Measuring the Power of Learning.®

Research Report
ETS RR–17-42

Use of Automated Scoring in Spoken Language Assessments for Test Takers With Speech Impairments

Anastassia Loukina

Heather Buzick

December 2017

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

RESEARCH REPORT

Use of Automated Scoring in Spoken Language Assessments for Test Takers With Speech Impairments

Anastassia Loukina & Heather Buzick

Educational Testing Service, Princeton, NJ

This study is an evaluation of the performance of automated speech scoring for speakers with documented or suspected speech impairments. Given that the use of automated scoring of open-ended spoken responses is relatively nascent and there is little research to date that includes test takers with disabilities, this small exploratory study focuses on one type of scoring technology, automatic speech scoring (the *SpeechRater*^{SM} automated scoring engine); one type of assessment, spontaneous spoken English by nonnative adults (six *TOEFL iBT*® test speaking items per test taker); and one category of disability, speech impairments. The results show discrepancies between human and SpeechRater scores for speakers with documented speech or hearing impairments who receive accommodations and for speakers whose responses were deferred to the scoring leader by human raters because the responses exhibited signs of a speech impairment. SpeechRater scores for these studied groups tended to be higher than the human scores. Based on a smaller subsample, the word error rate was higher for these groups relative to the control group, suggesting that the automatic speech recognition system contributed to the discrepancies between SpeechRater and human scores.

Keywords Automated scoring; fairness; disabilities; constructed response scoring; speech; validity

doi:10.1002/ets2.12170

Technology is increasingly being used to score open-ended items in a variety of assessment contexts. The development and refinement of automated scoring engines offer opportunities to gather operational data from test takers with and without disabilities to evaluate the fairness of automated scoring prior to its use in high-stakes settings. Given that the use of automated scoring of open-ended responses is relatively nascent and there is little published research to date that includes test takers with disabilities, we undertook a small exploratory study focused on automatic speech scoring for rating spoken language assessment responses for test takers with documented or suspected speech impairments.

We evaluated the performance of ETS's *SpeechRater*^{SM} automated scoring engine (Zechner, Higgins, Xi, & Williamson, 2009) for spontaneous (i.e., open-ended) spoken responses from nonnative English speakers when scoring speaking items from the *TOEFL iBT*® test (Educational Testing Service [ETS], 2017a, 2017b). Our study explored whether an automatic scoring algorithm trained and validated on the general population of test takers is transferrable to speakers who show uncommon speech patterns like those introduced by speech impairments.

Our focal groups are defined as two separate groups. Group 1 comprises test takers who submitted documentation from a qualified evaluator (ETS, 2017a, 2017b) and were deemed by an independent panel of experts to have a speech or hearing impairment and who were eligible to receive the appropriate testing accommodation(s). Group 2 comprises test takers who either did not request or did not receive accommodations for a speech impairment but whose responses were deferred to the scoring leader by human raters because one or more of their responses exhibited signs of a speech impairment. Our control group (Group 3) comprised a sample of test takers who neither received accommodations nor exhibited signs of a speech impairment.

This study addresses the following research questions:

1. Is there a difference in agreement between automatic and human scores for speakers among the three groups at the item and speaker levels?
2. If there is degradation in SpeechRater performance among the groups, is it related to different accuracy of automatic speech recognition among those groups?

Corresponding author: A. Loukina, E-mail: aloukina@ets.org

Background

In this report, we follow the approach to evaluating the fairness of automated scoring described by Bridgeman, Trapani, and Attali (2012) and Williamson, Xi, and Breyer (2012). In this approach, evidence for fairness of automated scoring is obtained by comparing the standardized mean differences between machine and human scores for different groups of test takers. It has been previously applied to evaluating the fairness of automated essay scoring by Buzick, Oliveri, Attali, and Flor (2016). They compared the scores from humans and the *e-rater*® automated scoring engine for essays (Attali & Burstein, 2006) for test takers with learning disabilities and/or attention-deficit hyperactivity disorder and reported that the machine and humans assigned similar scores, even though there were average differences among groups with respect to essay length and spelling errors.

Zechner, Evanini, and Laitusis (2012) explored automatically measuring the oral reading proficiency of middle school students with reading-based learning disabilities. Students read text passages aloud, and the researchers developed an automated speech recognition (ASR) system based on a sample of students with and without reading-based learning disabilities to measure fluency, pronunciation, and reading accuracy. Automated scores were correlated over .9 with human scores for the number of correctly read words per minute and were moderately correlated ($r = .7$) with the percentage of correctly read words per passage. There were no subgroup-specific analyses.

To date, there has been no research into the effect of speech impairments on automated speech scoring. Yet it is highly likely that acoustic properties of impaired speech may affect the performance of different components of the automated scoring engine and ultimately reduce the accuracy of the final score computed by the system.

