Using Item Response Theory to Measure Gender and Racial Bias of a BERT-based Automated English Speech Assessment System

Alexander Kwako
University of California, Los Angeles
akwako@ucla.edu

Elaine Wan
University of California, Los Angeles
elaine1wan@ucla.edu

Jieyu Zhao
University of Maryland, College Park
jieyuz@umd.edu

Kai-Wei Chang
University of California, Los Angeles
kwchang@cs.ucla.edu

Li Cai
University of California, Los Angeles
cai@cresst.org

Mark Hansen
University of California, Los Angeles
markhansen@ucla.edu

Abstract

Recent advances in natural language processing and transformer-based models have made it easier to implement accurate, automated English speech assessments. Yet, without careful examination, applications of these models may exacerbate social prejudices based on gender and race. This study addresses the need to examine potential biases of transformer-based models in the context of automated English speech assessment. For this purpose, we developed a BERT-based automated speech assessment system and investigated gender and racial bias of examinees' automated scores. Gender and racial bias was measured by examining differential item functioning (DIF) using an item response theory framework. Preliminary results, which focused on a single verbal-response item, showed no statistically significant DIF based on gender or race for automated scores.

1 Introduction

Automated speech assessment systems have become prominent at the K-12 and post-secondary levels (Collier and Huang, 2020; Educational Testing Service, 2005). Scores produced by automated systems are used for high stakes decisions, such as allocating public funds and determining university admissions decisions. Compared to human raters, automated assessments may be more efficient and affordable (Evanini et al., 2017), and they may improve reliability (Zechner, 2020). Yet automated assessments have a unique set of challenges (Williamson et al., 2012), and it is important that test developers and researchers continue to improve the overall enterprise of automated speech assessment.

Researchers have recently begun applying transformer-based models (Devlin et al., 2018) to English speech assessment. Largely, these research efforts have been directed towards improving the accuracy of automated scoring systems. For instance, Ormerod et al. (2021) has conducted research on BERT-based methods in automated essay scoring. In English speech assessment, Wang et al. (2021) compared the performance of BERT and XLNet for the purpose of scoring examinees' transcribed responses. Results have demonstrated that transformer-based models are highly accurate and correlate strongly with human ratings.

Although transformer-based models can produce accurate scores, less attention has been devoted to examining the biases of these models. In the field of English speech assessment, no such analyses have been conducted to date. In the broader field of natural language processing (NLP), research has demonstrated that transformer-based models can propagate and, in some cases, exacerbate gender and racial prejudice (e.g. Zhao et al., 2017; Kiritchenko and Mohammad, 2018). Biased scoring models certainly have the potential to cause allocational harms (Blodgett et al., 2020), underscoring the importance of conducting detailed analysis prior to implementation.

Beyond text modeling, there are additional sources of potential bias in audio processing. Audio speech recognition (ASR), in particular, may be

Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), pages 1 - 7
July 15, 2022 ©2022 Association for Computational Linguistics

less accurate for certain language-minority groups (e.g. Koenecke et al., 2020). Less accurate transcripts, in turn, could lead to biased scores.

There are multiple ways to measure bias, and the most appropriate method varies depending on the specific research context. Most techniques, however, are similar in that they deteremine the extent to which language modeling outputs—whether word embeddings (e.g. Dev et al., 2020) or inferences (e.g. Zhang et al., 2020)—conform to pro-stereotype expectations. This study takes a similar overall approach, but is unique in using measurement tools from educational assessment.

This study examines a type of bias known as differential item function (DIF), which is defined as the systematic difference (in scores) between a reference group and a focal (minority) group, while controlling for overall proficiency (Angoff, 1993). Although bias and fairness are conceptually distinct in educational testing, detection of DIF may provide evidence for a larger claim about fairness for certain groups of examinees (Camilli, 2006). Although analysis of DIF is common in educational assessment, it has not been applied to studies of bias in NLP.

In order to detect DIF, we use the Improved Wald Test, which is rooted in item response theory (IRT) (Cai, 2012). There are a variety of methods used to detect DIF but, in general, IRT tends to offer the most statistical power (Osterlind and Everson, 2009). The Improved Wald Test, in particular, has gained widespread adoption because it is sensitive to small group differences, holding constant examinees' overall proficiency (Woods et al., 2013).

