A Fairness Evaluation of Automated Methods for Scoring Text Evidence Usage in Writing

Diane Litman[1(✉)], Haoran Zhang[1], Richard Correnti[1], Lindsay Clare Matsumura[1], and Elaine Wang[2]

[1] University of Pittsburgh, Pittsburgh, PA 15260, USA
{dlitman,colinzhang,rcorrent,lclare}@pitt.edu
[2] RAND Corporation, Pittsburgh, PA 15213, USA
ewang@rand.org

Abstract. Automated Essay Scoring (AES) can reliably grade essays at scale and reduce human effort in both classroom and commercial settings. There are currently three dominant supervised learning paradigms for building AES models: feature-based, neural, and hybrid. While feature-based models are more explainable, neural network models often outperform feature-based models in terms of prediction accuracy. To create models that are accurate and explainable, hybrid approaches combining neural network and feature-based models are of increasing interest. We compare these three types of AES models with respect to a different evaluation dimension, namely algorithmic fairness. We apply three definitions of AES fairness to an essay corpus scored by different types of AES systems with respect to upper elementary students' use of text evidence. Our results indicate that different AES models exhibit different types of biases, spanning students' gender, race, and socioeconomic status. We conclude with a step towards mitigating AES bias once detected.

Keywords: Automated essay scoring · Fairness · Argumentation

1 Introduction

With the deployment of automated essay scoring (AES) systems in both summative and formative scenarios (e.g., high-stakes testing and classroom instruction, respectively), it is important that a student's membership in a demographic group does not impact AES accuracy. While the study of AES fairness/bias has been of increasing interest, prior work has often focused on simulated rather than actual student data [22]. Also, an open question is whether AES fairness results generalize across different AI methods commonly used to build AES systems.

The research reported here was supported, in whole or in part, by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A160245 to the University of Pittsburgh. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

I. Roll et al. (Eds.): AIED 2021, LNAI 12748, pp. 255–267, 2021.
https://doi.org/10.1007/978-3-030-78292-4_21

Table 1. RTA source article, writing prompt, and an essay (evidence score of 3).

Source Excerpt: Today, Yala Sub-District **Hospital has medicine**, **free of charge**, **for all of the most common diseases**. **Water is connected to the hospital**, which also has a **generator for electricity**. **Bed nets are used** in every sleeping site in Sauri
Essay Prompt: The author provided one specific example of how the quality of life can be improved by the Millennium Villages Project in Sauri, Kenya. Based on the article, did the author provide a convincing argument that winning the fight against poverty is achievable in our lifetime? Explain why or why not with 3–4 examples from the text to support your answer
Essay: In my opinion I think that they will **achieve it in lifetime**. During the years threw **2004 and 2008 they made progress**. People didn't have the money to buy the stuff in 2004. **The hospital was packed with patients** and they didn't have a lot of treatment in 2004. In 2008 it changed the **hospital had medicine**, **free of charge**, and **for all the common diseases**. **Water was connected to the hospital** and has a **generator for electricity**. **Everybody has net** in their site. **The hunger crisis has been addressed** with **fertilizer and seeds**, as well as the **tools needed to maintain the food**. **The school has no fees** and **they serve lunch**. To me that's sounds like it is going achieve it in the lifetime

Currently three supervised learning methods dominate the AES field. *Feature-based* models require hand-crafted features for essay representation and off-the-shelf learning algorithms for model training [1,10,24,26]. While feature-based models are typically explainable and can be tightly tied to a scoring rubric, *neural network* models are increasingly popular as they often outperform feature-based models in terms of scoring accuracy and furthermore do not require any human feature engineering [9,11,23,32,37]. To create models that are both accurate and transparent, *hybrid* models combining neural network and feature-based models are also being developed [8,17,33].

In this paper, we compare these AES model types with respect to a different evaluation dimension than scoring accuracy or model transparency, namely *algorithmic fairness*. We apply three fairness measures tailored to AES [19] that have previously been used to analyze whether native language [19] or wearing face masks [18] introduces bias when English speaking proficiency is scored in an ETS testing context. We instead use these measures to analyze whether gender, socioeconomic status, and race introduce bias when essays produced by upper elementary school students are automatically scored for text evidence usage in a classroom context. Our results indicate that when evaluated using the same fairness measure, the feature-based, neural, and hybrid AES models exhibit different types of biases. We conclude with a simple example illustrating how certain AES models make it easier to mitigate AES bias once detected.

