Lessons Learned about Evaluating Fairness from a Data Challenge to Automatically Score NAEP Reading Items

Maggie Beiting-Parrish
The Federation of American Scientists

John Whitmer
The Federation of American Scientists

Follow this and additional works at: https://www.ce-jeme.org/journal

Part of the Accessibility Commons, Educational Assessment, Evaluation, and Research Commons, Educational Methods Commons, Educational Psychology Commons, Educational Technology Commons, and the Language and Literacy Education Commons

Recommended Citation
Beiting-Parrish, Maggie and Whitmer, John (2023) "Lessons Learned about Evaluating Fairness from a Data Challenge to Automatically Score NAEP Reading Items," Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语期刊: Vol. 4: Iss. 3, Article 5.
DOI: https://doi.org/10.59863/NKCJ9608
Available at: https://www.ce-jeme.org/journal/vol4/iss3/5
Lessons Learned about Evaluating Fairness from a Data Challenge to Automatically Score NAEP Reading Items

Maggie Beiting-Parrish a, and John Whitmer a

a The Federation of American Scientists

Abstract
Natural language processing (NLP) is widely used to predict human scores for open-ended student assessment responses in various content areas (Johnson et al., 2022). Ensuring algorithmic fairness based on student demographic background factors is crucial (Madnani et al., 2017). This study presents a fairness analysis of six top-performing entries from a data challenge involving 20 NAEP reading comprehension items that were initially analyzed for fairness based on race/ethnicity and gender. This study describes additional fairness evaluation including English Language Learner Status (ELLs), Individual Education Plans, and Free/Reduced-Price Lunch. Several items showed lower accuracy for predicted scores, particularly for ELLs. This study recommends considering additional demographic factors in fairness scoring evaluations and that fairness analysis should consider multiple factors and contexts.

Keywords
Automated Scoring; Reading Assessment; K-12 Assessment; National Assessment of Education Progress; Algorithmic Fairness

Acknowledgements: We would like to thank all of the original contest winners for their contributions: Scott Crossley, Joon Suh Choi, Yanisa Scherber (Georgia State University); Andrew Lan (UMASS Amherst), Nigel Fernandez, Aritra Ghosh, Naiming Liu, Zichao Wang, Benoît Choffin, Richard Baraniuk; Chris Ormerod, Amir Jafari, Mackenzie Young, Susan Lottridge (Cambium); Prathic Sundarajan (Georgia Tech); Suraj Rajendran (Weill Cornell); Arianto Wibowo, David Vaughan, Derek Justice, Yong He, Corey Palermo (Measurement Incorporated); F. Zehner, S. Gombert, N. Andersen, & U. Kröhner (DIPF).

1 Context and Literature Review

1.1 Natural Language Processing & Automated Scoring

Using Natural Language Processing (NLP) to automatically score (AS) essays and constructed response items is one of the most widely deployed uses of artificial intelligence in education and is the most frequently researched area of machine learning in educational measurement (Zheng et al., 2023). For over a decade, NLP has been demonstrated to meet (or, in some cases, exceed) human inter-rater reliability (Hewlett Foundation, 2012; Lottridge & Young, 2022) and AS is used operationally in state-level and individual high-stakes assessment to score written examinee work products in the form of short responses and longer essays, for example, as seen within the Duolingo English Test (LaFlair & Settles, 2019) and the GMAT (Burstein et al., 1998). As assessments increasingly seek to use more authentic and complex items, researchers and assessment service providers have begun to integrate these approaches into the item development workflow and use NLP results to provide insights useful for item development and piloting (Rotou & Rupp, 2020). The following article demonstrates the use of AS for accurately scoring constructed responses from the NAEP Reading test as well as presents some additional fairness analyses of these approaches.

Contact: John Whitmer. The Federation of American Scientists. Email: jwhitmer@fas.org
1.2 NAEP Reading Items Challenge Overview

The National Assessment of Education Progress (NAEP) is the largest nationally representative and continuing assessment of what students in the United States know and can do across multiple subjects (U.S. Department of Education, Institute of Education Sciences, 2017). Since 1969, NAEP has measured student achievement in mathematics, reading, science, writing, arts, and civics. NAEP uses a mix of conventional multiple-choice items, short and extended constructed responses, simulations, hands-on activities, and other item types. The number of open-ended responses varies between administrations but in large administrations includes millions of student responses that are currently scored by humans. In addition to a large number of responses, the same items are often used across administrations, creating a large and potentially repeatable context for using an NLP model.

