferguson, M. (2016). *Listen to Us: Teacher Views and Voices*. Center on Education Policy. https://eric.ed.gov/?id=ED568172

RStudio Team (2020). *RStudio: Integrated Development for R.* RStudio, PBC. http://www.rstudio.com/

Sireci, S.G., & Faulkner-Bond, M. (2015). Promoting Validity in the Assessment of English Learners. *Review of Research in Education*, *39*(1), 215–252. https://doi.org/10.3102/0091732X14557003

Skiba, R.J., Simmons, A.B., Ritter, S., Gibb, A.C., Rausch, M.K., Cuadrado, J., & Chung, C.-G. (2008). Achieving equity in special education: History, status, and current challenges. *Exceptional Children*, *74*(3), 264–288. https://doi.org/10.1177/001440290807400301

Smith, T. (2018). *More states opting to 'robo-grade' student essays by computer*. National Public Radio (NPR). https://www.npr.org/2018/06/30/624373367/more-states-opting-to-robo-grade-student-essays-by-computer

Sullivan, A.L., Sadeh, S., & Houri, A.K. (2019). Are school psychologists' special education eligibility decisions reliable and unbiased? A multi-study experimental investigation. *Journal of School Psychology*, *77*, 90–109. https://doi.org/10.1016/j.jsp.2019.10.006

TEA (2018). *House Bill 1164 Texas Writing Pilot Program. Report to the Governor and the Texas Legislature*. https://tea.texas.gov/sites/default/files/Texas%20Writing%20Pilot%20Year-Two%20September%202018%20Legislative%20Report%2010.26.18%20FINAL.pdf

TEA (2020). *Enrollment in Texas public schools, 2019-20* (GE20 601 12). https://tea.texas.gov/sites/default/files/enroll_2019-20.pdf

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Warne, R.T., Yoon, M., & Price, C.J. (2014). Exploring the various interpretations of “test bias”. *Cultural Diversity and Ethnic Minority Psychology*, *20*(4), 570–582. https://doi.org/10.1037/a0036503

Wilson, J., & Rodrigues, J. (2020). Classification accuracy and efficiency of writing screening using automated essay scoring. *Journal of School Psychology*, *82*, 123–140. https://doi.org/10.1016/j.jsp.2020.08.008

Yang, Y., Buckendahl, C.W., Juszkiewicz, P.J., & Bhola, D.S. (2002). A Review of Strategies for Validating Computer-Automated Scoring. *Applied Measurement in Education*, *15*(4), 391–412. https://doi.org/10.1207/S15324818AME1504_04

Zieky, M.J. (2016). Fairness in test design and development. In N.J. Dorans & L.L. Cook (Eds.), *Fairness in educational assesment and measurement* (pp. 9–32). Routledge.