2 Essay Corpus

All AES models are trained using 2970 essays written by students in upper elementary school classrooms, using the response-to-text assessment (RTA) protocol [6]. After reading an article from *Time for Kids* about a United Nations effort to end poverty in a Kenyan village, students wrote an essay in response to a prompt encouraging them to use evidence from the article to support their claims. Table 1 shows a source article excerpt, the RTA prompt, and a student essay. After collection, essays were manually scored on a scale of 1 to 4 (low to

Table 2. Student Demographics (left)/Essay Scores (right) as "count (%)" (n = 818).

Male	Black	Free/Reduced	Score = 1	Score = 2	Score = 3	Score = 4
389 (47.6)	556 (68.0)	451 (55.1)	242 (29.6)	315 (38.5)	165 (20.2)	96 (11.7)

high) on five dimensions[1]. In particular, a team of undergraduates independently scored randomly ordered student essays from the corpus after extensive training by experts and guided by a rubric [21,30]. Here we focus only on the evidence dimension (inter-rater reliability ICC = 0.656, n = 735 essays [7]). The evidence dimension evaluates students' ability to find and use evidence from the source article (e.g., bolded phrases in the table) to support their ideas.

For our fairness evaluation, we report test results using only the sample of 818 student essays from the full corpus where we have information on student demographic characteristics (collected from the school district) in addition to the evidence scores. We focus specifically on whether the AES models might disadvantage particular groups, specifically African Americans, males, and students receiving free or reduced-price lunch. Table 2 shows the distributions of the student demographic characteristics to be investigated and the evidence scores for this sample.[2] Note that the demographics of students in our sample are roughly similar to that of the larger school district, where about 80% of students identified as Black and about 56% received free or reduced-price lunch.

3 AES Models

To score the essays in our corpus for text-based evidence usage, we use three different approaches to AES: 1) a feature-based supervised learning approach, which we refer to as AES_{rubric}, 2) a neural network approach, which we refer to as AES_{neural}, and 3) a hybrid approach combining a neural network and hand-crafted features, which we refer to as AES_{hybrid}.

AES_{rubric} uses traditional supervised machine learning (a random forest classification algorithm with max-depth = 5, implemented in Weka) with features hand-designed to align with the RTA evidence grading rubric. As detailed in [29,36], the features are automatically computed using natural language processing:

Number of Pieces of Evidence: the number of topics in the source article that are (semantically) mentioned in the essay.
Concentration: whether an essay elaborates on the source article topics.
Specificity: for each article topic, the number of specific examples (semantically) mentioned in the essay.
Word Count: the number of words in the essay.

[1] Analysis, Evidence, Organization, Style, Mechanics/Usage/Grammar/Spelling.

[2] Students in our sample also identified as Hispanic (22.0%), Native American (11.5%), Asian (4.3%), Hawaiian (2.0%) and White (12.1%). These categories are not mutually exclusive. We focus on African American students in our study as this was the only subgroup that was large enough (had sufficient data) for our analyses.