In Fall 2021, The Institute of Education Sciences (IES) of the U.S. Department of Education conducted a data challenge\(^1\) to evaluate the potential of using AS techniques to score open-ended responses to released reading assessment items for 4th and 8th grade students. The purpose of the challenge was to determine the existing capabilities, accuracy metrics, underlying validity evidence of assigned scores, and costs and efficiencies of using automated scoring. Over two dozen teams participated in the Challenge. The results of the six awards made in the item-specific challenge are described in more detail below and the evaluation measures used are also available\(^2\). The results from these teams are used in this study and the researchers are included as contributors to the paper. The winners were identified based on the accuracy of automated scores compared to human agreement at the response level. In addition to predictive accuracy, entries were required to provide technical reports that addressed transparency, explainability, and fairness analysis of the score predictions.

1.3 Fairness in Automated Scoring

Fairness is a foundational element of education measurement American Educational Research Association et al., 2014) and therefore is a critically important aspect of any AS deployment or use of NLP- driven text modeling. Biases and unfairness within the AS pipeline can occur at many stages, ranging from initial data collection to the final model predictions (Nielsen, 2022). One of the most notable features that potentially contribute to bias is the original human rating used to train these models given by the human raters themselves; scorers frequently give lower or higher scores to student groups based on preconceived beliefs that are not relevant to the construct being measured (Amorim et al., 2018; Guerra et al., 2011). These human biases may then be incorporated into the AS model during training, leading to unfairly low scores for one or more demographic groups (Wind et al., 2018). This creates models that may perpetuate, or possibly exacerbate, the biases held by the human raters which is especially problematic for scoring engines that are used for multiple assessments or contexts. Many of these assessments have high-stakes consequences for examinees so these results could damage educational opportunities for any student groups that are scored below their true ability level (Williamson et al., 2012).

There are many examples of AS engines, as well as assessment overall, being biased against one or more demographic subgroups across different kinds of content and modalities for demonstrating knowledge such as long-form essays, short written responses, and spoken utterances. For example, one of the testing modalities that frequently demonstrates bias is speech recognition for testing English Language proficiency with some engines unfairly over-awarding points to German speakers (Loukina et al., 2013), Chinese speakers (Burstein & Chodorow, 1999) and Korean speakers (Bridgeman et al., 2012); with other engines under-scoring Spanish-speakers (Burstein & Chodorow, 1999). Turning to AS for scoring written student work, engines have been found to be biased against African American examinees (Lottridge & Young, 2022), examinees with special needs (Lottridge & Young, 2022), and English Language Learners (ELLs) (He et al., 2022; Justice, 2022; Lottridge & Young, 2022). While AS is a promising tool for grading examinee responses across a wide variety of modalities, more needs to be done to improve AS so that it is fairly awarding scores that truly capture student ability levels for students of all background groups.

\(^1\)https://github.com/NAEP-AS-Challenge/reading-prediction
\(^2\)https://github.com/NAEP-AS-Challenge/reading-prediction/blob/952ee82cc52163b17bccc1f12174ce246b8df9b/results.md
2 Methods

2.1 Challenge Design and Sample

Twenty NAEP reading assessment items that would no longer be used in assessment administrations were selected for the challenge. Table 1 provides summary statistics about the items. There was a mix of short constructed response (SCR) and extended constructed response (ECR) items. The items were chosen largely as a convenience sample with heuristic criteria to select suitable items for AS with roughly 20,000 responses set as the minimum number of student responses per item. These items represented questions for fourth and eighth grade students and were administered digitally. For each item, students were asked to read a short text selection, then use writing to respond to a question which assessed their comprehension of that text.

Table 1: Description of items included in the challenge dataset including average word count per response and the breakdown of the responses into their training, cross-validation, and test sub-samples.