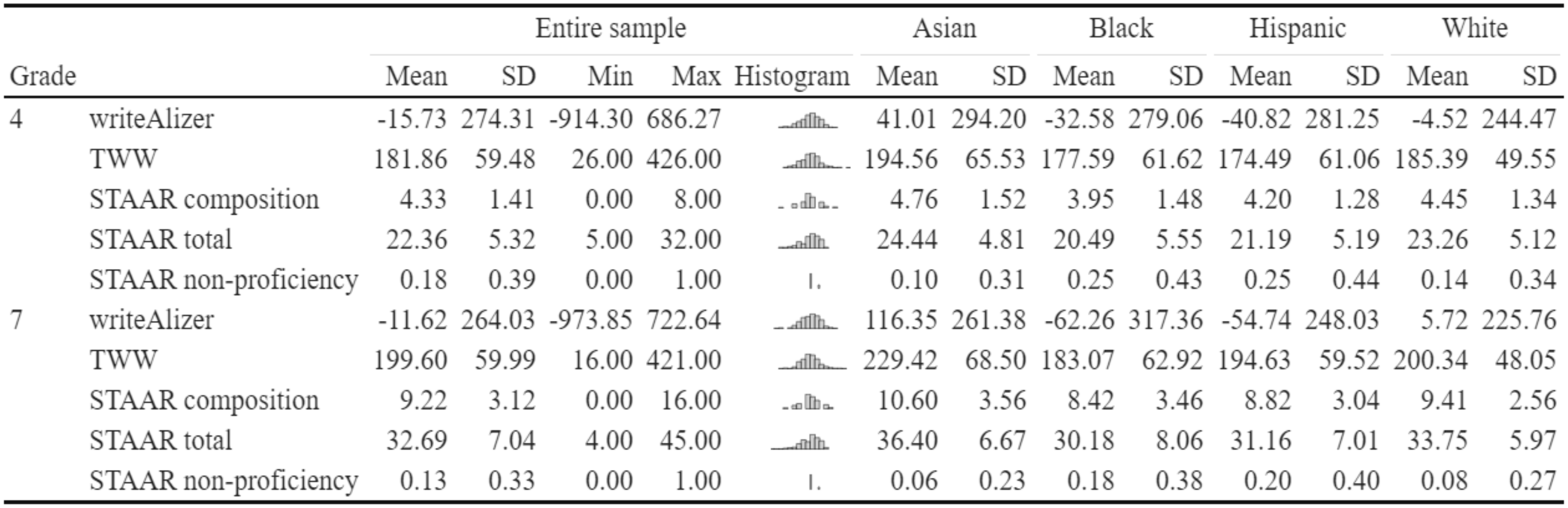
Table 1

Sample demographics

	Entire Sample	Asian	Black	Hispanic	White	Biracial
Sample size, *n*	421	69	73	118	140	21
Sex, *n* (%)						
Boys	220 (52.3)	41 (59.4)	33 (45.2)	66 (55.9)	68 (48.6)	12 (57.1)
Girls	201 (47.7)	28 (40.6)	40 (54.8)	52 (44.1)	72 (51.4)	9 (42.9)
Special Education, *n* (%)	30 (7.1)	3 (4.3)	7 (9.6)	8 (6.8)	10 (7.1)	2 (9.5)
English Learners, *n* (%)	44 (10.5)	18 (26.1)	2 (2.7)	16 (13.6)	8 (5.7)	0
Eligible for free or reduced-price meals, *n* (%)	115 (27.3)	15 (21.7)	36 (49.3)	45 (38.1)	15 (10.7)	4 (19)
Gifted, *n* (%)	62 (14.7)	23 (33.3)	2 (2.7)	8 (6.8)	25 (17.9)	4 (19)

Table 2

Descriptive statistics for the entire sample and by racial and ethnic group



		Entire sample					Asian		Black		Hispanic		White	
Grade		Mean	SD	Min	Max	Histogram	Mean	SD	Mean	SD	Mean	SD	Mean	SD
4	writeAlizer	-15.73	274.31	-914.30	686.27		41.01	294.20	-32.58	279.06	-40.82	281.25	-4.52	244.47
	TWW	181.86	59.48	26.00	426.00		194.56	65.53	177.59	61.62	174.49	61.06	185.39	49.55
	STAAR composition	4.33	1.41	0.00	8.00		4.76	1.52	3.95	1.48	4.20	1.28	4.45	1.34
	STAAR total	22.36	5.32	5.00	32.00		24.44	4.81	20.49	5.55	21.19	5.19	23.26	5.12
	STAAR non-proficiency	0.18	0.39	0.00	1.00		0.10	0.31	0.25	0.43	0.25	0.44	0.14	0.34
7	writeAlizer	-11.62	264.03	-973.85	722.64		116.35	261.38	-62.26	317.36	-54.74	248.03	5.72	225.76
	TWW	199.60	59.99	16.00	421.00		229.42	68.50	183.07	62.92	194.63	59.52	200.34	48.05
	STAAR composition	9.22	3.12	0.00	16.00		10.60	3.56	8.42	3.46	8.82	3.04	9.41	2.56
	STAAR total	32.69	7.04	4.00	45.00		36.40	6.67	30.18	8.06	31.16	7.01	33.75	5.97
	STAAR non-proficiency	0.13	0.33	0.00	1.00		0.06	0.23	0.18	0.38	0.20	0.40	0.08	0.27

Note. TWW = Total Words Written.