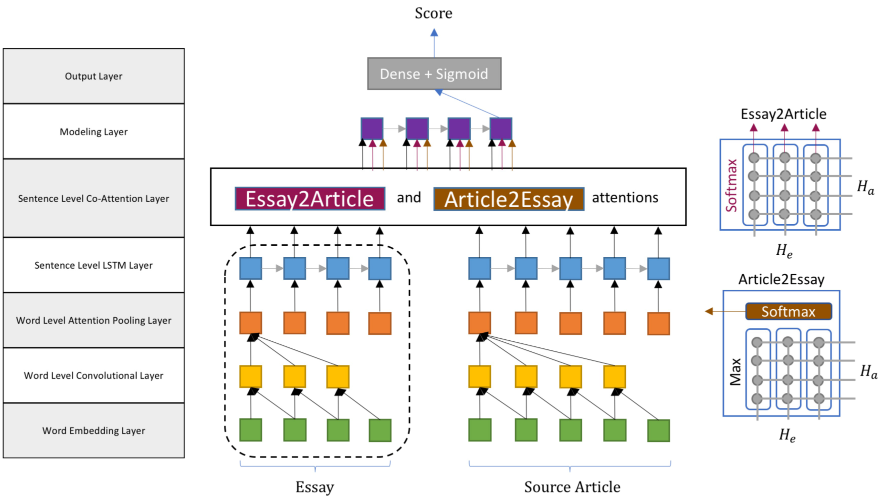

Fig. 1. Architecture of AES_{neural}, a co-attention based neural network [37].

Although the hand-crafted features of AES_{rubric} provide useful information for generating formative feedback in an accompanying automated writing evaluation (AWE) system [38], in order to improve stand-alone AES performance, a neural approach requiring no manual feature engineering and not restricted to the RTA was later developed [37]. As shown in Fig. 1, this model (AES_{neural}) uses a hierarchical neural network with a self-attention mechanism (in the dashed rounded box, originally designed for holistic scoring [9]), and adds a co-attention mechanism to support source-based scoring [37].[3]

To achieve high scoring performance yet provide some model transparency, in this paper we introduce AES_{hybrid},[4] a variant of AES_{neural} that enables the combination of hand-crafted features on any level of the hierarchical self-attention model. AES_{hybrid} offers the neural network the ability to model the features, and also no longer requires a source article. Figure 2 shows the combination of a hand-crafted feature at the word-level of the neural hierarchy. Since the released code computes hand-crafted linguistic features applicable to many AES tasks [13], we use feature selection to pick the following subset of 4 features:[5]

[3] https://github.com/Rokeer/co-attention.

[4] https://github.com/Rokeer/hybrid.

[5] We select one subset of features (from the set computed by the code release) that works for general AES purposes. Specifically, we introduce data from more prompts, including a second RTA prompt and eight prompts from the ASAP dataset (https://www.kaggle.com/c/asap-aes/). Then, we train models with only one combined hand-crafted feature for each prompt. Last, we select features that significantly improve the base neural model on the development set for at least 6 (out of 10) prompts. The intuition is that we want to select multiple features and combine each into the best level of the model hierarchy to create a version of AES_{hybrid} that is robust, while still preserving a reasonable number of features for our experiment.

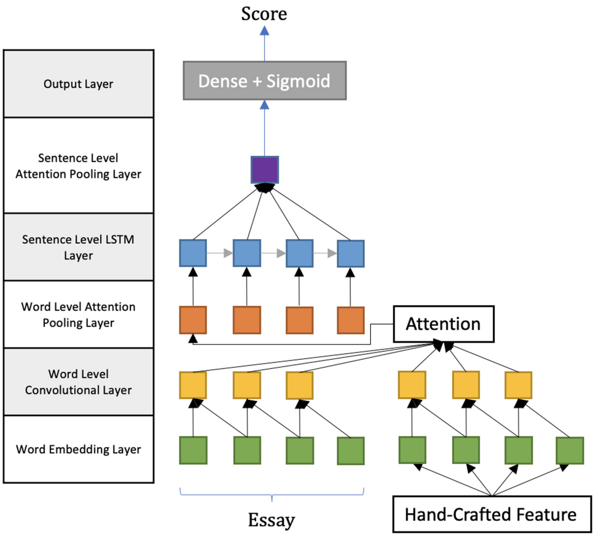
Score
Dense + Sigmoid
Output Layer
Sentence Level Attention Pooling Layer
Sentence Level LSTM Layer
Word Level Attention Pooling Layer
Word Level Convolutional Layer
Word Embedding Layer
Attention
Essay
Hand-Crafted Feature

Fig. 2. Architecture of AES_{hybrid}, a self-attention based neural network that can be combined with hand-crafted features.

Discourse Connectives: Word categories rather than words are often used to reduce feature space dimensionality. This feature labels each word as to whether it belongs to a PDTB discourse connective category [28].

Readability: This feature computes an essay's readability using the Flesch Reading Ease Test [14].