<table>
<thead>
<tr>
<th>Item Title</th>
<th>Grade</th>
<th>Format</th>
<th>Avg. No. Words</th>
<th>Training N</th>
<th>Cross-Validation N</th>
<th>Test N</th>
<th>Total N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Imp. of Fast Delivery.</td>
<td>8</td>
<td>SCR</td>
<td>28.11</td>
<td>533</td>
<td>20,207</td>
<td>426</td>
<td>21,166</td>
</tr>
<tr>
<td>Support Argument</td>
<td>8</td>
<td>ECR</td>
<td>47.14</td>
<td>495</td>
<td>19,043</td>
<td>396</td>
<td>19,934</td>
</tr>
<tr>
<td>Thoreau Quotation</td>
<td>8</td>
<td>SCR</td>
<td>28.28</td>
<td>502</td>
<td>18,989</td>
<td>402</td>
<td>19,669</td>
</tr>
<tr>
<td>Describe Merchant</td>
<td>8</td>
<td>SCR</td>
<td>34.59</td>
<td>539</td>
<td>20,307</td>
<td>432</td>
<td>21,197</td>
</tr>
<tr>
<td>Innkeeper Changes</td>
<td>8</td>
<td>ECR</td>
<td>51.29</td>
<td>531</td>
<td>20,214</td>
<td>425</td>
<td>21,170</td>
</tr>
<tr>
<td>I’m Ruined</td>
<td>8</td>
<td>SCR</td>
<td>28.80</td>
<td>529</td>
<td>20,135</td>
<td>424</td>
<td>21,088</td>
</tr>
<tr>
<td>Most Imp. Char.</td>
<td>8</td>
<td>ECR</td>
<td>34.78</td>
<td>527</td>
<td>19,920</td>
<td>422</td>
<td>20,909</td>
</tr>
<tr>
<td>Reader Interested</td>
<td>8</td>
<td>SCR</td>
<td>44.20</td>
<td>543</td>
<td>20,403</td>
<td>434</td>
<td>21,380</td>
</tr>
<tr>
<td>People Care Violin</td>
<td>8</td>
<td>SCR</td>
<td>32.82</td>
<td>538</td>
<td>20,340</td>
<td>431</td>
<td>21,309</td>
</tr>
<tr>
<td>2 Pieces</td>
<td>8</td>
<td>ECR</td>
<td>44.74</td>
<td>529</td>
<td>20,038</td>
<td>423</td>
<td>20,990</td>
</tr>
<tr>
<td>Text Box Helps</td>
<td>8</td>
<td>SCR</td>
<td>32.95</td>
<td>513</td>
<td>19,665</td>
<td>411</td>
<td>20,589</td>
</tr>
<tr>
<td>Study of Mummies</td>
<td>4</td>
<td>SCR</td>
<td>18.52</td>
<td>690</td>
<td>26,564</td>
<td>552</td>
<td>27,806</td>
</tr>
<tr>
<td>Describe Merchant</td>
<td>4</td>
<td>SCR</td>
<td>18.63</td>
<td>715</td>
<td>26,977</td>
<td>572</td>
<td>28,264</td>
</tr>
<tr>
<td>Innkeeper Changes</td>
<td>4</td>
<td>ECR</td>
<td>24.63</td>
<td>682</td>
<td>26,234</td>
<td>546</td>
<td>27,462</td>
</tr>
<tr>
<td>I’m Ruined</td>
<td>4</td>
<td>SCR</td>
<td>14.55</td>
<td>666</td>
<td>25,389</td>
<td>533</td>
<td>26,588</td>
</tr>
<tr>
<td>Most Imp. Char.</td>
<td>4</td>
<td>ECR</td>
<td>16.67</td>
<td>634</td>
<td>24,509</td>
<td>508</td>
<td>25,651</td>
</tr>
<tr>
<td>Reader Interested</td>
<td>4</td>
<td>SCR</td>
<td>23.90</td>
<td>707</td>
<td>27,034</td>
<td>566</td>
<td>28,307</td>
</tr>
<tr>
<td>People Care Violin</td>
<td>4</td>
<td>SCR</td>
<td>18.53</td>
<td>694</td>
<td>26,283</td>
<td>556</td>
<td>27,533</td>
</tr>
<tr>
<td>2 Pieces</td>
<td>4</td>
<td>ECR</td>
<td>20.55</td>
<td>641</td>
<td>24,806</td>
<td>513</td>
<td>25,960</td>
</tr>
<tr>
<td>Text Box Helps</td>
<td>4</td>
<td>SCR</td>
<td>14.56</td>
<td>583</td>
<td>22,670</td>
<td>467</td>
<td>23,720</td>
</tr>
</tbody>
</table>

