

Rapid #: -19446244

CROSS REF ID: **1851899**

LENDER: **NNY :: Main Library**

BORROWER: **AZS :: Main Library**

TYPE: Article CC:CCG

JOURNAL TITLE: Assessment in education

USER JOURNAL TITLE: Assessment in Education: Principles, Policy & Practice

ARTICLE TITLE: Complementary strengths? Evaluation of a hybrid human-machine scoring approach for a test of oral academic English

ARTICLE AUTHOR: Davis, Larry

VOLUME: 28

ISSUE: 4

MONTH:

YEAR: 2021

PAGES: 437-455

ISSN: 0969-594X

OCLC #:

Processed by RapidX: 8/26/2022 1:24:47 PM

This material may be protected by copyright law (Title 17 U.S. Code)

Full Terms & Conditions of access and use can be found at
<https://www.tandfonline.com/action/journalInformation?journalCode=caie20>



Complementary strengths? Evaluation of a hybrid human-machine scoring approach for a test of oral academic English

Larry Davis  and Spiros Papageorgiou 

Center for Language Education and Assessment Research, Educational Testing Service, Princeton, NJ, USA

ABSTRACT

Human raters and machine scoring systems potentially have complementary strengths in evaluating language ability; specifically, it has been suggested that automated systems might be used to make consistent measurements of specific linguistic phenomena, whilst humans evaluate more global aspects of performance. We report on an empirical study that explored the possibility of combining human and machine scores using responses from the speaking section of the TOEFL iBT® test. Human raters awarded scores for three sub-constructs: delivery, language use and topic development. The SpeechRaterSM automated scoring system produced scores for delivery and language use. Composite scores computed from three different combinations of human and automated analytic scores were equally or more reliable than human holistic scores, probably due to the inclusion of multiple observations in composite scores. However, composite scores calculated solely from human analytic scores showed the highest reliability and reliability steadily decreased as more machine scores replaced human scores.

ARTICLE HISTORY

Received 31 May 2020

Accepted 31 August 2021

KEYWORDS

Speaking assessment;
automated scoring; hybrid
scoring; analytic scoring

Introduction

Performance in tests of speaking ability has traditionally been evaluated by human raters, who judge responses on various criteria as defined in scoring rubrics and other materials such as exemplar responses. Well-trained raters, supported by appropriate scoring aids, are assumed to be capable of evaluating relevant aspects of performance so that the resulting scores are consistent and reflect the construct of language ability measured by the assessment. In recent years machine scoring has become increasingly common and when evaluating oral language proficiency, machine scoring has advantages over human scoring in terms of speed and ease of scalability, as well as consistency in the way specific aspects of performance are evaluated. However, construct under-representation is a persistent criticism of automated scoring, that is, machine scores may not support inferences regarding the full range of abilities that make up the language construct being measured. This criticism is based in part on the fact that machine scoring systems have so far been limited in their ability to evaluate the full range of phenomena that characterise

good performance in spontaneous spoken responses, particularly features that are more abstract, such as effectiveness of content, or that manifest over longer stretches of text, such as discourse organisation (Zechner, 2020).

This limitation has been an important consideration for developers of speaking tests. For example, at the time the TOEFL iBT® test was developed, technology for the automated scoring of speech was limited in the range of language features that could be measured. Given that automated systems could not fully replicate the construct coverage of human holistic scores, it was decided to use human raters exclusively for operational scoring (Xi et al., 2012) and only implement automated scoring of speech in a low-stake TOEFL iBT practice test (Xi et al., 2008). In the case of the practice test, a benefit of automated scoring technology was providing users with rapid, inexpensive feedback on their speaking performance, even if such feedback did not address the full range of the academic speaking construct evaluated in the operational test.

Where automated scoring is unable to fully address all aspects of the construct, another option is to include both human and machine evaluations in the scoring process. Three approaches in hybrid human-machine scoring have been discussed in the literature: a confirmatory hybrid approach, a parallel contributory approach and a divergent contributory approach. In a confirmatory hybrid approach, the automated score confirms a human score. If a discrepancy between human and machine scores is detected, the response is routed to a second human rater; the machine output is not used in calculating the final score. Such a confirmatory method is used in the analytical writing section of the GRE® General test (Bridgeman, 2013; Educational Testing Service, 2017). The benefit of this method is that the efficiency of machine scoring makes it practical to have greater monitoring of raters, helping to ensure score quality.

In a parallel contributory approach, human and machine holistic scores are combined and both scores contribute to the final score. This method has been used in the writing section, and more recently, in the speaking section of the TOEFL iBT test (Enright & Quinlan, 2010; Educational Testing Service, 2020a, 2020b; Ramineni et al., 2012). Machine and human scores are intended to measure the same construct and so the machine score essentially serves as a second rater. Such double scoring increases the reliability of the test and machine scores may contribute to more robust construct coverage by ensuring that specific features of performance are consistently measured.

A third approach is a divergent contributory method where raters and automated scoring systems each focus on different aspects of performance. The rationale behind the divergent contributory technique is that raters and machine scoring systems may have different and complementary strengths when it comes to scoring. Enright and Quinlan (2010) suggested that for scoring writing, humans might be better at evaluating ideas, content, and organisation, while machines might be better at evaluating specific linguistic phenomena. For speaking, it has been demonstrated that humans can have difficulty in consistently detecting specific linguistic phenomena such as word-level pronunciation accuracy (Loukina, Lopez et al., 2015). Based on such factors, Isaacs (2018a, 2018b) suggested that machines might focus on scoring spectral and durational aspects of speech (pronunciation and fluency), while humans focus on features such as task accomplishment or appropriateness. Isaacs also speculated that a further benefit of this approach might be to reduce the cognitive load on raters by simplifying the decision task. This

speculation is potentially supported by limited empirical evidence showing that the use of analytic rubrics encourages raters to evaluate specific features more consistently (Harsch & Martin, 2013).