Automated scoring engines for spoken responses use ASR systems to obtain the transcription of the response, and previous studies have shown that lower ASR accuracy often leads to degradation in performance of the automated scoring engines (Higgins, Chen, Zechner, Evanini, & Yoon, 2011; Tao, Evanini, & Wang, 2014). Automatic recognition of unconstrained nonnative speech for the general population of test takers is a challenging task in itself, and therefore the accuracy of ASR is typically lower for nonnative speakers than for native speakers (Wang, Schulz, & Waibel, 2003). Previous studies in which researchers used a recognizer similar to the one used in this study reported a word error rate (WER) of approximately 30% (Cheng, Chen, & Metallinou, 2015; Mulholland, Lopez, Evanini, Loukina, & Qian, 2016). Notably, this accuracy is comparable to the error rate of naive human transcribers (Evanini, Higgins, & Zechner, 2010; Mulholland et al., 2016). Furthermore, in previous research, speech recognition engines trained on a general population typically performed worse on impaired speech (e.g., Christensen, Cunningham, Fox, Green, & Hain, 2012; Czyzewski, Kaczmarek, & Kostek, 2003; Tolba & El Torgoman, 2009). Christensen et al. (2012) tested different ASR systems on speech from dysarthric speakers and showed that the average WER of the system trained on the general population was approximately 80%. This number could be reduced to 50% by training a new system customized for this population.

Automated scoring engines have evaluated language proficiency using a variety of features, including fluency indicators such as length of pauses or durational patterns (Bernstein, Cheng, Suzuki, Ave, & Alto, 2010; Cucchiarini, Strik, & Boves, 1997; Franco, Neumeyer, Digalakis, & Ronen, 2000; Higgins, Xi, Zechner, & Williamson, 2011). These features have also been identified as distinguishing between pathological and control speech in studies on automatic evaluation of the intelligibility of pathological speech for clinical purposes (e.g., Kim, Kumar, Tsiartas, Li, & Narayanan, 2015; see also, for a review, Middag, Clapham, van Son, & Martens, 2014).

Because in this study the same ASR engine is used for all groups of test takers, we expected lower ASR performance for test takers with speech impairments. Furthermore, the acoustic properties of impaired speech may lead to different distributions of feature values between the control population and the test takers with disabilities and, as a result, affect the accuracy of the scoring engine. Our method tests these hypotheses indirectly by comparing the agreement between human and automated scores and the WER from the ASR engine across groups with and without speech impairments.

Method

TOEFL iBT Speaking Test

The TOEFL iBT is designed to measure the ability of nonnative English speakers to use their listening, reading, speaking, and writing skills together to respond to university-level academic tasks (ETS, n.d.-a). The computer-delivered online assessment includes four separately scored sections—listening, reading, speaking, and writing—but each task requires more than one of these skills. For example, a task in the writing section may require test takers to read, listen, and then write a response. A total score is also reported. Reported scaled scores for each section range from 0 to 30, and the reported total scaled score ranges from 0 to 120. Total testing time is 4 hours.

The speaking section includes six tasks, and testing time is approximately 20 minutes. Two of the speaking tasks are *independent speaking tasks*, requiring responses generated from test takers' opinions, perspectives, or experiences, and the remaining four are *integrated speaking tasks*, which require test takers to combine multiple language skills (e.g., listening and speaking). The spoken responses have high linguistic entropy; that is, the sequence of spoken words is largely unpredictable. Each speaking task is scored by a human rater, who uses a holistic scale ranging from 0 to 4 for each task (for a description of the TOEFL iBT assessment and scoring rubrics for speaking section tasks, see ETS, n.d.-a, n.d.-b). Approximately 10% of responses are double-scored to evaluate interrater reliability. Reported speaking section scores are the sum of scores for each of the six tasks scaled to the range of 0–30.

Automated Scoring System

At the time of this writing, automated scoring was not currently used in scoring the TOEFL iBT speaking section operationally, but SpeechRater (Zechner et al., 2009) had been employed for automatically scoring the speaking section of the *TOEFL Practice Online (TPO*™) practice test for almost a decade. SpeechRater is designed to score nonnative responses to spoken items on English proficiency assessments. A SpeechRater score represents the broad construct of English speaking proficiency, defined by a combination of speech delivery (e.g., pronunciation and fluency) and language use (grammar and vocabulary; Higgins, Xi, et al., 2011).

The system begins with an ASR system, which translates audio files into text and also outputs timing data and confidence levels. The SpeechRater version used for this study uses a commercially licensed state-of-the-art ASR system trained on 800 hours of nonnative English speech data. Natural language processing tools are then employed to extract elements from the text, which are combined into linguistic features. The final score is derived by assigning weights to the features based on a model developed from input from content experts and empirically via linear regression using data from a previous set of responses scored by human raters (Higgins, Xi, et al., 2011; Xi, Higgins, Zechner, & Williamson, 2012).