The research design of this study involves three principal components: (1) constructing an ASR system, (2) training a transformer-based scoring model, and (3) investigating potential gender and racial bias based on these automated scores. Our analyses focus on a single speaking item. Although we found no statistically significant result in the automated scores for this item, analyses will soon be expanded to a larger pool of items and multiple grade bands which may be more susceptible to automated scoring bias.

2 Methods

Below, we describe the key methodological aspects of the research project. These include (1) the source of data used in analyses, (2) the design and development of our automated English speech assessment system, and (3) the statistical techniques used to measure gender and racial bias.

2.1 Data

This study draws on data from the English Language Proficiency Assessment for the Twenty-First Century (ELPA21), a collaborative of 7 state education agencies in the United States (Huang and Flores, 2018). Approval for this research project was granted by the consortium and the a university institutional review board. Confidentiality agreements and ethical considerations prevent sharing test items or student-level data publicly.

For test items in the speaking domain, students speak into a microphone, and their responses are recorded and subsequently sent to a third party to be scored. Currently, all verbal responses are scored by human raters.

For this study, we selected a single speaking item that was administered to students in grade 2-3. This particular item elicited responses that were short in duration (median response length = 4.8 seconds). Responses were scored 0, 1, or 2, with the highest score being given to examinees who correctly answered the question, even if small grammatical mistakes were made. A score of 0 indicated that the question was not addressed at all.

Home language was used as an indicator of race because it afforded several advantages. First, it was more fine-grained, i.e., included more categories, than the alternative indicator of race. Second, it was more related to examinees' speech, which was a focal point of the study. Home language does not necessarily indicate cultural identity, however, or native language. Respondents whose home language had fewer than 200 responses were removed from analysis.

2.2 Automated Speech Assessment

Chen et al. (2018) enumerate four components of automated speaking assessment systems. These include (1) an automated speech recognition (ASR) system, which includes speech-to-text transcription, (2) the extraction of linguistic features from audio and text data, (3) a filter model to remove non-scorable responses, and (4) a scoring model to combine linguistic features into a single score. Below, we discuss each of these components in turn.

ASR System We compared the performance of several ASR systems, based on both accuracy and efficiency (see Appendix A for details). Ultimately,

	n	%
GENDER		
Male	$4,988$	52.5
Female	$4,517$	47.5
LANGUAGE		
Spanish	$6,881$	72.4
Russian	858	9.0
Vietnamese	440	4.6
Chinese	420	4.4
Ukrainian	381	4.0
Arabic	321	3.4
Persian	204	2.1

Table 1: Descriptive Statistics of the Sample

we opted to use Google's speech-to-text service to generate text transcripts from examinees' speech. Of the $10,147$ total responses, Google produced $9,505$ non-blank transcripts, all of which were included in analyses. Descriptive statistics of the sample, disaggregated by gender and home-language, are presented in Table 1.

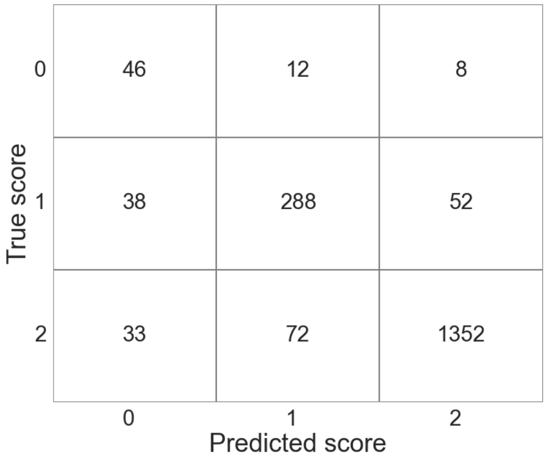
To assess Google's transcription accuracy for young, non-native speakers, we sampled 100 responses, listened to examinees' audio recordings, and manually transcribed them. Treating our own annotations as ground truth, we measured the word error rate (WER) of the Google-generated transcripts. We determined the average WER to be 22.3%—close to human parity for non-native speech, which typically ranges from 15-20$\%$ (Zechner, 2009).

Feature Extraction Linguistic features were not manually specified, but were embedded latently in the BERT scoring model.

Filtering Blank transcripts were not included in model training or analysis of bias. In some cases, blank transcripts were the result of silent audio files; in other cases, however, Google returned blank transcripts when it failed to detect speech (e.g. when examinees whispered into the microphone). 642 blank transcripts were removed from analyses.