Note: SCR = Short Constructed Response; ECR = Extended Constructed Response

The assessment texts were a mixture of literary and informational passages which required students to respond to prompts around identifying literary elements or themes and defending claims with evidence and information from the text. The assessment item types were either Short Constructed Responses (SCR) which require the examinee to write a paragraph or Extended Constructed Responses (ECR) which require the examinee to write a longer response, such as an essay. These items are also broken down into three levels of thinking that are needed to answer the question; in order of increasing perceived difficulty and complexity of thinking these are “locate and recall”, “integrate and
interpret,” and “critique and evaluate.” The challenge items contained representative items for all three types of reading comprehension.

This challenge allowed for participants to submit either a generic model for scoring all twenty items and/or item-specific models for each question. The current challenge found that item-specific models were far more accurate, so the top six teams were all those who submitted item-specific models. Of these challenger groups, two represented commercial testing companies, and four represented teams from universities, with a mixture of experts in the field as well as graduate students.

The challenge invited participants to create predictive models of the human scores assigned to the twenty reading items. The original raw data contained only student responses that were double scored by humans. This dataset was randomly split by the NAEP data science team into a training dataset (with a small percentage of items double-scored), a cross-validation set, and a test dataset which had human scores redacted. The test and cross-validation datasets were initially about the same average size (~N=1,181); however, an average of 23,000 student responses that had only been single-scored were added to each cross-validation set to increase the number of human-scored samples to validate engines with. The test dataset was examined to ensure it contained appropriate distributions of the demographic categories and was resampled if any demographic category was lower than the original sample. Challenge participants were also provided with detailed item information that included scoring guides used by human scorers, anchor papers that were used as examples for how score points were awarded and reading item framework information. Each response contained: the item ID, text entered by the student, assigned score, student race/ethnicity, and student sex/gender.

The test dataset was used to evaluate the degree of agreement between the human and machine scores, using Quadratic Weighted Kappa (QWK) to compensate for unbalanced score distributions, which is a common approach used in automated scoring (Williamson et al., 2012). Accuracy was evaluated at the student response level across items. In addition to score predictions, participants were also required to submit a technical report which provided training data results, including detailed descriptions of the data transformations and algorithmic choices made as well as fairness analysis based on race/ethnicity and gender. Additional demographic variables (English Language Learner Status, Individualized Educational Plan Status, Free and Reduced-Price Lunch) were available to the challenge administration team but were withheld from participants due to IES disclosure restrictions; these additional demographics were used for part of the present analysis.

2.2 Response-Level Accuracy Results

The submissions were scored for the accuracy of each predicted response compared to the human score assigned to that same response. Figure 1 shows the accuracy levels of predicted responses in relationship to the QWK between human raters for the top six challengers with accuracy scores closest to the human scores. The star symbol denotes grand prize winners. The agreement of automated scores with human scores is typically slightly lower than agreement between the two human scorers. A decrease of less than 0.05 is considered acceptable to use automated models in practice (Williamson et al., 2013). All awarded submissions have accuracy levels within this range.