To date, hybrid human-machine scoring has been relatively little tried in the assessment of oral language proficiency. Within one operational assessment, a low-stakes test of English for use in teaching, an automated scoring system was supplemented by human raters who scored responses flagged as unscorable by the automated system (Yoon & Zechner, 2017; Zechner et al., 2015). More recently, hybrid human-machine scoring following a parallel contributory approach was instituted for the speaking section of the TOEFL iBT test (Educational Testing Service, 2020a, 2020b). Beyond operational use, Luo et al. (2016) conducted a research study where human raters were provided with 10 machine-generated measures of pronunciation for a read aloud task, with data presented in the form of a radar chart. Novice raters were trained to recognise the ‘shape’ of plots associated with different levels of performance; then, during scoring they listened to the response, reviewed the chart and finally awarded an overall score for pronunciation. Machine scores thus informed raters’ decision-making, rather than directly contributing to the final score. Using this approach, inter-rater correlations between novices and expert raters increased to a level similar to expert-expert correlations.

Context of the study: the TOEFL iBT speaking test

For the sake of convenience, the current investigation was carried out using data collected from the TOEFL iBT® test. Specifically, a corpus of speaking responses was already available, which had been scored by human raters in a previous study. Additionally, access was available to an automated scoring system that had been developed for evaluating these types of responses (SpeechRater). The TOEFL iBT® test, the most recent iteration of the Test of English as a Foreign Language, is a four-skills test of academic English for university admission purposes. Speaking ability is measured via a series of monologic speaking tasks, including independent tasks where test takers speak about familiar topics drawing on personal experience and background knowledge and integrated tasks where test takers first read and/or listen to materials and then discuss these materials in their responses. The data used in our study were collected from this version of the test. However, in 2019 one independent and one integrated task was removed to reduce the test taking time, and automated scoring was introduced in a parallel contributory approach to increase reliability (Educational Testing Service, 2020b).

The academic speaking construct targeted by the test is most obviously captured in the scoring rubric (Educational Testing Service, 2019). Each task is scored separately, by different raters, who provide a holistic score on a scale of 0–4. Although responses are scored holistically-by-task, raters are asked to consider performance in three sub-constructs: delivery, language use, and topic development. The sub-construct of delivery includes features of fluency and pronunciation and terms like ‘flow’, ‘pacing’, ‘fluidity’, and ‘choppiness’ are used in the scoring rubric to capture features of speech rate and pausing in a way that will be accessible to raters. Several aspects of segmental and suprasegmental pronunciation

are also included through descriptors referencing features such as ‘articulation’, ‘intonation’, and ‘rhythm.’ The sub-construct of language use encompasses features related to vocabulary and grammar and scoring descriptors reference the range and complexity of lexis and syntax, as well as accuracy and precision of use. Other features mentioned include ‘automaticity’ in producing grammatical structures and coherence at the sentence level. The sub-construct of topic development includes the organisation and development of ideas. For independent tasks, the relevance of the ideas to the question asked is considered, whilst for integrated tasks, judgement is made of the accuracy and sufficiency of content reported from the stimulus materials. Coherence at the text level and precision of ideas are also addressed in this sub-construct. To the extent possible, machine scoring incorporates linguistic phenomena that reflect the same sub-constructs and in operational use, holistic scores from human and machine raters are combined to help ensure adequate construct coverage in the final scores. However, the ability of automated scoring systems to evaluate propositional content or discourse features remains relatively limited (Zechner, 2020); accordingly, under divergent contributory scoring, a possible scenario might be for humans to evaluate topic development, whilst the machine evaluates delivery and/or language use. This is the scenario we evaluate.

The current study

Although it has been suggested that a divergent contributory method has the potential to improve the quality of scores, to our knowledge, no attempt has yet been made to empirically evaluate such an approach. This paper reports an initial effort along these lines, where we evaluated the relative reliability of hybrid scores that were based on various combinations of human and machine analytic scores for different speaking sub-constructs. However, whilst the current study focuses particularly on the issue of reliability, other questions of validity related to automated scoring loom large. First is whether the automated system awards scores that are similar to what a human rater would assign in the same situation; accordingly, agreement between human and machine scores used in the study will be described. Second, and perhaps more important, is the degree to which the machine evaluates performance phenomena in the way that a human might. Making interpretations regarding the construct equivalence of machine and human scores can be difficult given the ways that information is combined in some machine learning algorithms, and the fact that the cognitive processes underlying human scoring decisions are generally not available for inspection. However, in this study, we use a machine learning approach that combines measures of linguistic phenomena in a relatively transparent way (linear regression). Additionally, we compare the linguistic measures used in generating automated scores to the language used in the scoring rubric. This analysis provides a rough indication of the extent to which automated scores for particular speaking sub-constructs focused on the same criteria expected for human raters and by extension, the extent to which machine-human hybrid scores might measure the same construct as scores generated solely by human raters.

Research questions

Given the potential benefits of hybrid scoring and lack of research on this approach within speaking assessment, we were interested in both evaluating the performance of a divergent contributory scoring approach, as well as investigating the supposition that humans and machine judgements have different strengths. Specifically, the study addressed the following research questions:

- (1) Does the reliability of analytic scores differ across analytic categories?
 - (a) Do human analytic scores for topic development differ in reliability from human analytic scores for delivery and language use?
 - (b) Do machine-generated analytic scores for delivery and language differ in reliability from human scores?
- (2) Are speaking scores constructed by combining human- and machine-generated analytic scores as reliable as human holistic-by-task scores?

Materials and methods

Test taker responses

A total of 1,200 previously scored responses to TOEFL iBT speaking tasks were used in the study, composed of 200 responses to each of the six items on the previous version of the test. The responses were collected from two different operational test forms and had received holistic scores from human raters using the standard operational scoring procedures.

The responses were divided into equal-sized training and evaluation sets. The training set was used to produce a linear regression model to predict human scores and the evaluation set was used to test the accuracy and generalisability of the machine scores produced. Testing performance on a separate dataset was used to detect regression model overfit and provided a more realistic measure of prediction accuracy under conditions of actual use (Williamson et al., 2012). The training set consisted of 600 responses from 571 unique test takers (100 responses for each item type), taken from a single TOEFL iBT speaking test form. For the training partition, a stratified random sample of responses was selected based on holistic scores awarded during the operational test administration such that, to the extent possible, an equal number of responses were taken from each score band (Table 1). The goal of this approach was to obtain a relatively flat distribution of scores to improve the precision of the model at both ends of the score range, as well as obtain independent observations from different individuals to satisfy the assumption of local independence required for regression analysis. (A total of 29 test takers were

Table 1. Responses used for scoring model training and evaluation.