Scoring Model We compared BERT and RoBERTa as two potential scoring models. Selection of the scoring model was based on the accuracy of models' predictions of examinees' scores on the test dataset. Because the particular speaking item that we studied was imbalanced (e.g., 76.6% of responses were scored a 2), we chose to use a cross-entropy loss function, weighted inversely to the marginal frequency of the scores. Scoring models were trained for 10 epochs. Batch size, dropout ratio, and learning rate were set to 128, 0.1 and $2 \cdot 10^{-5}$, respectively. Data were split $80\%/20\%$ for training and testing sets.

Averaged across 3 random seeds, the most accurate model was the BERT model. Test set accuracy for BERT was 88.85%, marginally higher than RoBERTa. Figure 1 presents the confusion matrix of true and predicted scores using the above scoring model for the test dataset. Details regarding the series of experiments to optimize model performance may be found in Appendix B.



True score \ Predicted score	0	1	2
0	46	12	8
1	38	288	52
2	33	72	1352

Figure 1: BERT Confusion Matrix.

The automated scoring model was found to be slightly more consistent than human raters. The Spearman Correlation Coefficient among human raters was calculated to be $\rho = 0.81$ (based on $n = 1,929$ doubly-scored responses). By comparison, the Spearman Correlation Coefficient of 2 BERT models, whose starting values and test-train splits were determined by 2 different random seeds, was found to be $\rho = 0.88$ (based on all $9,505$ responses).

2.3 Measurement of Bias

To measure bias, we used the Improved Wald Test to examine differential item functioning (DIF) using an item response theory (IRT) framework (Cai, 2012; Woods et al., 2013). In IRT, the Wald Test is used to measure and compare differences in item parameters between two groups of examinees. For the particular test item examined in this paper, IRT parameters included one discrimination parameter, a, and two item difficulty parameters, $\mathbf{b}$. The discrimination parameter captures the variability of scores, whereas the item difficult parameters capture how difficult the item is (in this case, how difficult it is for examinees to receive a score of 1 or 2). See Cai et al. (2016) for a review of the Graded 2PL model, which was used to model this item. When weighted by the inverse of the variance-covariance matrix, the difference in a (or $\mathbf{b}$) is asymptotically distributed as χ^2.

If there is a statistically significant difference between groups' item parameters based on χ^2 values, this may indicate that scores are biased against certain groups of examinees, holding constant examinees' proficiency (Holland et al., 1993; Osterlind and Everson, 2009). In mathematical notation, DIF is present (i.e. bias against examinees is present) if and only if

$$P(correct\,response|\theta, g = 0) > \\ P(correct\,response|\theta, g = 1),$$

where $g = 0$ refers to the reference group, $g = 1$ refers to the focal group, and θ is overall proficiency. For multiple-group comparisons, multiple pairs are tested separately against the same reference group.

To take an example, if the automated system was excessively harsh toward female examinees, we would see higher $\mathbf{b}$ for female examinees (as compared to male examinees). If the automated system was less reliable for female examinees, then we would see higher a (as compared to male examinees). Since these scaled differences are distributed as χ^2, we can calculate observed p-values for each comparison.

The false discovery rate of multiple comparisons was controlled using the Benjamini-Hochberg technique (Benjamini and Hochberg, 1995), which has been shown to limit Type 1 errors to the nominal level while also maximizing statistical power (Williams et al., 1999). This approach is common in analysis of DIF using IRT (Edwards and Edelen, 2009).

3 Results

Table 2 shows the results of DIF for automated scores of one speaking item, based on gender and race differences. Reference groups were "Male" and "Spanish" as these were the two majority groups for gender and race, respectively. Results were originally ordered in decreasing value of p-observed ($\mathbf{p_{obs.}}$), as required by the Benjamini-Hochberg adjustment; however, for ease of interpretation, rows have been rearranged. In no comparison was $\mathbf{p_{obs.}}$ found to be lower than p-critical ($\mathbf{p_{crit.}}$), which indicates that none of the comparisons were statistically significant.

Two Wald Tests were conducted for each DIF comparison: one to test the significance of the discrimination parameter, a, and the other to test the significance of the difficulty parameters, $\mathbf{b}$. $\mathbf{b}$ is written in bold to indicate that it is a vector of difficulty parameters. There are two degrees of freedom for tests of differences of $\mathbf{b}$, corresponding to the two difficulty parameters. Observed p-values were calculated based on χ^2 and df.