Data

We derived our sample from operational administrations of TOEFL iBT between the years 2009 and 2015. Our focal groups of interest comprised test takers with abnormal speech patterns, and we identified those test takers in three ways. We selected test takers who submitted documentation of a speech impairment when they requested testing accommodations prior to taking TOEFL iBT. We also selected test takers who submitted documentation of a hearing impairment, because they may also have deficits in speech production (Blamey et al., 2001). Finally, we selected test takers with no documented disabilities whose responses were deferred to the scoring leader by human raters because one or more of their responses exhibited signs of a speech impairment.

For our analyses, we grouped together test takers with speech or hearing impairments who received accommodations because the sample of test takers with documented speech impairments who had digital audio files available for automated was too small to report alone (Group 1). Group 2 comprises the test takers whose responses were deferred to the scoring leader. We analyzed their results separately from test takers in Group 1 because they did not receive accommodations and did not submit documentation for a disability prior to testing. Our third and final group is a control population of test takers who had no documented disabilities and who did not exhibit signs of a speech impairment (Group 3). For the purposes of this exploratory study, we treated repeat test takers as independent; we use the term test taker herein to refer to a single test session for ease of exposition. Note that test takers in the Group 1 and Group 2 samples represent less than 1% of the total test taker population across 6 years.

Group 1: Test Takers With Documented Speech or Hearing Impairments Who Received Accommodations

The speakers in this group provided documented evidence of their speech and/or hearing impairments and were approved to receive test accommodations. Qualified test takers who have speech impairments may be granted an extended time accommodation for the TOEFL iBT and/or have the speaking section omitted altogether (ETS, 2015). Qualified test takers who are deaf or hard of hearing may be granted an extended time accommodation and/or may omit one or both of the listening and speaking section (ETS, 2015). All test takers who take the speaking section have their speaking tests scored by human raters based on the same scoring rubric described earlier.

Between 2009 and 2014, a total of 1,015 test takers received testing accommodations when taking TOEFL iBT and gave permission for their responses to be used for research. This number includes test takers who submitted documentation across the full range of disabilities, not only for speech or hearing impairments. We further obtained a separate database of test takers' disabilities and accommodation(s) approved for use on TOEFL iBT between 2009 and 2014. We matched 996 test takers, representing a 98% match rate. We separated out test takers with documented speech or hearing impairments from other types of disabilities. Among the 996 test takers, 31 test takers had documented speech impairments and 86 test takers had documented hearing impairments. Recall that speech impairments can be comorbid with hearing impairments, but in our data set, individuals would be listed as having a hearing impairment whether or not they had a speech impairment as well. Note that we had data on test takers' disabilities only if they requested accommodations.

Group 2: Test Takers Who Exhibited Signs of a Speech Impairment Based on Rater Scoring Logs

The information we used to identify test takers who exhibited signs of a speech impairment but did not request accommodations came from two sources. One was a log kept by scoring leaders that covered test administrations during two different periods: (a) between November 2012 and January 2013 and (b) between September 2014 and January 2015. Together, 858 responses from 627 speakers were flagged during these two time periods. The log contained test-taker ID, item number, and qualitative notes of varying detail about the observed signs of a speech impairment. The second source was the online scoring system, which included a flag for responses deferred by raters because of a suspected speech impairment as well as test-taker ID and item number. Data from the online scoring system covered the time period between November 2013 and December 2014 (no data are available for earlier test takers) and included flags for 2,903 responses from 2,345 speakers. The two sources covered different periods of time, with a short overlap during the period between September and December 2014. A total of 319 test takers appeared in both data sets, so the total number of test takers whose scores were submitted to scoring leaders because of signs of a speech impairment was 2,653 during the time period of study. We included all scored responses from test takers in this group in our analysis.

Group 3: Control Population of Test Takers Who Had No Documented Disabilities and Who Did Not Exhibit Signs of a Speech Impairment

The data for Groups 1 and 2 include responses to a large number of different test forms as well as test takers with different first languages: The numbers of speakers in these groups were too small for further sampling. Therefore, rather than matching the control population by form and test-taker first language, we selected a random sample of 2,500 test takers from the general population from a corpus of 11,000 test takers constructed for previous research studies. That corpus contains responses to 10 forms used in six operational TOEFL iBT administrations in 2012–2013.

We also obtained the scores assigned by human raters during operational scoring to all responses from test takers in all three groups. The majority of responses were assigned a numeric score on a scale ranging from 0 to 4. A small percentage of responses were classified as *non-scoreable* owing to the poor quality of the recording or other technical issues. Table 1 shows the number of test takers with speaking scores available and the number of scored responses. The number of responses with numeric nonzero scores is also shown in Table 1.

Finally, we obtained the digital recording of each response created during test administration. Such digital audio files were available for almost all responses of test takers in Groups 2 and 3. For Group 1, depending on the nature of the accommodations, some responses had been recorded in analog form. We excluded such responses from further analysis because the difference in audio quality between the two types of recordings might have introduced an additional confound into our study.