Table 3

Scoring bias of writeAlizer scores

Variable	STAAR composition (Grade 4)					STAAR composition (Grade 7)				
	B	*SE*	*t*	*df*	*p*	*B*	*SE*	*t*	*df*	*p*
Intercept	0.06	0.07	0.96	388.45	.34	0.02	0.06	0.35	341.77	.72
writeAlizer (wA)	0.64	0.07	8.86	383.35	<.001	0.65	0.07	8.82	327.35	<.001
Race/Ethnicity (Hispanic = 1)	-0.12	0.10	-1.21	388.69	.23	0.00	0.10	0.03	329.89	.98
Race/Ethnicity (Black = 1)	-0.28	0.11	-2.48	384.98	.01	-0.15	0.11	-1.33	309.20	.18
Race/Ethnicity (Asian = 1)	0.09	0.12	0.74	388.70	.46	0.03	0.12	0.25	303.10	.80
wA x Race/Ethnicity (Hispanic = 1)	-0.18	0.10	-1.76	386.89	.08	0.08	0.10	0.73	326.92	.46
wA x Race/Ethnicity (Black= 1)	0.04	0.12	0.38	375.07	.71	0.06	0.11	0.57	328.71	.57
wA x Race/Ethnicity (Asian = 1)	0.11	0.11	0.98	387.61	.33	0.08	0.12	0.67	315.26	.50

Note. The variables were entered as predictors of the STAAR composition within each grade level.

Table 4

Predictive bias of writeAlizer scores

Variable	STAAR composition (Grade 7)					STAAR total (Grade 7)				
	B	*SE*	*t*	*df*	*p*	*B*	*SE*	*t*	*df*	*p*
Intercept	0.05	0.08	0.65	351.61	.52	0.15	0.07	1.99	363.42	.05
writeAlizer (wA)	0.42	0.09	4.66	325.54	<.001	0.43	0.08	5.07	341.71	<.001
Race/Ethnicity (Hispanic = 1)	-0.13	0.12	-1.12	344.50	.26	-0.32	0.11	-2.89	356.71	.004
Race/Ethnicity (Black = 1)	-0.26	0.14	-1.90	325.88	.06	-0.46	0.13	-3.49	342.59	.001
Race/Ethnicity (Asian = 1)	0.21	0.15	1.43	314.63	.15	0.17	0.14	1.26	333.60	.21
wA x Race/Ethnicity (Hispanic = 1)	0.04	0.12	0.35	331.40	.73	0.01	0.12	0.09	346.18	.93
wA x Race/Ethnicity (Black= 1)	0.04	0.14	0.31	338.76	.75	0.09	0.13	0.66	345.03	.51
wA x Race/Ethnicity (Asian = 1)	0.08	0.14	0.53	315.84	.59	0.07	0.14	0.55	323.53	.59

Table 5

Diagnostic bias of writeAlizer scores

	STAAR non-proficiency (Grade 7)							
Variable	*B*	*SE*	*t*	*df*	*p*	*OR*	*LL*	*UL*
Intercept	-2.82	0.44	-6.41	347.25	<.001	0.06	0.02	0.14
writeAlizer (wA)	-1.27	0.39	-3.24	340.47	.001	0.28	0.13	0.61
Race/Ethnicity (Hispanic = 1)	1.20	0.53	2.28	349.93	.02	3.31	1.18	9.34
Race/Ethnicity (Black = 1)	1.05	0.60	1.75	338.19	.08	2.86	0.88	9.33
Race/Ethnicity (Asian = 1)	-0.09	0.82	-0.11	337.80	.91	0.91	0.18	4.53
wA x Race/Ethnicity (Hispanic = 1)	0.43	0.47	0.91	346.76	.36	1.53	0.61	3.84
wA x Race/Ethnicity (Black= 1)	0.31	0.54	0.58	337.06	.56	1.37	0.47	3.95
wA x Race/Ethnicity (Asian = 1)	0.09	0.70	0.13	282.04	.89	1.10	0.28	4.31

Note. OR = Odds Ratio, LL = Lower Limit, UL = Upper Limit.