Critical p-values were determined a-priori using the Benjamini-Hochberg adjustment. Although not shown, p-values would have been significant if any $\mathbf{p_{obs.}}$ had been lower than its corresponding $\mathbf{p_{crit.}}$. Although not presented here, there were also no significant differences found in human-rated scores, based on gender or race.

4 Conclusion and Next Steps

Transformer-based models have gained widespread attention due to their highly accurate predictions and correlations with human ratings, yet it is important that issues of fairness be addressed concurrently. Our study constitutes a step forward in automated English speech assessment by examining bias in BERT-based scoring models. Our study also demonstrates how item response theory can be used to identify differential item functioning (DIF) in the context of automated scoring—a practice that is common in educational assessment, yet uncommon in the field of natural language processing.

Although our analysis did not find any gender or race DIF in automated scores produced by our BERT-based model, we refrain from drawing general conclusions about the bias of such models for English speech assessment. In this instance, we found no evidence of bias, yet it is possible that such biases are more prominent in lengthier speaking items, for older groups of examinees, or for different language minorities. Indeed, based on research of implicit bias (Spencer et al., 2016), we might expect more bias in lengthier items or for older students. The next step of our research project is to take up these challenges by expanding DIF analyses to different types of speaking items, multiple age groups, and respondents with different home-languages.

Attribute	Ref. Group	Focal Group	Parameter	χ^2	df	$p_{obs.}$	$p_{crit.}$
Gender	Male	Female	a	0.0	1	0.8438	0.0232
			b	4.4	2	0.1082	0.0107
Language	Spanish	Persian	a	0.0	1	0.9500	0.0250
			b	9.1	2	0.0104	0.0036
		Ukrainian	a	4.9	1	0.0262	0.0071
			b	1.4	2	0.5055	0.0214
		Arabic	a	5.1	1	0.0241	0.0054
			b	2.2	2	0.3396	0.0196
		Vietnamese	a	1.0	1	0.3186	0.0179
			b	11.9	2	0.0025	0.0018
		Chinese	a	2.0	1	0.1555	0.0125
			b	2.8	2	0.2523	0.0161
		Russian	a	1.8	1	0.1747	0.0143
			b	6.8	2	0.0327	0.0089

Table 2: Differential Item Functioning of Automated Scores by Gender and Language

In addition to expanding the scope of the current analysis, next steps also include experimenting with a wider variety of transformer-based models and ASR systems. Incorporating audio data into the scoring model, for instance, may improve accuracy yet also change the behavior of the automated scoring system. If biases are detected, then there will be further opportunities to explore sources of bias and to apply debiasing techniques that have been developed for other applications of transformer-based models (Sun et al., 2019).

References

William H. Angoff. 1993. Perspectives on Differential Item Functioning Methodology. In Paul W. Holland and Howard Wainer, editors, *Differential Item Functioning*, pages 3–23. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300. Publisher: Wiley Online Library.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

L. Cai. 2012. flexMIRT: Flexible multilevel item factor analysis and test scoring [Computer software]. *Seattle, WA: Vector Psychometric Group, LLC*.

Li Cai, Kilchan Choi, Mark Hansen, and Lauren Harrell. 2016. Item response theory. *Annual Review of Statistics and Its Application*, 3:297–321. Publisher: Annual Reviews.

Gregory Camilli. 2006. Test fairness. *Educational measurement*, 4:221–256.

Lei Chen, Klaus Zechner, Su-Youn Yoon, Keelan Evanini, Xinhao Wang, Anastassia Loukina, Jidong Tao, Lawrence Davis, Chong Min Lee, Min Ma, Robert Mundkowsky, Chi Lu, Chee Wee Leong, and Binod Gyawali. 2018. Automated scoring of nonnative speech using the¬†speechratersm v. 5.0 engine. *ETS Research Report Series*, 2018(1):1–31.

Jo-Kate Collier and Becky Huang. 2020. Test Review: Texas English Language Proficiency Assessment System (TELPAS). *Language Assessment Quarterly*, 17(2):221–230. Publisher: Taylor & Francis.

Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7659–7666.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

Educational Testing Service. 2005. Test and Score Data Summary: 2004-05 Test Year Data Test of English as a Foreign Language. Technical report.

Michael C. Edwards and Maria Orlando Edelen. 2009. Special topics in item response theory. *The SAGE handbook of quantitative methods in psychology*, pages 178–198. Publisher: Sage Publications New York, NY.

Keelan Evanini, Maurice Cogan Hauck, and Kenji Hakuta. 2017. Approaches to automated scoring of speaking for k‚Äì12 english language proficiency assessments. *ETS Research Report Series*, 2017(1):1–11.