SpeechRater Feature Extraction and Performance Evaluation

All responses with digital audio files were processed by SpeechRater using the same configuration parameters (e.g., ASR settings and reference models) as used in the operational setting for TPO. Those parameters were selected to optimize

Table 1 Total Number of Test Takers in the Original Sample and the Sample Selected for Further Analysis

	Group 1	Group 2	Group 3
Original sample			
No. of test takers in the original sample	117	2,645	2,500
No. of item responses assigned a numeric score by a human rater	685	15,464	14,995
Percentage of double-scored responses	5	10	10
No. of test takers with digital audio files[a]	59	2,640	2,500
Selected for analysis			
No. of test takers who had six numeric SpeechRater scores and six numeric nonzero human scores[a]	52	2,418	2,495
No. of responses[a]	312	14,508	14,970

[a]For our analyses of SpeechRater scores, we combined results for test takers with documented hearing or speech impairments because the samples were too small to report separately.

model performance for the general test-taker population. Because no operational scoring model exists for TOEFL iBT, we used the scoring model SpeechRater described in Loukina, Zechner, Chen, and Heilman (2015) to compute the final (continuous) score. Loukina et al.'s model was also developed using responses from the general test-taker population.

Loukina et al.'s (2015) model includes 24 features that cover different aspects of language proficiency, including fluency, pronunciation, rhythm, and language use. The predicted score is computed as a linear combination of these 24 features. The coefficients for each feature were estimated using ordinary least squares linear regression on 10,000 responses from the general test-taker population. There was no overlap in speakers or prompts between the corpus used to train the model and the corpus used in this study.

We computed the agreement between the SpeechRater scores and scores assigned from each rater using the metrics recommended by Williamson et al. (2012). These included the standardized mean difference (*d*) between human and automated scores (standardized by pooling the standard deviations from both groups) and Pearson's correlation between human and continuous automated scores (*r*). We also computed additional metrics not covered by Williamson et al. (2012). These included the coefficient of determination, R^2, which shows the proportion of the variance explained by the model, and two measures of reliability, ρ^2_{human} and $\rho^2_{\text{SpeechRater}}$, suggested by Haberman, Yao, and Sinharay (2015). The latter two measures rely on the concept of latent true scores and represent the proportional reduction in mean squared error when predicting such true scores from the observed human or SpeechRater score relative to the prediction by the constant. Higher values for either ρ^2_{human} or $\rho^2_{\text{SpeechRater}}$ indicate that the observed scores are better predictors of the true scores.

Because scores are reported operationally as the total score for the whole speaking section, we evaluated the agreement for these section-level scores (denoted speaker-level scores herein). We computed raw speaker-level scores as the sum of item-level scores for both humans and SpeechRater. Because SpeechRater currently does not assign 0 as a score, and to avoid imputation of missing scores, we included only the test takers with six numeric nonzero scores from both SpeechRater and human raters. Therefore the raw speaker-level scores in this study range from 6 to 24 rather than from 0 to 24. Under operational conditions, these raw scores are then converted to a 0–30 scale, and the final score is then reported to the test taker.

Results

Descriptive statistics for each of the three groups are reported in Table 2. Averages are reported at the speaker level, on a scale of 6–24, for both human and SpeechRater scores. The focal groups scored lower than the control group with both modes, on average, $p < .01$ in all cases (see Table 2 for *t* statistics for each comparison).

Speaker-level agreement statistics are reported in Table 3. These statistics represent agreement between human raters and SpeechRater on total speech section scores (aggregated across all six items). We report agreement statistics for all speakers who received six numeric scores. Because responses for double-scoring are selected at random, we did not have a sufficient number of double-scored responses to compute human–human agreement for the focal groups at the speaker level. Relative to the control group, the agreement statistics show less human–SpeechRater agreement and lower reliability of SpeechRater scores for Group 2 relative to Groups 1 and 3. The standardized mean difference was large for Group

Table 2 Speaker-Level Scores, by Group

Test-taker groups	n^a	SpeechRater	Human raters
Group 1: Speech or hearing impairment	52	14.28 (2.48), $t = -4.03$	13.86 (4.03), $t = -4.09$
Group 2: Deferred for signs of speech impairment	2,418	14.85 (2.29), $t = -12.22$	13.83 (2.93), $t = -26.28$
Group 3: Control (no disabilities)	2,495	15.73 (2.3)	16.16 (3.28)[b]

Note. The values in parentheses show standard deviations. For Groups 1 and 2, we also give t values for comparison between Group 3. All differences are significant after applying Bonferroni correction for multiple comparisons at $\alpha = 0.01$.
[a] Denotes test takers with six nonzero human scores and SpeechRater scores on the speaking section. [b] The average score for the total testing population between January and December 2014 was 20.2 ($SD = 4.6$; ETS, 2014), which corresponds to the raw score of 15.9.