Table S1

Scoring and predictive validity coefficients for STAAR written composition scores

Grade	Measure	*r*	*LL*	*UL*
4 → 4	writeAlizer	.638	.576	.693
	TWW	.508	.431	.577
7 → 7	writeAlizer	.712	.655	.762
	TWW	.654	.588	.712
4 → 7	writeAlizer	.478	.394	.555
	TWW	.363	.268	.452
	STAAR composition	.468	.382	.546
	STAAR total	.617	.544	.681

Note. Temporal stability of writeAlizer *r* = .472, 95% CI [.385, .550] and TWW *r* = .411, 95% CI [.315, .499]

Table S2

Predictive validity and diagnostic accuracy coefficients for STAAR total scores (grade 7)

Measure	*r*	*LL*	*UL*	*AUC*	*LL*	*UL*
writeAlizer	.489	.407	.563	.764	.739	.788
TWW	.363	.271	.449	.682	.654	.710
STAAR composition	.512	.433	.584	.743	.594	.851
STAAR total	.762	.712	.804	.895	.876	.912

Table S3

Scoring and predictive validity coefficients for STAAR composition scores by racial and ethnic group

Grade	Measure	Asian	Black	Hispanic	White
4 → 4	writeAlizer	.746	.670	.521	.608
	TWW	.646	.581	.384	.418
7 → 7	writeAlizer	.668	.761	.711	.668
	TWW	.593	.771	.598	.624
4 → 7	writeAlizer	.477	.442	.477	.441
	TWW	.332	.334	.358	.305
	STAAR composition	.525	.332	.496	.421
	STAAR total	.628	.602	.630	.562

Table S4

Predictive validity and diagnostic accuracy coefficients for STAAR total scores (grade 7) by racial and ethnic group

Measure	Asian		Black		Hispanic		White	
	r	*AUC*	*r*	*AUC*	*r*	*AUC*	*r*	*AUC*
writeAlizer	.559	.776	.489	.739	.450	.759	.455	.770
TWW	.423	.685	.358	.642	.288	.681	.304	.653
STAAR composition	.556	.809	.390	.719	.528	.716	.488	.784
STAAR total	.754	.939	.738	.899	.794	.871	.704	.883

Table S5

Scoring bias of TWW scores

Variable	STAAR composition (grade 4)					STAAR composition (grade 7)				
	B	*SE*	*t*	*df*	*p*	*B*	*SE*	*t*	*df*	*p*
Intercept	0.05	0.07	0.75	388.86	.45	0.06	0.07	0.84	347.47	.40
TWW	0.47	0.09	5.36	369.32	<.001	0.64	0.08	7.53	314.80	<.001
Race/Ethnicity (Hispanic = 1)	-0.11	0.11	-1.01	388.95	.32	-0.12	0.10	-1.19	336.28	.23
Race/Ethnicity (Black = 1)	-0.28	0.12	-2.21	386.26	.03	-0.10	0.12	-0.79	314.07	.43
Race/Ethnicity (Asian = 1)	0.11	0.13	0.89	388.93	.38	0.05	0.13	0.41	317.13	.68
TWW x Race/Ethnicity (Hispanic = 1)	-0.13	0.12	-1.15	378.32	.25	-0.05	0.11	-0.42	315.67	.67
TWW x Race/Ethnicity (Black= 1)	0.11	0.13	0.80	367.56	.42	0.17	0.13	1.33	311.59	.19
TWW x Race/Ethnicity (Asian = 1)	0.15	0.13	1.19	380.56	.23	-0.07	0.12	-0.54	316.02	.59

Note. The variables were entered as predictors of the STAAR composition score within grade level.

Table S6

Predictive bias of TWW scores

Variable	STAAR composition (grade 7)					STAAR total (grade 7)				
	B	*SE*	*t*	*df*	*p*	*B*	*SE*	*t*	*df*	*p*
Intercept	0.05	0.08	0.55	355.25	.58	0.14	0.08	1.80	366.26	.07
TWW	0.31	0.10	3.07	332.91	.002	0.31	0.10	3.17	344.96	.002
Race/Ethnicity (Hispanic = 1)	-0.12	0.12	-1.01	347.92	.31	-0.32	0.12	-2.69	359.91	.01
Race/Ethnicity (Black = 1)	-0.27	0.15	-1.83	333.94	.07	-0.46	0.14	-3.33	349.35	.001
Race/Ethnicity (Asian = 1)	0.24	0.15	1.59	322.92	.11	0.20	0.15	1.39	340.95	.17
TWW x Race/Ethnicity (Hispanic = 1)	0.03	0.14	0.22	323.57	.82	-0.03	0.13	-0.23	338.61	.82
TWW x Race/Ethnicity (Black= 1)	0.03	0.15	0.21	344.67	.84	0.06	0.14	0.43	351.46	.67
TWW x Race/Ethnicity (Asian = 1)	0.02	0.15	0.15	329.50	.88	0.06	0.14	0.41	340.46	.68

Note. The variables were entered as predictors of the STAAR composition score in grade 4.