Paul W. Holland, Howard Wainer, and William H. Angoff. 1993. *Perspectives on Differential Item Functioning Methodology*, page 3–23. Lawrence Erlbaum Associates.

Becky H. Huang and Belinda Bustos Flores. 2018. The English language proficiency assessment for the 21st century (ELPA21). *Language Assessment Quarterly*, 15(4):433–442. Publisher: Taylor & Francis.

Svetlana Kiritchenko and Saif M. Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689. Publisher: National Acad Sciences.

Christopher M. Ormerod, Akanksha Malhotra, and Amir Jafari. 2021. Automated essay scoring using efficient transformer-based language models. *arXiv preprint arXiv:2102.13136*.

Steven J. Osterlind and Howard T. Everson. 2009. *Differential item functioning*. SAGE.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

Steven J. Spencer, Christine Logel, and Paul G. Davies. 2016. Stereotype threat. *Annual review of psychology*, 67:415–437. Publisher: Annual Reviews.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Xinhao Wang, Keelan Evanini, Yao Qian, and Matthew Mulholland. 2021. Automated Scoring of Spontaneous Speech from Young Learners of English Using Transformers. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 705–712. IEEE.

Valerie SL Williams, Lyle V. Jones, and John W. Tukey. 1999. Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of educational and behavioral statistics*, 24(1):42–69. Publisher: Sage Publications.

David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational measurement: issues and practice*, 31(1):2–13. Publisher: Wiley Online Library.

Carol M. Woods, Li Cai, and Mian Wang. 2013. The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement*, 73(3):532–547. Publisher: Sage Publications Sage CA: Los Angeles, CA.

Klaus Zechner. 2009. What did they actually say? Agreement and disagreement among transcribers of non-native spontaneous speech responses in an English proficiency test.

Klaus Zechner. 2020. Summary and Outlook on Automated Speech Scoring. In Klaus Zechner and Keelan Evanini, editors, *Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech*, volume 3 of *Innovations in Language Learning and Assessment at ETS*, pages 192–204. Routledge, New York, NY.

Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: quantifying biases in clinical contextual word embeddings. In *proceedings of the ACM Conference on Health, Inference, and Learning*, pages 110–120.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2979–2989. Association for Computational Linguistics.

A ASR Systems Comparison

We explored two different approaches to the automated speech recognition (ASR) task. First, we looked into publicly-accessible transcribing services provided by Cloud computing platforms. Specifically, we tried services provided by Amazon Web Service (AWS) and Google Cloud Platform (GCP). Second, we considered implementing our own ASR system, trained on our own audio data. We experimented with the Librispeech ASR Chain 1d model, a pre-trained Factorized Deep Tensor Neural Network (DTNN-F)-based chain model specifically targeting speech recognition tasks provided by Kaldi, an open-source speech recognition toolkit for speech recognition and signal processing tasks (Povey et al., 2011). Based on accuracy and speed of transcription, we opted to use Google's speech-to-text service to generate text transcripts based on examinees' speech.

B Scoring Model Optimization

We divided cleaned data into train and test datasets with proportions of 0.8 and 0.2 using sickit-learn's train-test split function for training and evaluating the NLP model. In order to get a better sense of the generality of model performance, we experimented with three different random seeds—0, 1, and 2.

We trained uncased, medium-sized BERT and RoBERT models for 10 epochs with three different random seeds during the training process. Hyperparameters batch size, dropout ratio and learning rate were set to 128, 0.1 and $2 \cdot 10^{-05}$, respectively. Accuracy on test set and training loss were averaged across the 3 different random seeds.

Model Name	**Seed**	**Test Acc (%)**	**Train Loss**
BERT	0	89.58	7.77
	1	88.74	9.47
	2	88.22	7.76
	Average	**88.85**	**8.33**
RoBERTa	0	88.69	10.24
	1	88.80	11.58
	2	88.22	10.47
	Average	88.57	10.76

Table 3: Model Performance on Score-stratified Dataset Split with Seed 0

According to Table 3, BERT performed (marginally) better than RoBERTa on both test accuracy and training loss. Overall accuracy of BERT, averaged across 3 different random seeds, was found to be 88.85% with training loss of 8.33 (compared to 88.17% and 12.42 for RoBERTa).

Therefore, we choose to use the uncased BERT base model for scoring examinees' transcripts in further experiments.