	Test Takers	Items	Responses	Score Distribution (holistic scores)			
				1	2	3	4
Training Set	571	6	600	42	226	226	106
Evaluation Set	100	6	600	27	174	311	88

Note: Holistic scores were obtained during were operational administration of the test.

Table 2. Background information reported by individuals in training and evaluation datasets.

Training Set		Evaluation Set	
Total <i>N</i> (individuals)	571		100
Gender			
Female	249 (44%)	Female	39 (39%)
Male	264 (46%)	Male	54 (54%)
Unreported	58 (10%)	Unreported	7 (7%)
TOEFL iBT speaking section score (CEFR level)			
25–30 (C1)	N/A	25–30 (C1)	21 (21%)
20–24 (B2)	N/A	20–24 (B2)	42 (42%)
16–19 (B1)	N/A	16–19 (B1)	25 (25%)
10–15 (A2)	N/A	10–15 (A2)	11 (11%)
0–9 (Below A2)	N/A	0–9 (Below A2)	1 (1%)
First Language			
Chinese	155 (27%)	Korean	18 (18%)
Korean	103 (18%)	Chinese	15 (15%)
Spanish	42 (7%)	Spanish	12 (12%)
Arabic	36 (6%)	Arabic	11 (11%)
German	32 (6%)	Turkish	6 (6%)
Japanese	31 (5%)	4 individuals each (4%): French, German, Portuguese	12 (12%)
15–21 individuals each (3–4%): French, Hindi, Telugu	55 (10%)	3 individuals each (3%): English, Tagalog	6 (6%)
10 individuals each (2%) Russian, Turkish	20 (4%)	2 individuals each (2%): Farsi, Hindi, Japanese, Vietnamese	8 (8%)
34 languages with eight or fewer individuals	93 (16%)	12 languages with one individual each	12 (12%)

Note: Because mapping to CEFR levels is based on performance on all tasks of the speaking section (Papageorgiou et al., 2015), CEFR levels are only provided for test takers of the evaluation set.

repeated in the data because all responses receiving a score of 1 were included due to a low frequency of responses at this level.) The evaluation set consisted of 600 responses obtained from 100 randomly selected test takers with each test taker providing a full set of six responses. In contrast to the training set, the distribution of scores in the evaluation set approximated the distribution of scores in the operational data. Half of the responses in the evaluation set came from the same test form used for training and the other half came from a different test form to allow evaluation of model performance on an unseen test form. Table 2 shows the gender, proficiency level, and first languages of participants whose speaking responses made up the training and evaluation datasets; the composition of both groups was similar.

Human analytic scores

Each response was double-scored on the sub-constructs identified in the TOEFL iBT scoring rubric: delivery, language use, and topic development. A total of 24 TOEFL iBT scoring leaders were recruited, with each rater scoring a total of 300 responses over six scoring sessions; scoring was done outside of the system used for operational scoring, using materials provided by the researchers. Most raters completed two sessions a day over a period of 3–5 days; in each session 50 responses were scored, with scoring done for a single dimension only to minimise halo effect. Scoring rubrics for each dimension were copied from the TOEFL scoring rubric, with only the relevant dimension from the full rubric shown during scoring. Raters were also provided with 7–8 benchmark responses and accompanying scoring commentaries and a practice set of 10 previously scored

Table 3. Counts of human analytic scores for scoring model partitions, average of two raters.

Avg. Score	Delivery		Language Use		Topic Development	
	Training	Evaluation	Training	Evaluation	Training	Evaluation
1	26	24	44	33	73	39
1.5	27	37	58	35	43	39
2	149	129	159	136	131	106
2.5	96	95	81	97	88	114
3	158	151	130	148	110	133
3.5	71	92	68	79	75	93
4	73	72	60	72	80	76

Table 4. Spearman correlations between human analytic scores.

	Language Use	Topic Development
Delivery	.71	.62
Language Use		.68

Note: Correlations are the average of inter-correlations for scores from two raters.

responses, again produced with the help of assessment development staff. Raters were asked to review the rubric and benchmark responses and then attempt the practice set; if more than 50% of a rater's practice scores disagreed with reference scores, then the rater was asked to go back to review the benchmarks. Averaged raw score distributions for these analytic scores are provided in Table 3; score distributions were very similar across both training and evaluation sets for all three analytic scales. Scores for topic development were somewhat more disperse than for delivery or language use and showed a greater frequency of half-point scores, suggesting a somewhat lower degree of inter-rater agreement. Average Spearman correlations between human analytic scores ranged from .62 to .71 (Table 4), suggesting that scores for the dimensions are related, as might be expected for sub-constructs of a general language ability (speaking).

Machine analytic scores

Machine analytic scores were produced using language measures generated by the SpeechRater automated scoring engine. First, a large set of feature values, or measures of language phenomena, were generated by the SpeechRater engine as described in Higgins et al. (2011). Lasso regression was used on the training dataset to select a subset of SpeechRater features to be used in computing machine scores, following the approach described in Loukina, Zechner et al. (2015) and implemented in Madnani et al. (2017). This approach was used to optimise the process of feature selection and has been demonstrated to produce scoring models that more accurately predict human scores for TOEFL iBT speaking responses (Loukina, Zechner et al., 2015). We then manually adjusted the collections of SpeechRater features obtained from lasso regression to help ensure adequate coverage of each sub-construct. The rationale for this step was that the regression procedure selects features based on correlation with the training criterion (human score), without regard to what the features actually measure. As a result, this process is vulnerable to selecting features that do not measure construct-relevant phenomena, but rather are correlated with other phenomena that are related to the criterion measure. Specifically,

the initial machine-generated model for the sub-construct of language use contained features related to fluency and rhythm; these features were omitted from the final scoring model as described in more detail in the following section. Finally, features were entered as predictors in a linear regression model to produce an algorithm for computing machine analytic scores.

Scoring models were constructed for the sub-constructs of delivery and language use. No attempt was made to produce a scoring model for topic development given that (1) relatively few automated features addressing this sub-construct were available and (2) an intent of the study was to evaluate the use of human scores for this sub-construct. For delivery and language use, generic scoring models were used that were trained on responses to all six items, that is, the scoring models were not optimised for specific item types. This approach was taken because our previous experience in machine holistic scoring of TOEFL iBT responses found little consistent difference in performance of generic and item-specific scoring models.