2.3 Response-Level Fairness Results

Evidence of Scoring Bias. All of the challengers provided results around analyses of the training data from their scoring models for different racial/ethnic groups and by gender. One of the most popular methods for assessing fair scoring in general in educational measurement (especially as related to Item Response Theory (IRT), which is used by NAEP) is the use of Differential Item Functioning (DIF) (Österlind & Everson, 2009). In this case, Standardized Mean Difference (SMD) was used to examine whether the engines were adding additional bias over and above the bias potentially present in the human scores. SMD is the absolute value of the difference between the mean of the automated score and the mean of the human score which is divided by the square root of the difference between the variance of automated score and the variance of the human score (Williamson et al., 2012). Standardized Mean Difference values that are as close to 0 as possible indicate that the automated scoring engine is not awarding scores in a more biased fashion than the original human scores; the original human scores may already reveal bias, but this measure of SMD is to ensure the engine is not adding additional bias into the measurement model. In the field, an SMD higher than 0.10 has been established as the
threshold for a biased score that should not be used in a test administration (Williamson et al., 2012). The challengers were looking for whether or not the scoring algorithm was unfairly biased (meaning, more inaccurate scores) against one or more racial/ethnic groups and whether it was biased against male or female examinees as these were the only demographic variables given to the challenge participants due to IES student data privacy requirements. The process of conducting these kinds of analysis is also related to the larger psychometric concept of construct-irrelevant variance; essentially, any element of the testing experience that may be more accessible to one or more groups of examinees while being less accessible to other group(s) (AERA et al., 2014). Thus, it is critical to examine where there are any differential effects of performance based on student characteristics so that researchers and test developers can better investigate item properties that may unfairly disadvantage any examinee subpopulation.

**Average Race/Ethnicity & Gender Results by Team.** All of the winning challengers used absolute Standardized Mean Difference (SMD) to test for a difference in scores between the machine learning-assigned scores and the original human scores. None of the challengers reached this threshold on average across response predictions for any of the primary demographic groups and in fact, all winning entries were below 0.05 with a range of 0.031-0.086. The highest values of SMD were observed for the “Other” Demographic category (Asian, Native Hawaiian or Pacific Islanders, American Indians or Alaskan Natives, and those with two or more ethnic/racial identities) with challengers finding SMD values of 0.057 - 0.086 for this category. However, these results are likely due to a small sample size. Turning to the results by gender, the top six winning challengers found an average SMD of 0.042-0.049 between genders which again is far below the threshold of 0.10.

**Additional Item-Level Analysis for Accuracy and Fairness.** Overall response accuracy was used to evaluate challenge submissions; however, in an operational administration, each item’s scoring model is evaluated for accuracy to ensure that it can be used reliably. Therefore, the NAEP challenge evaluation team conducted additional analysis at the item level across the top-scoring submissions using additional demographic criteria. These criteria included English Language Learner (ELL) Status, Individual Education Plan (IEP), and Free and Reduced-Price Lunch status (removed as there was no item for which DIF exceeded the threshold value).

It was anticipated that overall predictive accuracy and SMD by demographic group would be relatively consistent
with each other. Additionally, in the case where a machine learning-predicted response is inaccurate, an additional consideration was that there could be variance between subgroups due to random error, not due to a systematic bias, so a deeper analysis was warranted. As demonstrated in Figure 2, this assumption held true for some items and was violated in others. The “Importance of Fast Delivery” item could not accurately be scored by any of the six teams (as found in the Overall Accuracy bar represented in blue in Figure 2), and for most teams there was a larger SMD for students with English Language Learner status (ELLs) and an Individualized Educational Plan (IEP) as compared with any other demographics. Similar findings were observed in several other items that were not scored accurately by many teams.

A surprising result was that the fairness analysis did not identify increased inaccuracy by the included (and most common) demographic categories of race/ethnicity and gender which were included in the dataset provided to participants (and used for most NAEP analyses), but instead ELL status had the greatest degree of inaccurate results, comprising a full 50% of the biased results. In hindsight, it seems reasonable that a student’s history of learning a language would be a salient factor in how they express their English reading comprehension with text responses also written in English.

Nonetheless, it is concerning that Automated Scoring produces inaccurate (not just different) results above the human-awarded scores. There are many possible explanations for this result: the models may not have been biased for the demographic categories of race/ethnicity and gender/sex, the challenge participants may have fixed other bias issues that were found during model training, or because there are very few ELLs and students with IEPs represented in the larger NAEP sample, the ML engines may just have had fewer examples of writing from these groups and thus have less training experience in grading these students’ essays.