Table 3 Speaker-Level Human–SpeechRater Agreement Statistics

Group	n	D	r	R^2	ρ^2_{human}	$\rho^2_{\text{SpeechRater}}$
1	52	.12	.83	.63	.92	.73
2	2,418	.39	.71	.38	.84	.57
3	2,495	−.15	.83	.66	.87	.76

Note. D is standardized mean difference (SpeechRater − human raters) divided by the pooled standard deviation; r is Pearson correlation coefficient. R^2 is the coefficient of determination for SpeechRater scores, while ρ^2_{human} and $\rho^2_{\text{SpeechRater}}$ show the reliability of human and SpeechRater scores, respectively, in predicting true scores.

2 but approached the threshold of .1 recommended by Williamson et al. (2012) for Groups 1 and 3. The positive values mean that, on average, SpeechRater assigned higher scores over the six speaking tasks relative to human raters for Groups 1 and 2.

Word Error Rate

We next evaluated whether the performance of the scoring engine might have been affected by the accuracy of the ASR engine. We obtained manual transcriptions for a subset of responses and used them as a reference for evaluating the hypothesis produced by the ASR engine.

Previous work on the impact of ASR accuracy on automated speech scoring has consistently shown a correlation between ASR accuracy and test-taker proficiency score with ASR WER higher for responses assigned lower proficiency scores (see, e.g., Loukina & Cahill, 2016, who reported a correlation of $r = -0.24$ between ASR accuracy and proficiency score). The accuracy of ASR may be further affected by the choice of stimulus as well as test-taker native language.

Because the goal of this experiment was to evaluate whether documented or suspected speech impairment may affect the performance of ASR in comparison to control population, when selecting responses for transcription, we tried to control for other factors that can affect ASR performance to the extent possible with our data. For Group 1, because of the small number of test takers in this group, we originally planned to obtain transcriptions for all test takers for whom we had a digital audio recording. As a result of a processing error, some test takers were omitted from this sample, and therefore we report the results for 48 out of 52 test takers considered in the previous section. For Group 2, we selected responses from 120 speakers so that the score levels for the first item were uniformly distributed with 30 responses per each score level. We also tried to minimize the total number of different forms. The final data set included 13 different forms with, on average, 9 speakers per form (min. = 5, max. = 15). The speakers represented 32 native languages, with Chinese speakers being the largest group (43 speakers), followed by Arabic (12 speakers) and Japanese (6 speakers). Finally, we selected 120 speakers from the general test-taker population. The speakers were selected from a different population than the control group described in the previous sections so that the distribution of scores, forms, and, where possible, native languages was similar or identical to the distribution of responses we selected from Group 2.

Tables 4 and 5 show the mean scores and the agreement between SpeechRater and human scores for this subset.

Comparing Table 4 to Table 2, average scores were about the same for the main and restricted samples across modes and groups. But our strategy to sample speakers to create a uniform distribution across the first item scores increased the percentage of responses scored as 1 and 4 in the samples for Groups 2 and 3. This is illustrated in Figure 1.

Table 4 Speaker-Level Scores, by Group, for Test Takers Whose Responses Were Selected for Transcription

Test-taker groups	*n*	SpeechRater	Human raters
Group 1: Speech or hearing impairment	48	14.3 (2.5)	13.5 (4.0)
Group 2: Deferred for signs of speech impairment	120	14.8 (2.8)	14.2 (4.3)
Group 3: Control (no disabilities)	120	15.4 (3.2)	15.51 (4.4)

Table 5 Speaker-Level Human – SpeechRater Agreement Statistics for Test Takers Whose Responses Were Selected for Transcription

Group	*n*	*D*	*r*	R^2		$\rho^2_{\text{SpeechRater}}$
1	48	0.18	0.83	0.62	0.91	0.73
2	120	0.16	0.84	0.65	0.92	0.74
3	120	−0.03	0.89	0.76	0.92	0.83

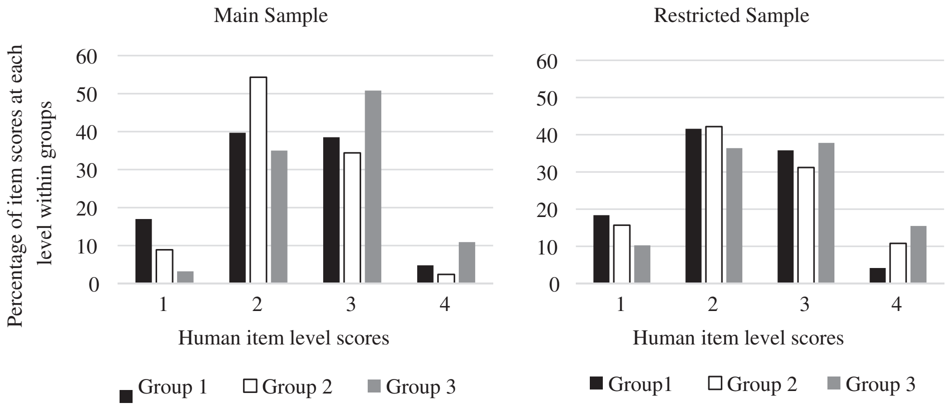

Figure 1 Distribution of item scores across groups for main and restricted samples.