Essay and Sentence Word Counts: These features count words at both the essay level (as in AES_{rubric}) and at the sentence level.

4 AES Fairness Measures

While a variety of measures can be used to examine algorithmic fairness in education [15], the measures chosen for our evaluation are recommended for automated scoring systems [19]. In particular, Loukina et al. [19] advocate for evaluating AES fairness along multiple dimensions, arguing that total algorithmic fairness may not be achievable and that addressing fairness problems may require different mitigation strategies for different fairness dimensions. They propose three measures – overall score accuracy (OSA), overall score difference (OSD), and conditional score difference (CSD) – to capture different fairness dimensions applicable to AES. We will use these three measures to compare the fairness of our three AES models.[6]

[6] Comparing to the broader fairness literature, Loukine et al. [19] state that OSA is similar in spirit to predictive accuracy [31], OSD to standardized mean difference [34] and treatment equality [3], and CSD to conditional procedure equality [3] and differential feature functioning [39].

Overall score accuracy (OSA) measures whether AES scores are *equally accurate* across student groups compared to human scores. First, the difference in squared error between human (H) and AES (S) scores are computed: $(S-H)^2$. Fairness is evaluated by fitting a linear regression with the squared error as the dependent variable and student demographic (e.g., male) as the independent variable. The regression R^2 is used as the OSA fairness value, with statistical significance suggesting AES bias. Further, a larger R^2 indicates more impact of student group membership on score accuracy and thus less fairness/more bias.

Overall score difference (OSD) measures whether AES and human scores are *consistently different* across student groups. In order to maintain the sign of the difference, this computation uses the absolute (rather than squared) error: $S - H$. The absolute difference is now the dependent variable in the regression, with student group again the independent variable. This regression model's R^2 is the OSD fairness value, with larger R^2 again indicating less AES fairness.

Conditional score difference (CSD) is similar to OSD, but first controls for student proficiency which is approximated using the human score H. This measure is computed by fitting a regression model with absolute difference $S-H$ as the dependent variable, first with only H as the independent variable, then with both H and student group. If the difference in R^2 between the two models is statistically significant, then student group membership is having an impact on AES accuracy beyond student proficiency.

5 Evaluating AES_{rubric}, AES_{neural}, and AES_{hybrid}

We first evaluate scoring performance. Based on Sects. 1 and 3, we hypothesize (H1) that AES_{rubric}, which is purely feature-based, will be outperformed by the other two models involving neural networks. We then evaluate the same models for fairness. Based on Sect. 4, we hypothesize (H2a) that for each AES model, different fairness measures will expose different biases. We in addition hypothesize (H2b) that using the same fairness measure, different biases for each type of AES algorithm will be identified. Next, we evaluate a simple method for mitigating detected bias in models involving hand-crafted features, and hypothesize (H3) that while mitigation can indeed improve fairness, there is a scoring tradeoff. Finally, we discuss the implications of our evaluations.

Evaluating Scoring Performance. We evaluate performance using QWK (Quadratic Weighted Kappa), a standard AES evaluation measure. All reported results are obtained by training each AES model on the full corpus of 2970 essays using a 5-fold cross-validation experimental setting. While AES_{rubric} uses 4 folds for training and 1 fold for testing in each round, both AES_{neural} and AES_{hybrid} use 3 folds for training, 1 fold for development and 1 fold for testing. All neural network models are built with TensorFlow 2.2.0, and trained on an RTX 5000 GPU. Table 3 shows the neural network hyper-parameters for both AES_{neural} and AES_{hybrid}, which are based on the original self-attention model [9].

Table 3. Hyper-parameters for neural training.

Layer	Parameter	Value	Layer	Parameter	Value
Embedding	Embedding dimension	50	Dropout	Dropout rate	0.5
Sent-LSTM	Hidden units	100	Modeling	Hidden units	100
Others	Epochs	50	Word-CNN	Kernel size	5
	Batch size	16		Number of filters	100
	Initial learning rate	0.001			
	Momentum	0.9			

Table 4. Quadratic weighted Kappa between AES and human gold-standard scores.