Automated scoring model composition

Table 5 shows the SpeechRater features used in the scoring models for delivery and language use, as well as the standardised beta values from the linear regression models. The R-squared value for the delivery scoring models was .544 while for Language Use the R-squared was .517. In addition, relative beta values were calculated by dividing each standardised beta value by the total of standardised values. The relative beta values therefore add to 100% and are presented to indicate the relative impact of each feature on the prediction of the analytic score.

The distribution of relative weights across dimensions was generally in keeping with the definitions of the sub-constructs. That is to say, the combination of different linguistic measures that contributed to an analytic score was similar to the scoring criteria mentioned in the TOEFL iBT scoring rubric. A total of 12 features were selected for the delivery analytic score model and 6 features were selected for the language use score model. The features in the scoring model for delivery covered a range of phenomena and included speaking rate (wpsecutt), pausing (silpwd, longpfreq), repair (ipc), rhythm (stetimemean, stresyllmdev), chunking (wdpchk, wdpchenkmeandev), fillers and repetitions (dpsec, repfreq), and segmental pronunciation (L6, conftimeavg). This selection of features is in line with the wording of the scoring rubric regarding delivery, where, for example, excellent performance (score = 4) on the independent tasks is described as

Generally well-paced flow (fluid expression). Speech is clear. It may include minor lapses, or minor difficulties with pronunciation or intonation patterns, which do not affect overall intelligibility (Educational Testing Service, 2019).

Poor performance (score = 1) is described as

Consistent pronunciation, stress and intonation difficulties cause considerable listener effort; delivery is choppy, fragmented, or telegraphic; frequent pauses and hesitations (Educational Testing Service, 2019).

Table 5. SpeechRater features used to compute analytic scores.

a. Delivery				
Feature Name	Description	Phenomenon	Std B	Rel. B
stretimemean	Average distance between stressed syllables.	Pronunciation – Rhythm	0.174	15%
wpsecutt	Speaking rate in number of words per second.	Fluency – Speaking rate	0.173	15%
wdpchk	Average length of uninterrupted speech (chunks) in words.	Fluency – Chunking	0.151	13%
wdpchkmeandev	Mean absolute deviation of chunk length in words.	Fluency – Chunking	0.146	13%
conftimeavg	Mean automated speech recogniser confidence score; confidence score is a fit statistic to a NNS reference pronunciation model.	Pronunciation – Segmental	0.133	12%
repfreq	Number of repetitions, where one or more words are repeated exactly.	Repair – Repetitions	0.098	8%
silpwd	Number of silences (>0.15 seconds) per word.	Fluency – Pause frequency	0.075	6%
ipc	Number of interruption points (IP) per clause, where a repetition or repair is initiated.	Repair – Pausing	0.073	6%
stresyllmdev	Mean deviation of distances between stressed syllables.	Pronunciation – Rhythm	0.058	5%
L6	Normalised acoustic Model (AM) score, where pronunciation is compared to a NS reference model.	Pronunciation – Segmental	0.029	3%
longpfreq	Number of long silences (>0.5 sec) per word.	Fluency – Pause length/frequency	0.029	3%
dpsec	Number of disfluencies (fillers such as ‘uh’ and ‘uhm’) per second.	Fluency – Fillers	0.017	1%
b. Language Use				
Feature Name	Description	Phenomenon	Std B	Rel B
types	Number of word types used in the response.	Vocabulary diversity	0.393	35%
poscvamax	Comparison of part-of-speech bigrams in the response with responses receiving the maximum score.	Grammar – Accuracy & complexity	0.201	18%
logfreq	The average frequency of word types, where frequency is the number of occurrences in a reference corpus.	Vocabulary sophistication	0.165	15%
lmscore	Language model score; compares the response to a reference model of expected word sequences.	Grammar – Accuracy	0.129	11%
tpsec	Number of word types per second.	Vocabulary diversity & speaking rate	0.124	11%
cvamax	Comparison of words used in the response with responses receiving the maximum score.	Vocabulary – Diversity & sophistication	0.113	10%

The scoring criteria for delivery mention speaking pace, pausing, rhythm, chunking, segmental pronunciation, and intonation. Of these, only intonation pattern is not covered in some way in the automated scoring model evaluated in this study.

The scoring model for language use included four features related to vocabulary use (types, logfreq, tpsec and cvamax) and two features related to grammatical accuracy/complexity (poscvamax and lmscore). Although fewer features were chosen than for the scoring model for delivery, features related to both lexis and grammar are present as consistent with the wording of the rubric, as exemplified by the description of excellent performance for language use for the independent task:

The response demonstrates effective use of grammar and vocabulary. It exhibits a fairly high degree of automaticity with good control of basic and complex structures (as appropriate). Some minor (or systematic) errors are noticeable but do not obscure meaning (Educational Testing Service, 2019).

The rubric descriptors for language use also mention automaticity, which might be observed through speech rate or pausing. One of the vocabulary features included in the language use scoring model also captures speech rate (tpsec) and the original model produced by lasso regression contained three fluency-related features, accounting for 27% of the prediction and three rhythm features, which accounted for an additional 9%; these features were removed in the expert manual adjustment step to ensure that the language use model focused primarily on features clearly related to the descriptors for language use in the scoring rubric.

Evaluation of analytic and composite human-machine scores

The scoring algorithms developed using the training set were used to output predicted analytic scores for responses in the evaluation set. The performance of the algorithms in predicting human scores was then evaluated in terms of correlation and agreement (quadratic-weighted kappa) with human analytic scores. Agreement was examined at the level of both individual responses ($n = 600$) and summed scores of six items representing the entire speaking section for each test taker ($n = 100$). Inter-rater human-human agreement was similarly evaluated for human analytic and holistic scores for the same responses. Although all responses in the evaluation set were double-scored by human raters, evaluations of machine-human agreement were made using a randomly selected single human score to be consistent with the results for human-human performance, where a human score was also compared with a single human score from the other rater.