Table 5 shows the human – SpeechRater agreement using the restricted sample. As can be seen from Table 5 compared to Table 3, the overall patterns of performance in terms of human – SpeechRater agreement for the restricted sample differed from what we had observed in the main corpus. Specifically, performance for Groups 2 and 3 is substantially better for the restricted sample than for the main corpus. The statistics in Table 5 show higher human – SpeechRater agreement for Group 3 relative to both Groups 1 and 2, using the restricted sample (cf. Table 3).

Computing the WER requires a comparison between the ASR output and human transcription of the digital audio responses. For Group 1, we obtained transcriptions for all responses in the sample. For Groups 2 and 3, we transcribed 360 responses per group (3 responses per speaker to Items 1, 3, and 5). The number of items was reduced from the original six items per speaker to maximize the number of speakers included in the sample, given a finite number of transcriptions. We employed a professional transcription agency to obtain expert human transcriptions for these 1,064 selected responses. We then calculated the WER between these transcriptions and the output of the ASR engine used in SpeechRater. Table 6 shows the average WER and WER per score level for each group in our study.

The WER for the control group (33%) was similar to what has been reported for this type of recognizer applied to other corpora of nonnative speech. We found that the WER was substantially greater for speakers in Groups 1 and 2 relative to the control group. Average WER was 44.4% for Group 2 and almost 50% for Group 1. This relationship was consistent across all score levels. Referring back to Table 5, the relative WER across groups is consistent with the accuracy of the automated scoring on this restricted corpus, which was better for Group 3 than for Groups 1 and 2.

Discussion

In this study, we explored the performance of automated speech scoring for speakers with documented or suspected speech impairments. We compared the control group of test takers without documented disabilities with test takers who

Table 6 Mean Automated Speech Recognition Word Error Rate for All Responses in the Transcribed Corpus as Well as the Responses at Each Score Level

Group	No. responses	Mean WER	*SD* WER	Score 1	Score 2	Score 3	Score 4
1	288	49.8	24.7	72.3	49.8	39.7	36.4
2	360	44.4	17.9	61.7	46.2	36.4	35.0
3	360	33.3	0.14	47.3	35.8	28.9	27.1

Note. The score levels are defined by human scores. WER = word error rate.

were granted accommodations due to speech or hearing impairments as well as an independent set of test takers whose scores were sent by raters to scoring leaders for review because of a suspected speech impairment. We found that the overall agreement between SpeechRater and human scores was lower for test takers with suspected speech impairments who did not receive accommodation and for test takers with documented speech or hearing impairments than for the control group.

We suggested that the accuracy of ASR may be one of the possible reasons for lower performance of the automated scoring engine. For the control group, the accuracy of the ASR was at the expected level, consistent with levels found in previous studies; yet the ASR error rate was substantially higher for test takers with documented speech or hearing impairments and for those with suspected speech impairments across all score levels. This is consistent with previous studies on ASR accuracy for impaired speech, which generally reported low performance of the engines trained on the general population. Because ASR output forms the basis for computing many features, lower accuracy of the ASR may have contributed to lower performance of the system as a whole, as measured by human – SpeechRater agreement.

Looking at the nature of disagreement between SpeechRater and human scores, we found that SpeechRater tended to assign higher scores than human raters to test takers with suspected speech impairments who did not receive accommodations. SpeechRater scores were also slightly higher on average than human scores for test takers with documented speech or hearing impairments. Several factors may have led to this pattern. These test takers have a relatively high proportion of responses scored as 1 by human raters. Such responses are likely to be overscored by SpeechRater, leading to the higher average SpeechRater score in comparison to the average human score. It is also possible that lower human scores to these responses were due to test takers failing to demonstrate the skills assessed by human raters but not currently measured by SpeechRater. We note, however, that similar to patterns observed for human scores, the average SpeechRater scores for these groups were lower than the average SpeechRater scores for the control group. In other words, while responses from test takers with suspected or documented disabilities received on average lower proficiency scores from both humans and SpeechRater than the control population, the difference between the groups was smaller for SpeechRater scores.