	AES_{rubric}	AES_{neural}	AES_{hybrid}
Full corpus (n = 2970)	0.653	0.697	0.692
Demographic sample (n = 818)	0.665	0.719	0.718

Table 4 shows that the results for all AES models support hypothesis H1, whether reporting test results using all essays or only those where we have associated demographics. AES_{neural} outperforms AES_{rubric}, while AES_{hybrid} is able to maintain AES_{neural}'s QWK while increasing model transparency. Model transparency will be exploited for bias mitigation as discussed below.

Evaluating Fairness. Table 5 shows the fairness results. Support for hypothesis H2a can be seen by comparing the 3 columns under each AES model, while keeping the row constant. For example, for AES_{rubric}, CSD significantly identifies (and OSD more weakly suggests) a bias in scoring males. While OSA is unable to detect any gender bias, it is instead the only measure to (weakly) identify a problem with AES_{rubric} and socioeconomic status (free/reduced in Table 5). For AES_{neural}, only OSA suggests a problem with scoring males, while only CSD suggests a problem with scoring students based on the other types of demographics. Finally, for AES_{hybrid}, OSA is the only fairness measure to identify any bias, here for males. In addition to the R^2 values shown in the table, the sign of the coefficients in each regression (not in the table) further indicate the direction of the bias. The male and free/reduced variables all have negative coefficients, while the black variable has a positive coefficient. This means, for example, that for OSD and CSD, the results suggest lower overall AES scores for male and free/reduced lunch students compared to the human scores. Our results support the need to evaluate a given AES model for a given demographic of interest using multiple dimensions of fairness, as each yields different insights [19].

Support for hypothesis H2b can be seen by comparing the 3 columns representing the same fairness measure across the three AES models. For example, evaluations along the single dimension measured by CSD show that while AES_{hybrid} is fair, the error of AES_{rubric} is impacted by a student's gender, while

Table 5. Fairness evaluation for each AES model, using the three measures representing different fairness dimensions. Cells for OSA and OSD contain adjusted R^2 values, while CSD cells contain Δ R^2 values. The values in each row show the percentage of variance for each AES model attributed to the membership of a student in the row's demographic (e.g. Male or Not). Larger values correspond to a greater impact of the demographic on scoring error. Cells marked 'ns' mean that the effect of the student demographic is not significant at $p < .05$. Cells with values in parentheses mean that while not significant, the demographic effect is a trend at $p < .1$.

	AES_{rubric}			AES$_{neural}$			AES$_{hybrid}$		
	OSA	OSD	CSD	OSA	OSD	CSD	OSA	OSD	CSD
Male	ns	(.002)	.009	(.003)	ns	ns	.009	ns	ns
Black	ns	ns	ns	ns	ns	.004	ns	ns	ns
Free/Reduced	(.002)	ns	ns	ns	ns	.005	ns	ns	ns

the error in AES_{neural} is instead impacted by the two other demographics. Overall, our results show that while all three AES models exhibit some dimension of bias, which fairness measures detect a bias, and for which student demographic varies for each model. AES_{hybrid} seems to be our fairest AES model, with only 1 of its nine cells suggesting a problem. This is also interesting since AES_{hybrid} evaluates best with respect to balancing QWK and model explainability.

Mitigating Detected Bias. Since Table 5 suggests that gender is the most significant bias issue for our models (in terms of number of cells as well as their values), we attempt to mitigate gender bias in our models, then examine the impact of this mitigation on both the scoring and fairness measures.

One source of model bias is often a very unequal demographic distribution in the training data. While this can potentially be mitigated by resampling to create more balance, Table 2 shows that imbalance is not the case for our gender demographic. Training demographic-specific models is another approach to handling bias, but we do not have a large enough training dataset to support splitting the data in half to train two separate models.

As an alternative to resampling training data or training demographic-specific models, Loukina et al. [19] also propose manipulating the feature representation of the data, by creating a 'fairer' feature subset. To be included in this subset, a feature's values should not differ across demographics of interest, even for the same proficiency level. Such features can be identified using the CSD computation from Sect. 4, but with the feature as the regression's dependent variable. We use this method to attempt to mitigate the gender biases detected above, by creating fairer feature subsets for AES_{rubric} and AES_{hybrid}. Note that we can not apply this mitigation to AES_{neural}, as no hand-crafted features are involved.