Table S7

Predictive bias of STAAR composition scores

Variables	STAAR composition (grade 7)					STAAR total (grade 7)				
	B	*SE*	*t*	*df*	*p*	*B*	*SE*	*t*	*df*	*p*
Intercept	0.03	0.08	0.40	351.44	.69	0.12	0.07	1.68	361.99	.09
STAAR composition	0.38	0.09	4.30	304.61	<.001	0.44	0.08	5.35	324.99	<.001
Race/Ethnicity (Hispanic = 1)	-0.10	0.12	-0.87	346.22	.38	-0.28	0.11	-2.55	356.77	.01
Race/Ethnicity (Black = 1)	-0.19	0.14	-1.36	327.17	.17	-0.37	0.13	-2.79	342.32	.01
Race/Ethnicity (Asian = 1)	0.16	0.15	1.11	315.04	.27	0.15	0.14	1.08	332.46	.28
STAAR composition x Race/Ethnicity (Hispanic = 1)	0.16	0.13	1.24	334.90	.22	0.14	0.12	1.19	347.94	.24
STAAR composition x Race/Ethnicity (Black= 1)	-0.04	0.14	-0.29	330.64	.78	-0.03	0.13	-0.26	344.64	.80
STAAR composition x Race/Ethnicity (Asian = 1)	0.16	0.14	1.16	317.17	.25	0.06	0.13	0.46	329.73	.64

Note. The variables were entered as predictors of the STAAR composition score in grade 4.

Table S8

Diagnostic bias of TWW scores

Variable	*B*	*SE*	*t*	*df*	*p*	*OR*	*LL*	*UL*
Intercept	-2.49	0.36	-6.93	348.97	<.001	0.08	0.04	0.17
TWW	-0.82	0.40	-2.06	322.86	.04	0.44	0.20	0.96
Race/Ethnicity (Hispanic = 1)	1.00	0.45	2.23	348.74	.03	2.72	1.12	6.58
Race/Ethnicity (Black = 1)	0.87	0.51	1.69	344.30	.09	2.39	0.87	6.57
Race/Ethnicity (Asian = 1)	-0.16	0.67	-0.23	350.98	.82	0.86	0.23	3.18
TWW x Race/Ethnicity (Hispanic = 1)	0.31	0.48	0.63	323.52	.53	1.36	0.53	3.50
TWW x Race/Ethnicity (Black= 1)	0.17	0.53	0.33	330.22	.74	1.19	0.42	3.36
TWW x Race/Ethnicity (Asian = 1)	0.05	0.65	0.07	291.42	.94	1.05	0.29	3.80

Note. OR = Odds Ratio, LL = Lower Limit, UL = Upper Limit.

Table S9

Diagnostic bias of STAAR composition scores

Variable	*B*	*SE*	*t*	*df*	*p*	*OR*	*LL*	*UL*
Intercept	-2.81	0.46	-6.12	326.96	<.001	0.06	0.02	0.15
STAAR composition	-1.35	0.47	-2.89	297.48	.004	0.26	0.10	0.65
Race/Ethnicity (Hispanic = 1)	1.19	0.54	2.19	335.50	.03	3.28	1.13	9.56
Race/Ethnicity (Black = 1)	1.04	0.61	1.72	332.99	.09	2.83	0.86	9.33
Race/Ethnicity (Asian = 1)	0.05	0.78	0.07	345.90	.95	1.06	0.23	4.91
STAAR composition x Race/Ethnicity (Hispanic = 1)	0.43	0.55	0.78	320.45	.43	1.54	0.52	4.59
STAAR composition x Race/Ethnicity (Black= 1)	0.64	0.57	1.13	315.52	.26	1.90	0.62	5.78
STAAR composition x Race/Ethnicity (Asian = 1)	0.14	0.79	0.17	301.05	.86	1.15	0.24	5.47

Note. OR = Odds Ratio, LL = Lower Limit, UL = Upper Limit.