Additionally, several items were administered to both 4th and 8th grade students. These items also demonstrated
differences in scorable entries across grades (see Figure 2). For example, the “Text Box Helps” item was used for both fourth and eighth grade but only the fourth-grade version of the exam had higher SMDs for ELLs while the eighth-grade version did not. In an additional example, the “Most Important Character” item was also used across both grades but demonstrated higher average SMDs in ELLs for five of the six winning groups for the eighth-grade responses whereas, for the fourth-grade response, there were higher average SMDs for White students’ responses to the item for one of the challenge groups. These findings were very interesting especially since the challengers developed item-specific models for grading each version of the items used across both grades. This suggests that perhaps these engines are focusing on different aspects of each grade level’s writing, even though these should theoretically represent the same question responses. More certainly needs to be understood about the differences between models used for the different samples of students for the same item, such as post hoc interpretability analyses in which the authors trace the variables used to make different score predictions.

Turning to a more detailed analysis of the mean (and standard deviation) SMD by each demographic group (Figure 3), the demographic category with the highest mean SMD values was the ELL group with four items demonstrating an average SMD across challengers over 0.10. The “Text Box Helps (4)” item and “Most Important Character (8)” item had very high average SMDS for ELLs over 0.18. The item with the most average SMD values over 0.10 was the “Thoreau Quotation Effective” item with an average SMD getting as high as 0.336 for the “Other Racial” category which is a mixture of examinees who are Asian, Native Hawaiian or Pacific Islanders, American Indians or Alaskan Natives, and those with two or more ethnic/racial identities. This group was created to represent these three other racial/ethnic groups because these were too small on their own for adequate sample sizes for statistical analysis. Clearly, more research needs to be done to understand these groups separately because examinees from these differing backgrounds may have very diverse experiences and knowledge, so they may respond to prompts differently and also possibly use different linguistic patterns, etc. which are being evaluated by the ML engine differently than human scorers. Additionally, since the challengers were able to access the sex/gender and race/ethnicity of the examinee data during their model development, it’s possible they were able to adjust their models to remove bias based on these categories. This is in contrast to the instances where the challenge participants had no access to the other characteristics, thus were not able to adjust for these, possibly leading to more instances of ML scoring bias for these subgroups.

3 Conclusions

Overall, these challenge submissions showed that AS can be used to accurately score student written work around Reading Comprehension for differing levels of text complexity from fourth and eighth grade examinees. Overall, this challenge evaluation also suggested that different kinds of student writing, both responses to fiction and informational texts, are able to be scored almost as accurately as human raters by item-specific ML engines. Additionally, the majority of these items were fairly scored for all demographic subgroups included in the challenge dataset, with the exceptions noted in Figures 2 and 3.

This study has several implications for the design and evaluation of AS engines for scoring open-ended student responses. The study identified modeling approaches that provide the most accurate results within reading items and broader implications around how to evaluate those models robustly for fairness. It also showed that item-specific ML engines can be designed to accurately score student work almost as well as humans in the majority of cases. All participants used transformer-based models or other deep learning architectures, demonstrating that these approaches have largely replaced prior approaches that rely on hand-engineered features. Transformer models are a unique architecture for Natural Language Processing in that these contain both an encoder and decoder layers that rely on self-attention mechanisms; these are more sophisticated than previous models because they allow for multi-head self-attention, which eliminates the need for recurrent layers (Vaswani et al., 2017) and are used in some of the more robust AI-powered systems, such as ChatGPT (OpenAI, 2023).