To create our 'fairer' feature subsets, we remove all features based on word counts. Although word count is often highly positively correlated with essay

Table 6. Effect of a simple gender bias mitigation on scoring (QWK) and fairness (OSD, CSD, OSA) for AES models allowing feature removal (n = 818).

	AES_{rubric}	AES_{hybrid}	AES_{rubric}	AES_{rubric}	AES_{hybrid}
	(QWK)	(QWK)	(OSD)	(CSD)	(OSA)
Original	0.665	0.718	(.002)	.009	.009
Mitigated	0.663	0.704	ns	.006	.008

quality and thus used by many feature-based AES systems [2,5,25,27,35], in our corpus, word count is not a 'fair' feature. In particular, essay word count is significantly smaller for students who are male (141.2) versus not (175.9), even after controlling for proficiency (145.2 vs 172.3). Essay word count is thus removed from both the AES_{rubric} and AES_{hybrid} feature sets; sentence word count (only used in the AES_{hybrid} feature set) is similarly removed.

After removing the word count features, we retrain the two models that use them, with the results shown in Table 6. As hypothesized (H3), although a simple mitigation method based on using a fairer feature subset indeed slightly reduces the previously detected gender bias across AES models and fairness measures, the use of fewer features also reduces each model's scoring performance.

6 Discussion

While the identified biases in Table 5 are small (although significant), they are similar in size to those found by Loukina et al. [19]. Specifically, the percentage of variance in AES error attributed to our investigated demographics is roughly similar to the percentage of variance in automated speech scoring error attributed to native language (with OSA, OSD, and CSD values of .002, .017, and .062, respectively [19]). Aligned in some respects to our research, other studies also have identified small, but significant algorithmic bias with respect to race and gender. As described in Bridgeman [4], for example, African American men tended to receive slightly lower scores from e-rater than from human raters.

While any level of algorithmic bias is concerning and undesirable, when a detected bias is large enough to warrant mitigation is an open question, particularly if there are tradeoffs. For example, one tradeoff could be between fairness and other evaluation dimensions such as AES reliability and validity (e.g., as in our work where increasing fairness reduced reliability). A different tradeoff could be between model interpretability (transparency) and fairness. If the purpose for using AES is to generate formative feedback to improve teaching and learning, then understanding how a score was derived is critical. In this case, the more transparent rubric-based scoring model would have an advantage over the neural net model. Similar explorations of model selection have been conducted outside of AES. For example, Kung and Yu [16] examined tradeoffs between accuracy, interpretability, and fairness when using different (non-neural) machine learning models to predict college success. While they did not find that their more interpretable models compromised accuracy or fairness, like us, they did find some (small) level of bias against student groups in even the fairest models. We emphasize that if AES is used for summative evaluation purpose, for example, to assign a course grade or make a more generalized inference about a student's skill and knowledge, then it would be important to include other measures, such as human evaluation, as a check to ensure that students in a particular group whose scores might show bias are evaluated fairly [4].

The 'fairer' feature approach to bias mitigation highlights the potentially limited utility of a given mitigation method across AES paradigms. For AES_{rubric} and AES_{hybrid}, some features might have high construct validity. For example, each of the AES_{rubric} features 'number of pieces of evidence', 'concentration', and 'specificity' capture scoring rubric criteria. 'Specificity' is in fact identified as unfair, but is undesirable to remove due to its construct validity; reconstituting the algorithm or mitigating some underlying component used to operationalize 'specificity' may be possible, but this suggests a more nuanced approach than removing the feature altogether. In contrast, the unfair features based on word counts that we did remove do not correspond to any explicit rubric criteria. Finally, for AES_{neural}, creating a fairer feature subset is not even applicable as the essay representation is learned rather than based on hand-crafted features.