Although this small exploratory study provides insights into possible human – SpeechRater scoring differences for test-taker groups who are likely to have lower ASR performance, there are several noteworthy limitations. Owing to small sample sizes for the two focal groups (they represent less than 1% of the test-taking population), sampling error likely influenced our statistics. In fact, we observed differences in the patterns of human – SpeechRater agreement statistics that depended on our sampling procedure; that is, human – SpeechRater agreement statistics differed between the main and restricted corpora for Groups 2 and 3, which had different score distributions across the samples (similar means but more 1 s and 4 s in the restricted sample). This has broader implications for studies on validity and fairness: Score distributions for different subgroups are an important consideration that can affect the metrics used to measure the accuracy of the automated scoring engines and therefore need to be taken into account when comparing different subgroups. Furthermore, we were unable to confirm whether those test takers suspected of having a speech impairment by raters did or did not have a speech impairment. Finally, although our focus on human – SpeechRater agreement statistics and WER contributes to one component of a broad-based validation of automated scoring (Bennett & Zhang, 2016), it cannot provide construct validity evidence; that is, test takers' speech impairments may directly interact with some components or constructs measured by oral language assessment. This is a potential threat to the validity of inferences drawn from overall speaking scores.

In terms of the use of automated speech scoring for operational testing, despite a higher WER, SpeechRater assigned higher scores than humans, on average, for two very small and nonrandom test-taker groups. Continuing to use human raters while also employing SpeechRater is one way to minimize the threats to validity for test takers with suspected or documented speech impairments: Human raters will be able to flag a suspected speech impairment and have all responses from such test takers rerouted for further review and human scoring.

Despite the limitations and exploratory nature of this study, the results serve as a catalyst for future research. One promising area is exploring the effect of speech impairments on different features extracted by the automated scoring engine, including the number of pauses and phrase repetitions. This area of inquiry might lead to the development of an automatic screening procedure to route responses to a scoring leader after administration of the test. Another possible line of research is to explore the utility and feasibility of asking test takers to read text aloud to distinguish speech impairments from language difficulties during test administration.

Acknowledgments

The authors thank Pamela Mollaun, Jakub Novak, and Patricia Freeland for their many comments and suggestions. The authors thank Pamela Mollaun and all raters for meticulously maintaining the log of people with suspected speech impairments. The authors also thank Klaus Zechner for his comments on an earlier draft of this study and Melissa Lopez, who assisted with the bibliography review. James Bruno, Brian Hansell, and Ruth Loew helped the authors to obtain the data from various databases. Matthew Mulholland and Chi Lu assisted with processing spoken responses.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V. 2. *The Journal of Technology, Learning, and Assessment, 4*(3), 1–30.

Bennett, R., & Zhang, M. (2016). Validity and automated scoring. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 142–173). New York, NY: Routledge.

Bernstein, J., Cheng, J., Suzuki, M., Ave, S. C., & Alto, P. (2010). Fluency and structural complexity as predictors of L2 oral proficiency. In *Proceedings of Interspeech 2010, Makuhari, Chiba, Japan* (pp. 1241–1244). Baixas, France: International Speech Communication Association.

Blamey, P. J., Sarant, J. Z., Paatsch, L. E., Barry, J. G., Bow, C. P., Wales, R. J., … Tooher, R. (2001). Relationships among speech perception, production, language, hearing loss, and age in children with impaired hearing. *Journal of Speech, Language, and Hearing Research, 44*, 264–285. https://doi.org/10.1044/1092-4388(2001/022)

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education, 25*, 27–40. https://doi.org/10.1080/08957347.2012.635502

Buzick, H., Oliveri, M. E., Attali, Y., & Flor, M. (2016). Comparing human and automated essay scoring for prospective graduate students with learning disabilities and/or ADHD. *Applied Measurement in Education, 29*, 161–172. https://doi.org/10.1080/08957347.2016.1171765

Cheng, J., Chen, X., & Metallinou, A. (2015). Deep neural network acoustic models for spoken assessment applications. *Speech Communication, 73*, 14–27. https://doi.org/10.1016/j.specom.2015.07.006

Christensen, H., Cunningham, S., Fox, C., Green, P., & Hain, T. (2012). A comparative study of adaptive, automatic recognition of disordered speech. In *13th annual conference of the International Speech Communication Association 2012, INTERSPEECH 2012* (Vol. *2*, pp. 1774–1777). Baixas, France: International Speech Communication Association.

Cucchiarini, C., Strik, H., & Boves, L. (1997). Automatic evaluation of Dutch pronunciation by using speech recognition technology. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings* (pp. 622–629). New York, NY: IEEE. https://doi.org/10.1109/ASRU.1997.659144

Czyzewski, A., Kaczmarek, A., & Kostek, B. (2003). Intelligent processing of stuttered speech. *Journal of Intelligent Information Systems, 21*, 143–171. https://doi.org/10.1023/A:1024710532716

Educational Testing Service. (n.d.-a). *Frequently asked questions about the TOEFL iBT test*. Retrieved from https://www.ets.org/toefl/ibt/faq/

Educational Testing Service. (n.d.-b). *TOEFL iBT speaking section scoring guide*. Retrieved from https://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf

Educational Testing Service. (2014). *Test and score data summary for TOEFL iBT tests*. Princeton, NJ: Author.