In thinking about the implications for fairness, while there were some scoring biases within these models for gender and ethnicity, these were small compared to those identified for other student criteria that were not included in the challenger-facing dataset. Namely, there were many items for which the scoring engine was much less accurate for English Language Learners (ELLs). This result has been observed in prior automated scoring studies of reading items
Figure 3: This heat map shows the mean and standard deviation SMD by demographic group with items highlighted in darker orange indicating higher values near and above 0.10 for the mean SMD and 0.05 for the standard deviation (SD). Darker blue indicates smaller values of Mean and SD with darker being smaller. These are the aggregated SMD values across all of the challengers’ submissions.
This present study did not investigate the cause(s) for these imbalanced scoring results between monolinguals and ELLs; this research provides evidence that suggests further investigation is needed for why ELLs consistently score lower than their monolingual English-speaking peers. A possible reason for these scoring issues is that since there are fewer examples of ELL writing compared with monolingual English samples, the engine may have a harder time assessing these examinees’ writing. There may also be an incompatibility between the kinds of passages we are asking examinees to read and respond to, and the kinds of background knowledge students must draw from to respond to these items. Finally, this dataset did not contain any information on the examinees’ prior English proficiency in reading and writing for the English Language Learner group nor did it describe whether these were recent immigrants or had been in the American school system for several years. Previous literature has found that first generation immigrants tend to achieve very low scores on standardized reading assessments (Teltemann & Schunck, 2020), which may be a possible additional contributing factor to these findings.

3.1 Next Steps and Future Directions

**Improving Demographic Variables.** There are many next steps to be considered within the context of the current findings for future analysis of NAEP constructed response items. The first and most obvious action is to include a wider range of demographic variables that are collected with consideration for the domain (e.g., language learning status to evaluate reading comprehension). This would also better help psychometricians and item developers to understand how appropriate different kinds of content are for different demographic groups with an eye toward focusing on eliminating items that may be unfairly penalizing one or more groups of students. Certainly, a reconsideration of identity that includes intersectionality and other contemporary approaches to these constructs should be considered as part of this work to redefine these categories and perform more thorough post hoc fairness analyses.

Additional future studies could also consider modeling that go beyond atomized individual criteria, following contemporary moves in the social sciences to reconsider identities as complex intersections between these criteria (Nielsen, 2022), for example an examinee could be African American, female, from an urban area, and have an IEP. There are likely a variety of different intersectional identities that all contribute to individual differences between examinees; biases in measurement are more robust when they consider multiple background identities as opposed to just one at a time (Belzack, 2023; Foulds et al., 2019). Additionally, since the variable for the “Other” demographic group was composed of several smaller ethnic/racial groups that had very small sample sizes individually, subsequent fairness analyses should examine the impact of the AS engine on these demographic groups individually whenever sample sizes permit. There is a further question as to whether we should consider student-generated categories of identity as the criteria to use to evaluate fairness (Belitz et al., 2023) since students are the ones reporting their own demographic categories for the NAEP and perhaps their understanding of these categories is different from the test development team definitions of these categories. Finally, as SMD tends to be the default metric for examining AS engine fairness post-hoc, more research could be performed on different metrics and their efficacy for detecting unfairness for different scoring engines and sample sizes.

**Improved Item Review.** In looking more into fairness as it relates to the qualities of the items themselves, it seems as though the same question used across different grades produced different biases in scoring for different demographic groups. This finding suggests that more analysis is needed about the differences in the ways that human raters score at different grade levels as well as the features of students’ writing at different grade levels as it relates to AS engine development. Another potential analysis may be exploring the development of student writing across demographic groups between grades to see if there are any significant differences there that may be contributing to these scoring discrepancies for the same item across fourth and eighth grade. It also appears that one of the items (Thoreau Quotation) was difficult to score for the challengers for three of the demographic groups; suggesting that AS could be used not only to attempt to score responses, but as a quality control check on items themselves. If these kinds of analyses were built more consistently into item development and piloting, this may eliminate these potentially biased items from the testing pool, leaving more items that can be fairly scored by the AS engine.

**Understanding the Raters.** Finally, needs to be understood about the human raters and their biases to ensure that the AS engines are not perpetuating any biases in the human-assigned scores that were used to train the model. This
may include additional studies in future to use data mining approaches to examine the scoring patterns of different raters to ensure they are not unfairly biased against one or more groups. Additionally, since some demographic groups were very small and had to be aggregated into the “Ethnic/Racial Other” category, more work needs to be done exploring how AS engines operate with smaller demographic samples; especially to identify what kinds of sample sizes are needed to accurately and fairly capture the work of these groups. Automated Scoring has the potential to be an incredible tool for educational use; however, much more needs to be done to ensure fairness and transparency to ensure that the scores are being awarded fairly across all demographic groups so that all students can achieve scores on standardized tests that accurately reflect what they know.

References