7 Summary and Future Work

Our main contribution is to use a multi-dimensional approach to evaluating AES fairness as the basis of a systematic *fairness comparison across three prominent machine learning-based AES methods*. A secondary contribution is the introduction of new hybrid model architecture for AES. Our AES methods vary both with respect to whether they use hand-crafted features (AES_{nubric}, AES_{hybrid}) or not (AES_{neural}), and when features are used, whether the features primarily encode rubric-specific (AES_{rubric}) or more general linguistic (AES_{hybrid}) constructs. Comparing results across AES models demonstrates that 1) all three AES models suffer from a small but significant bias on at least one fairness dimension with respect to at least one demographic, 2) when evaluated along a single fairness dimension, the biases vary across the AES models, and 3) the utility of a fairer feature strategy for bias mitigation also varies across the AES models. Also, by comparing results within a single AES model while varying fairness measures, we generalize prior findings (namely, that multiple fairness dimensions are needed as they provide different insights) from speech scoring in a testing context [19] to the very different context of evidence scoring in elementary school classrooms.

As with similar studies of algorithmic fairness, our bias conceptualization assumed that human scores represent the gold standard by which to compare AES models. We note, however, that human ratings are not necessarily bias free and may also warrant investigation. Past research, for example, has noted that trained raters react differently to the linguistic features in the essays of African American, English learners, and standard American English writers and to student characteristics such as gender and socioeconomic background (e.g., [12,20]).

An interesting future direction would be to flip the conceptualization, by exploring whether differences with a consistent and replicable AES might be a useful method for identifying bias in human scores.

References

1. Amorim, E., Cançado, M., Veloso, A.: Automated essay scoring in the presence of biased ratings. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (Long Papers), vol. 1, pp. 229–237 (2018)
2. Attali, Y., Burstein, J.: Automated essay scoring with e-rater® v. 2. J. Technol. Learn. Assess. **4**(3) (2006)
3. Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in criminal justice risk assessments: the state of the art. Sociol. Methods Res. 0049124118782533 (2018)
4. Bridgeman, B.: 13 human ratings and automated essay evaluation. In: Handbook of Automated Essay Evaluation: Current Applications and New Directions, p. 221 (2013)
5. Chen, H., He, B.: Automated essay scoring by maximizing human-machine agreement. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1741–1752 (2013)
6. Correnti, R., Matsumura, L.C., Hamilton, L., Wang, E.: Assessing students' skills at writing analytically in response to texts. Elem. Sch. J. **114**(2), 142–177 (2013)
7. Correnti, R., Matsumura, L.C., Wang, E., Litman, D., Rahimi, Z., Kisa, Z.: Automated scoring of students' use of text evidence in writing. Read. Res. Q. **55**(3), 493–520 (2020)
8. Dasgupta, T., Naskar, A., Dey, L., Saha, R.: Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In: Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pp. 93–102 (2018)
9. Dong, F., Zhang, Y., Yang, J.: Attention-based recurrent convolutional neural network for automatic essay scoring. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), pp. 153–162 (2017)
10. Ghosh, D., Khanam, A., Han, Y., Muresan, S.: Coarse-grained argumentation features for scoring persuasive essays. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Short Papers), vol. 2, pp. 549–554 (2016)
11. Jin, C., He, B., Hui, K., Sun, L.: TDNN: a two-stage deep neural network for prompt-independent automated essay scoring. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), vol. 1, pp. 1088–1097 (2018)
12. Johnson, D., VanBrackle, L.: Linguistic discrimination in writing assessment: how raters react to African American "errors," ESL errors, and standard English errors on a state-mandated writing exam. Assess. Writ. **17**(1), 35–54 (2012)
13. Ke, Z., Ng, V.: Automated essay scoring: a survey of the state of the art. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence, pp. 6300–6308. AAAI Press (2019)
14. Kincaid, J.P., Fishburne Jr, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel (1975)