To create composite scores, analytic scores for the three TOEFL iBT speaking sub-constructs (delivery, language use, and topic development) for each response were averaged. Machine scores were rounded to whole numbers before averaging given that potential hybrid scores would be reported as whole numbers, although results were similar when using either rounded or unrounded values. Descriptive statistics and reliability (Cronbach's alpha) were then calculated. The following combinations were evaluated:

- Delivery (machine) + Language use (human) + Topic development (human)
- Delivery (human) + Language use (machine) + Topic development (human)
- Delivery (machine) + Language use (machine) + Topic development (human)
- Delivery (human) + Language use (human) + Topic development (human)

Results

Given that the scoring models were trained to approximate human analytic scores, the performance of the automated scoring models was first evaluated in terms of predicting human scores for the evaluation dataset (Table 6). Considering all responses individually, the correlations between machine and human scores were .65 for delivery and .61 for language use, compared to human-human inter-rater correlations of .70 and .73, respectively. Quadratic-weighted kappa values for machine-human agreement at the item level were .49 for delivery and .51 for language use, whilst human-human kappa values were .70 and .73. A drop in correlation of .1 or more from machine-human to human-human comparisons is not unusual in speech scoring (e.g. Zechner et al., 2015), although

Table 6. Performance of automated scoring models on the evaluation set (unseen data).

	Machine-Human		Human-Human		
	Delivery	Language Use	Delivery	Language Use	Topic Dev.
Item-level weighted kappa	.49	.51	.70	.73	.71
Item-level correlation	.65	.61	.70	.73	.70
Speaker-level correlation	.72	.75	.92	.92	.92

Note: Speaker-level correlation refers to the correlation at the level of total score on the speaking section. Consistency in section-level score is of interest because section scores are reported to score users and may be used to support decision making.

Loukina, Zechner et al. (2015) reported machine-human correlations for TOEFL iBT speaking data that were actually several hundredths higher than human-human correlations. However, the study by Loukina, Zechner, et al. used a much larger corpus of responses to train the scoring models (10,000 responses), the scoring models contained a larger number of predictors (from 25–75 features) and human-human agreement was lower (rater item-level correlation of .61 compared to roughly .70 in the current study).

The reliability of human and machine analytic scores were then examined (Table 7). Human analytic scores for the areas of delivery, language use, and topic development were all equally or more reliable than human holistic scores: a Cronbach's alpha of .88 was observed for human holistic scores and human scores for topic development, whereas alpha for human delivery and language use scores was higher at .92. Reliability of machine scores for both delivery and language use was lower than the comparable human scores, with a drop in alpha of .036 (4%) for delivery and .074 (8%) for language use. This drop is probably the result of a restriction in range among machine scores associated with the use of a linear regression model for predicting scores, as will be described in more detail in the discussion section. The standard deviation of the SpeechRater item scores was lower than that for human scores and fewer machine scores were observed at the extremes of the score distributions (Table 8).

Finally, speaking scores produced through four different combinations of automated and human analytic scores were evaluated. The reliability of speaker-level scores varied across different combinations of human and automated analytic scores, but in all cases, speaking section reliability was higher for composite scores than for human holistic scores (Table 9). This difference is perhaps not unexpected given that the composite scores incorporate more observations than the holistic scores. Among the various types of composite scores, those incorporating only human analytic scores showed the highest alpha values (.94) and reliability consistently decreased as more SpeechRater scores replaced human scores. Nonetheless, rounded composite scores incorporating

Table 7. Descriptive statistics for individual analytic scores for the evaluation dataset.

Scores	No. Items	Mean	SD	Cronbach's Alpha
Human holistic score	6	16.6	3.56	.88
Human analytic scores				
Delivery	6	16.4	4.27	.92
Language Use	6	16.1	4.42	.92
Topic development	6	16.2	4.33	.88
Machine analytic scores				
Delivery	6	17.3	2.83	.88
Language use	6	15.9	2.56	.85

Note: Reliability for machine scores was computed from values rounded to whole numbers.

Table 8. Score distributions of human and machine scores for individual responses from the evaluation dataset.

	<i>N</i>	Score			
		1	2	3	4
Holistic Score – Human	600	27 (5%)	174 (29%)	311 (52%)	88 (15%)
Language Use – Human	1199 ^a	105 (9%)	400 (33%)	468 (39%)	226 (19%)
Language Use – Machine	600	6 (1%)	228 (38%)	339 (57%)	27 (5%)
Delivery – Human	1200	90 (8%)	387 (32%)	483 (40%)	240 (20%)
Delivery – Machine	600	7 (1%)	124 (21%)	398 (66%)	71 (12%)
Topic Development – Human	1199 ^a	125 (10%)	360 (30%)	463 (39%)	251 (21%)

Note: Human analytic scores were obtained from two raters, thus *N* is double.

^aOne score of zero is not included in results.

Table 9. Reliability of composite scores (6 responses) for the evaluation dataset.

	No. Items	Mean	SD	Cronbach's Alpha
Holistic human score	6	16.6	3.56	.88
Composite scores				
Human analytic scores only	6	16.1	4.35	.94
Delivery (machine) + Language use (human) + Topic development (human)	6	16.5	3.81	.92
Delivery (human) + Language use (machine) + Topic development (human)	6	16.1	3.66	.92
Delivery (machine) + Language use (machine) + Topic development (human)	6	16.3	3.00	.90

Note: Composite scores are the average of three analytic scores; machine analytic scores were rounded to whole numbers before averaging.

SpeechRater analytic scores for both delivery and language use were slightly more reliable than human holistic scores (.90 vs. .88). There was a relatively high degree of exact agreement (82%) between all-human composite scores and hybrid scores incorporating a single machine analytic score (Table 10). This is not surprising given that human analytic scores made up two thirds of the measure. When a second analytic score was added, exact agreement decreased to 66%, similar to the level of exact agreement between the all-human hybrid score and the human holistic score (63%). This latter result suggests that the hybrid score generated from using two machine scores was similar in accuracy to the single human holistic score.

Table 10. Agreement with all-human composite scores (600 responses) for the evaluation dataset.