Educational Testing Service. (2015). *Bulletin supplement for test takers with disabilities or health-related needs*. Retrieved from http://www.ets.org/s/disabilities/pdf/bulletin_supplement_test_takers_with_disabilities_health_needs.pdf

Educational Testing Service. (2017a). *Test and score data summary for TOEFL iBT tests*. Retrieved from https://www.ets.org/s/toefl/pdf/94227_unlweb.pdf

Educational Testing Service. (2017b). *Disabilities and health-related needs*. Retrieved from https://www.ets.org/disabilities/

Evanini, K., Higgins, D., & Zechner, K. (2010). Using Amazon Mechanical Turk for transcription of non-native speech. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk* (pp. 53–56). Stroudsburg, PA: Association for Computational Linguistics.

Franco, H., Neumeyer, L., Digalakis, V., & Ronen, O. (2000). Combination of machine scores for automatic grading of pronunciation quality. *Speech Communication, 30*, 121–130. https://doi.org/10.1016/S0167-6393(99)00045-X

Haberman, S. J., Yao, L., & Sinharay, S. (2015). Prediction of true test scores from observed item scores and ancillary data. *British Journal of Mathematical and Statistical Psychology, 68*, 363–385. https://doi.org/10.1111/bmsp.12052

Higgins, D., Chen, L., Zechner, K., Evanini, K., & Yoon, S. (2011). *The impact of ASR accuracy on the performance of an automated scoring engine for spoken responses*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Higgins, D., Xi, X., Zechner, K., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech & Language, 25*, 282–306. https://doi.org/10.1016/j.csl.2010.06.001

Kim, J., Kumar, N., Tsiartas, A., Li, M., & Narayanan, S. S. (2015). Automatic intelligibility classification of sentence-level pathological speech. *Computer Speech & Language*, *29*, 132–144. https://doi.org/10.1016/j.csl.2014.02.001

Loukina, A., & Cahill, A. (2016). Automated scoring across different modalities. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, San Diego, California* (pp. 130–135). Denver, CO: Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-0514

Loukina, A., Zechner, K., Chen, L., & Heilman, M. (2015). Feature selection for automated speech scoring. In *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 12–19). Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.3115/v1/W15-0602

Middag, C., Clapham, R., van Son, R., & Martens, J.-P. (2014). Robust automatic intelligibility assessment techniques evaluated on speakers treated for head and neck cancer. *Computer Speech & Language*, *28*(2), 467–482. https://doi.org/10.1016/j.csl.2012.10.007

Mulholland, M., Lopez, M., Evanini, K., Loukina, A., & Qian, Y. (2016) A comparison of ASR and human errors for transcription of non-native spontaneous speech. In *Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5855–5859). Piscataway, NJ: IEEE. https://doi.org/10.1109/ICASSP.2016.7472800

Tao, J., Evanini, K., & Wang, X. (2014). The influence of automatic speech recognition accuracy on the performance of an automated speech assessment system. In *2014 IEEE Spoken Language Technology Workshop (SLT)* (pp. 294–299). Piscataway, NJ: IEEE. https://doi.org/10.1109/SLT.2014.7078590

Tolba, H., & El Torgoman, A. S. (2009, August). Towards the improvement of automatic recognition of dysarthric speech. In *ICCSIT 2009, 2nd IEEE International Conference on Computer Science and Information Technology* (pp. 277–281). Piscataway, NJ: IEEE. https://doi.org/10.1109/ICCSIT.2009.5234947

Wang, Z., Schulz, T., & Waibel, A. (2003). Comparison of acoustic models adaptation techniques for non-native speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 540–543). Piscataway, NJ: IEEE.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Xi, X., Higgins, D., Zechner, K., & Williamson, D. (2012). A comparison of two scoring methods for an automated speech scoring system. *Language Testing, 29*, 371–394. 10.1177/0265532211425673

Zechner, K., Evanini, K., & Laitusis, C. (2012). Using automatic speech recognition to assess the reading proficiency of a diverse sample of middle school students. In *Proceedings of the Third Workshop on Child, Computer Interaction, Third Workshop on Child, Computer and Interaction (WOCCI 2012),* (pp. 45–52). Portland, OR: International Speech Communications Association.

Zechner, K., Higgins, D., Xi, X., & Williamson, D. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Science Communication, 51*, 883–895. https://doi.org/10.1016/j.specom.2009.04.009

Suggested citation:

Loukina, A., & Buzick, H. (2017). *Use of automated scoring in spoken language assessments for test takers with speech impairments* (Research Report No. RR-17-42). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12170

Action Editor: Beata Beigman Klebanov

Reviewers: Klaus Zechner and Brent Bridgeman