15. Kizilcec, R.F., Lee, H.: Algorithmic fairness in education. In: Holmes, W., Porayska-Pomsta, K. (eds.) Ethics in Artificial Intelligence in Education. Taylor and Francis (forthcoming)
16. Kung, C., Yu, R.: Interpretable models do not compromise accuracy or fairness in predicting college success. In: Proceedings of the Seventh ACM Conference on Learning@ Scale, pp. 413–416 (2020)
17. Liu, J., Xu, Y., Zhao, L.: Automated essay scoring based on two-stage learning. arXiv preprint arXiv:1901.07744 (2019)
18. Loukina, A., Evanini, K., Mulholland, M., Blood, I., Zechner, K.: Do face masks introduce bias in speech technologies? The case of automated scoring of speaking proficiency. In: Meng, H., Xu, B., Zheng, T.F. (eds.) Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25–29 October 2020, pp. 1942–1946. ISCA (2020)
19. Loukina, A., Madnani, N., Zechner, K.: The many dimensions of algorithmic fairness in educational applications. In: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 1–10. Association for Computational Linguistics, Florence (2019)
20. Malouff, J.M., Thorsteinsson, E.B.: Bias in grading: a meta-analysis of experimental research findings. Aust. J. Educ. **60**(3), 245–256 (2016)
21. Matsumura, L.C., Correnti, R., Wang, E.: Classroom writing tasks and students' analytic text-based writing. Read. Res. Q. **50**(4), 417–438 (2015)
22. Mayfield, E., Black, A.W.: Should you fine-tune BERT for automated essay scoring? In: Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 151–162 (2020)
23. Nadeem, F., Nguyen, H., Liu, Y., Ostendorf, M.: Automated essay scoring with discourse-aware neural models. In: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 484–493 (2019)
24. Nguyen, H.V., Litman, D.J.: Argument mining for improving the automated scoring of persuasive essays. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
25. Östling, R., Smolentzov, A., Hinnerich, B.T., Höglin, E.: Automated essay scoring for Swedish. In: Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 42–47 (2013)
26. Persing, I., Ng, V.: Modeling argument strength in student essays. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Long Papers), vol. 1, pp. 543–552 (2015)
27. Phandi, P., Chai, K.M.A., Ng, H.T.: Flexible domain adaptation for automated essay scoring using correlated linear regression. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 431–439 (2015)
28. Pitler, E., Nenkova, A.: Using syntax to disambiguate explicit discourse connectives in text. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pp. 13–16 (2009)
29. Rahimi, Z., Litman, D., Correnti, R., Wang, E., Matsumura, L.C.: Assessing students' use of evidence and organization in response-to-text writing: using natural language processing for rubric-based automated scoring. Int. J. Artif. Intell. Educ. **27**(4), 694–728 (2017). https://doi.org/10.1007/s40593-017-0143-2
30. Rahimi, Z., Litman, D.J., Correnti, R., Matsumura, L.C., Wang, E., Kisa, Z.: Automatic scoring of an analytical response-to-text assessment. In: Trausan-Matu, S., Boyer, K.E., Crosby, M., Panourgia, K. (eds.) ITS 2014. LNCS, vol. 8474, pp. 601–610. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07221-0_76

31. Ramineni, C., Williamson, D.M.: Automated essay scoring: psychometric guidelines and practices. Assess. Writ. **18**(1), 25–39 (2013)
32. Tay, Y., Phan, M.C., Tuan, L.A., Hui, S.C.: SkipFlow: incorporating neural coherence features for end-to-end automatic text scoring. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
33. Uto, M., Xie, Y., Ueno, M.: Neural automated essay scoring incorporating handcrafted features. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 6077–6088 (2020)
34. Williamson, D.M., Xi, X., Breyer, F.J.: A framework for evaluation and use of automated scoring. Educ. Meas. Issues Pract. **31**(1), 2–13 (2012)
35. Zesch, T., Wojatzki, M., Scholten-Akoun, D.: Task-independent features for automated essay grading. In: Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 224–232 (2015)
36. Zhang, H., Litman, D.: Word embedding for response-to-text assessment of evidence. In: Proceedings of ACL 2017, Student Research Workshop, pp. 75–81 (2017)
37. Zhang, H., Litman, D.: Co-attention based neural network for source-dependent essay scoring. In: Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 399–409 (2018)
38. Zhang, H., et al.: eRevise: using natural language processing to provide formative feedback on text evidence usage in student writing. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 9619–9625 (2019)
39. Zhang, M., Dorans, N., Li, C., Rupp, A.: Differential feature functioning in automated essay scoring. In: Test Fairness in the New Generation of Large-Scale Assessment, pp. 185–208 (2017)