	Difference from all-human composite				
	–2	–1	0	1	2
Delivery (machine) + Language use (human) + Topic development (human)	0 (0%)	73 (12%)	492 (82%)	35 (6%)	0 (0%)
Delivery (human) + Language use (machine) + Topic development (human)	0 (0%)	54 (9%)	493 (82%)	53 (9%)	0 (0%)
Delivery (machine) + Language use (machine) + Topic development (human)	1 (0%)	111 (19%)	398 (66%)	90 (15%)	0 (0%)
Human holistic score	0 (0%)	88 (15%)	378 (63%)	130 (22%)	4 (1%)

Note: Negative values indicate the score was higher than the all-human composite. Composite scores were rounded to whole numbers for comparison.

Discussion

Our study explored the question of whether hybrid human-machine scores were more reliable than holistic scores produced by human raters. We found that machine analytic scores predicted human analytic scores with accuracy similar to that seen in other automated speech scoring contexts. Moreover, all of the human-machine composite scores we examined were slightly more reliable than human holistic scores, although the most reliable composite scores were constructed solely from human analytic scores. These findings suggest that in this instance there was relatively little benefit in using a hybrid approach, compared to holistic scoring.

The reliability of human analytic scores was higher or equal to the reliability of human holistic scores, although the differences were modest. One possible explanation for this finding is that the raters used in the current study, who were all scoring leaders, were more proficient in scoring than the raters who produced the operational holistic scores. Another possibility is that raters found it easier to make consistent decisions about separate dimensions either because there was a reduced cognitive load when scoring a single dimension or it was simpler to make decisions in cases where performance was uneven across dimensions. Uneven performance might have been common; Xi (2007) reported a difference in analytic speaking scores of 1 or more in roughly a third of cases for a sample of TOEFL iBT test takers. Similarly, Poonpon and Jamieson (2013), in a separate study of analytic scoring using a decision-tree rubric based on the TOEFL iBT scoring rubric, observed that roughly two-thirds of responses from a sample of TOEFL iBT test takers showed an uneven score profile across the sub-constructs of delivery, language use, and topic development. If human analytic scoring does indeed increase consistency, then this approach might be considered in place of human-machine hybrid scoring. However, human analytic scoring would have significant impacts on cost and operational demands, particularly if each response is scored by three different raters as done in the current study. We also note that the reliability of holistic scores was already .88, making it challenging to justify the additional cost and operational requirements that human analytic scoring would entail. Moreover, it is often difficult to demonstrate that sub-scores actually carry additional psychometric information beyond that contained in an overall score (Sinharay et al., 2011) and at the time the TOEFL iBT test was launched, Xi (2007) found that relatively little benefit was gained from human analytic scoring of speaking.

Although humans undoubtedly are more adept at judging higher-level language phenomena compared to automated scoring systems, human raters were not more reliable in scoring topic development compared to delivery or language use. Inter-rater agreement was similar across all three speaking sub-constructs and reliability of analytic scores for topic development (.88) was actually somewhat lower than for delivery and language use (.92). Evaluating topic development requires judgements to be made about higher-order language features such as organisation, where there may be many ways to successfully accomplish a communicative goal and the difference between stronger and weaker responses may be subtle. So, it is not surprising if reaching agreement on the quality of topic development would be relatively difficult for raters.

It was also hypothesised that the automated scoring engine might perform relatively well in scoring delivery or language use, but in actuality machine analytic scores were somewhat less reliable than human analytic scores. This lower reliability may be due in part to the fact that machine scores were less variable than human scores; essentially there was a tendency to over-predict scores near the mean. This restriction in range probably resulted from the use of regression to train the scoring model, where automated scores are equivalent to regression model predicted values. For regression, the variance of observed values is equal to the sum of the variance of predicted values plus the variance of residuals (Pedhazur, 1997, p. 23). That is, the variance of observed values (human scores) is equal to the variance of predicted values (machine scores) plus the residual variance (unaccounted variables plus error). If the residual variance is greater than zero, then the variance of human scores will always be greater than the variance of machine scores. Reduced variability in machine scores may have been a disadvantage in this instance, and in an operational context, might reduce item discrimination and therefore require that additional items are included to ensure a given degree of section-level reliability. However, reduced variability may actually be desirable in some cases in that it tends to minimise large prediction errors (Higgins et al., 2011).

The findings are also generally in line with the way that the academic speaking construct has been operationalised in the *TOEFL iBT* scoring rubric, where speaking ability is conceptualised as three distinct but closely related sub-constructs. Consistent with this view, inter-correlations between human analytic scores for different sub-constructs were moderate to high, ranging from .62 to .71. These correlations are also within the range of intercorrelations between listening, reading, speaking and writing test sections reported for the *TOEFL iBT* test (.54 to .76; Sawaki & Sinharay, 2013). They are lower than the correlations between *TOEFL iBT* speaking sub-constructs reported by Xi (2007), where observed correlations for different item types and sub-construct comparisons varied from .69 to .90 and averaged approximately .80. We note in the present study, raters scored only a single sub-construct for a given response, whilst in Xi's study, all sub-constructs were scored simultaneously, so it is possible that the higher correlations in her study may reflect some degree of the halo effect. Moreover, Xi's study investigated a pilot version of the *TOEFL iBT* test and so raters were relatively less experienced in applying the scoring criteria compared to the scoring leaders used in the current study, and Xi's raters may have found it more challenging to independently evaluate specific aspects of performance. Although we attempted to optimise the unique information captured by the different analytic scores through using highly experienced raters who scored sub-constructs independently, producing analytic scores that capture distinct information is inevitably a challenge given that language abilities are often moderately to highly correlated.

Interestingly, Xi (2007) noted that a majority of raters perceived 'some overlap' or 'much overlap' between the three sub-constructs, with delivery and language use being the most similar and delivery and topic development being the most distinct; our results show a similar pattern, where the inter-correlation between delivery and language use was highest (.71) and the correlation between delivery and topic development was lowest (.62). Also, as noted earlier, the initial version of the scoring model for language use included several fluency features and the scoring rubric description of this category essentially appeals to fluency features when using terms like 'automaticity'. However,

fluency features were purposefully excluded from the machine scoring model for language use, given that the machine measures fluency globally throughout the response, not just fluency or disfluency associated with producing challenging syntactic structures. This issue illustrates the challenge inherent in interpreting human scores and in designing automated systems when specific performance features are relevant to different speaking sub-constructs.

Limitations and conclusion

We found only a very small benefit of hybrid scoring in terms of score reliability, but this result should be interpreted in light of a number of limitations to the study, several of which might have reduced the performance of the automated scoring system. First, the corpora of test taker responses used to train and evaluate the automated scoring system was modest in size for this type of machine learning application. We used a corpus of 1,200 responses (600 responses for model training and 600 for model evaluation) which is smaller than would be advisable to build an operational scoring system, where corpus size may extend into thousands of responses (e.g. Chen et al., 2018), or potentially, hundreds of thousands of responses (e.g. Loukina & Yoon, 2020). This relatively small corpus reflects the cost of obtaining data for an experimental study where each response received six independent scores. However, this was an exploratory study and the sample provided a minimum of 50 observations per predictor, which we believe was adequate for the linear regression procedure used (Tabachnick & Fidell, 2013). Nonetheless, a larger dataset might have resulted in a scoring algorithm that better approximated human scores as well as more precisely evaluated scoring performance.

Language features were also intentionally removed from the automated scoring models to help ensure that automated scores measured phenomena relevant to the intended sub-construct and not other variables. This approach helped to ensure that machine analytic scores reflected the intended sub-constructs, but it might not have been optimal for accurate prediction of human analytic scores. However, given that the purpose of this study was to investigate the possibility of having machine and human scorers evaluate different sub-constructs, we felt it was important that the information going into a machine score reflected only the intended sub-construct. Similarly, we chose to use linear regression as the machine learning approach, rather than other approaches that could potentially produce better-performing scoring models, because linear regression has the advantage of transparency in the relative weighting of each linguistic feature used in computing the final score. Understanding construct coverage of each analytic score was a major concern, so we felt it was important to use a machine learning approach in which the relative importance of different language phenomena was as clear as possible. Further, the results were also limited by the technology available at the time of the study; additional linguistic measures continue to be added to the SpeechRater system (Chen et al., 2018), and there have been significant improvements to automated speech recognition technology (Qian et al., 2020), a key component in generating many SpeechRater features.

Finally, the use of human scores to train and evaluate automated scoring models may be questioned. An unsupervised machine learning approach or other means of synthetically creating a score might produce cleaner results, with scores that are more reliable or

more distinct across sub-constructs, and such an approach is something that might be investigated in the future. However, it is not a simple matter to create an unsupervised scoring model that is interpretable in terms of the construct as defined by test designers and documented in scoring rubrics or other materials, and so the validity of the resulting scores is a potential concern. We also note that in most cases, humans are the final arbiters of performance in the real world and use of human judgements to train scoring models helps to ensure that machine scores reflect human perceptions of ability. The use of human scores to train scoring models remains the standard in deployments of automated scoring of speaking performance (Chen et al., 2018; LaFlair & Settles, 2020; Pearson, 2019; Xu et al., 2020) and we believe that training the automated system on human analytic scores was a reasonable place to start.

To conclude, we believe that our study is a useful addition to the language testing literature in that it describes a first attempt within a speaking assessment context to evaluate the performance of a divergent contributory scoring approach where machine and human scores are combined to capture different aspects of the speaking construct. We found that composite human-machine scores were slightly more reliable than human holistic scores and in this instance, human raters were no less reliable in scoring foundational aspects of speaking (e.g. fluency, pronunciation, grammatical range and accuracy) compared to discourse-level phenomena.

For the context we investigated (the speaking section of the TOEFL iBT test), the modest improvement in score reliability we observed would probably not justify the extra expense and complication required for hybrid scoring based on analytic scores. However, the current study represents only a first step in the exploration of divergent contributory scoring. Decisions were made that prioritised the transparency of the automated measures over maximising scoring model performance and an effort focused on maximising performance might have produced more promising results. As technology continues to improve, automated scoring will probably feature more accurate measurement of a wider range of language phenomena, contributing to more accurate and reliable scoring and facilitating hybrid scoring approaches. There is considerable scope for additional evaluation of hybrid scoring using different machine learning approaches and increasingly powerful technologies.

The human element of hybrid scoring might also be further optimised. It is conceivable that if human raters focused their attention entirely on a single sub-construct, then scoring might be faster, or the consistency or accuracy of rater judgements might be improved. For example, if raters were focused on topic development, then it might be possible to employ detailed scoring criteria or provide focused training to support more consistent scoring judgements. The feasibility of divergent contributory scoring relies on the quality of both human and machine scores, and improvements to human scoring of targeted aspects of performance aspects is an important subject for future research.

Acknowledgments

We thank Anastassia Loukina for her assistance in producing the automated scoring models.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Ethics


This research was deemed exempt from review by the ETS Committee for Prior Review of Research. De-identified data was obtained for test takers who, at the time of test registration, had indicated that their data could be used for research purposes. Human raters used in the study were ETS employees working within the scope of their employment.

Funding

This work was supported by Educational Testing Service.

ORCID

Larry Davis  <http://orcid.org/0000-0002-1656-1123>

Spiros Papageorgiou  <http://orcid.org/0000-0002-7940-3472>

References

- Bridgeman, B. (2013). Human ratings and automated essay evaluation. In M. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 221–232). Routledge.
- Chen, L., Zechner, K., Yoon, S. Y., Evanini, K., Wang, X., Loukina, A., Tao, J., Davis, L., Lee, C. M., Ma, M., Mundkowsky, R., Lu, C., Leong, C. W., & Gyawali, B. (2018). *Automated scoring of nonnative speech using the SpeechRaterSM v. 5.0 engine* (Research Report No. RR-18-10). Educational Testing Service. <https://doi.org/10.1002/ets2.12198>
- Educational Testing Service. (2017). *How the test is scored*. https://www.ets.org/gre/revised_general/scores/how/
- Educational Testing Service. (2019). *TOEFL iBT speaking section scoring guide*. https://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf
- Educational Testing Service. (2020a). *TOEFL Research Insight Series Volume 2: TOEFL Research*. https://www.ets.org/s/toefl/pdf/toefl_ibt_insight_s1v2.pdf
- Educational Testing Service. (2020b). *TOEFL Research Insight Series Volume 3: Reliability and comparability of TOEFL iBT scores*. https://www.ets.org/s/toefl/pdf/toefl_ibt_research_s1v3.pdf
- Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing*, 27(3), 317–334. <https://doi.org/10.1177/0265532210363144>
- Harsch, C., & Martin, G. (2013). Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice*, 20(3), 281–307. <https://doi.org/10.1080/0969594X.2012.742422>
- 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(2), 282–306. <https://doi.org/10.1016/j.csl.2010.06.001>
- Isaacs, T. (2018a). Shifting sands in second language pronunciation teaching and assessment research and practice. *Language Assessment Quarterly*, 15(3), 273–293. <https://doi.org/10.1080/15434303.2018.1472264>

- Isaacs, T. (2018b). Fully automated speaking assessment: Changes to proficiency testing and the role of pronunciation. In O. Kang, R. I. Thomson, & J. Murphy (Eds.), *The Routledge Handbook of English Pronunciation* (pp. 570–584). Routledge.
- LaFlair, G. T., & Settles, B. (2020). *Duolingo English Test: Technical manual*. Duolingo. <https://duolingo-papers.s3.amazonaws.com/other/det-technical-manual-current.pdf>
- Loukina, A., Lopez, M., Evanini, K., Suendermann-Oeft, D., & Zechner, K. (2015). Expert and crowdsourced annotation of pronunciation errors for automatic scoring systems. In *Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association* (pp. 2809–2813). International Speech Communication Association. <http://dx.doi.org/10.21437/Interspeech.2015-591>
- Loukina, A., & Yoon, S. Y. (2020). Scoring and filtering models for automated speech scoring. In K. Zechner & K. Evanini (Eds.), *Automated speaking assessment: Using language technologies to score spontaneous speech* (pp. 192–204). Routledge.
- Loukina, A., Zechner, K., Chen, L., & Heilman, M. (2015). Feature selection for automated speech scoring. In J. Tetreault, J. Burstein, & C. Leacock (Eds.), *Proceedings of the Tenth workshop on innovative use of NLP for building educational applications* (pp. 12–19). Association for Computational Linguistics. <http://dx.doi.org/10.3115/v1/W15-06>
- Luo, D., Gu, W., Luo, R., & Wang, L. (2016). Investigation of the effects of automatic scoring technology on human raters' performances in L2 speech proficiency assessment. In *10th International Symposium on Chinese Spoken Language Processing* (pp. 1–5). <https://doi.org/10.1109/ISCSLP.2016.7918378>
- Madnani, N., Loukina, A., von Davier, A., Burstein, J., & Cahill, A. (2017). Building better open-source tools to support fairness in automated scoring. In D. Hovy, S. Spruit, M. Mitchell, E. Bender, M. Strube, & H. Wallach (Eds.), *Proceedings of the First Workshop on Ethics in Natural Language Processing* (pp. 41–52). Association for Computational Linguistics. <https://aclanthology.org/W17-1605.pdf>
- Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015). *The association between TOEFL iBT® test scores and the Common European Framework of Reference (CEFR) levels* (Research Memorandum No. RM-15–06). Educational Testing Service. <https://www.ets.org/Media/Research/pdf/RM-15-06.pdf>
- Pearson. (2019). *Pearson Test of English Academic: Automated scoring*. https://assets.ctfassets.net/yqwtwibiobs4/018RxttvPWsMkkGIQJ5Gg3/6f410437ceb2c6f2762fbcdfa8a28e8c/2021_PTEA_White_Paper_Institutions_Automated_Scoring_White_Paper-May-2018.pdf
- Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed). Harcourt Brace.
- Poonpon, K., & Jamieson, J. (2013). *Developing analytic rating guides for TOEFL iBT's integrated speaking tasks* (Research Report No. RR-13–13). Educational Testing Service. <https://doi.org/10.1002/j.2333-8504.2013.tb02320.x>
- Qian, Y., Lange, P., & Evanini, K. (2020). Summary and outlook on automated speech scoring. In K. Zechner & K. Evanini (Eds.), *Automated speaking assessment: Using language technologies to score spontaneous speech* (pp. 61–74). Routledge.
- Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012). *Evaluation of the e-rater® scoring engine for the TOEFL® independent and integrated prompts* (Research Report No. RR-12–06). Educational Testing Service. <https://doi.org/10.1002/j.2333-8504.2012.tb02288.x>
- Sawaki, Y., & Sinharay, S. (2013). *Investigating the value of section scores for the TOEFL iBT test* (Research Report No. RR-13–35). Educational Testing Service. <https://doi.org/10.1002/j.2333-8504.2013.tb02342.x>
- Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice*, 30(3), 29–40. <https://doi.org/10.1111/j.1745-3992.2011.00208.x>
- Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson.

- Williamson, D., Xi, X., & Breyer, 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. (2007). Evaluating analytic scoring for the TOEFL Academic Speaking Test (TAST) for operational use. *Language Testing*, 24(2), 251–286. <https://doi.org/10.1177/0265532207076365>
- Xi, X., Higgins, D., Zechner, K., & Williamson, D. (2008). *Automated scoring of spontaneous speech using SpeechRaterSM v. 1.0* (Research Report No. RR-08–62). Educational Testing Service. <https://doi.org/10.1002/j.2333-8504.2008.tb02148.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(3), 371–394. <https://doi.org/10.1177/0265532211425673>
- Xu, J., Brenchley, M., Jones, E., Pinnington, A., Benjamin, T., Knill, K., Seal-Coon, G., Robinson, M., & Geranpayeh, A. (2020). *Linguaskill: Building a validity argument for the speaking test*. Cambridge Assessment English. <https://www.cambridgeenglish.org/Images/589637-linguaskill-building-a-validity-argument-for-the-speaking-test.pdf>
- Yoon, S. Y., & Zechner, K. (2017). Combining human and automated scores for the improved assessment of non-native speech. *Speech Communication*, 93, 43–52. <https://doi.org/10.1016/j.specom.2017.08.001>
- Zechner, K. (2020). Summary and outlook on automated speech scoring. In K. Zechner & K. Evanini (Eds.), *Automated speaking assessment: Using language technologies to score spontaneous speech* (pp. 192–204). Routledge.
- Zechner, K., Chen, L., Davis, L., Evanini, K., Lee, C. M., Leong, C. W., Wang, X., & Yoon, S. Y. (2015). *Automated scoring of speaking tasks in the Test of English-for-Teaching (TEFT[™])* (Research Report No. RR-15–31). Educational Testing Service. <https://doi.org/10.1002/ets2.12080>