Listening. Learning. Leading.®

Research Report
ETS RR–15-31

Automated Scoring of Speaking Tasks in the Test of English-for-Teaching (*TEFT*™)

Klaus Zechner

Lei Chen

Larry Davis

Keelan Evanini

Chong Min Lee

Chee Wee Leong

Xinhao Wang

Su-Youn Yoon

August 2015

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.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Automated Scoring of Speaking Tasks in the Test of English-for-Teaching (*TEFT*™)

Klaus Zechner, Lei Chen, Larry Davis, Keelan Evanini, Chong Min Lee, Chee Wee Leong, Xinhao Wang, & Su-Youn Yoon

Educational Testing Service, Princeton, NJ

This research report presents a summary of research and development efforts devoted to creating scoring models for automatically scoring spoken item responses of a pilot administration of the Test of English-for-Teaching (*TEFT*™) within the *ELTeach*™ framework. The test consists of items for all four language modalities: reading, listening, writing, and speaking. This report only addresses the speaking items, which elicit responses ranging from highly predictable to semipredictable speech from nonnative English teachers or teacher candidates. We describe the components of the system for automated scoring, comprising an automatic speech recognition (ASR) system, a set of filtering models to flag nonscorable responses, linguistic measures relating to the various construct subdimensions, and multiple linear regression scoring models for each item type. Our system is set up to simulate a hybrid system whereby responses flagged as potentially nonscorable by any component of the filtering model are routed to a human rater, and all other responses are scored automatically by our system.

Keywords Automated speech scoring; language assessment; automated scoring of non-native speech

doi:10.1002/ets2.12080

As English has become increasingly important as a language of international business, trade, science, and communication, efforts to promote teaching English as a foreign language (EFL) have increased in many non-English-speaking countries worldwide in recent years. In addition, the prevailing trend in English pedagogy has been to promote the use of spoken English in the classroom, as opposed to the respective native languages of the EFL learners. However, due to the high demand for EFL teachers in many countries, the training of these teachers has not always caught up with these high expectations, so there is a need for both governmental and private institutions involved in the employment and training of EFL teachers to develop, improve, and assess their competence in the English language as well as in English pedagogy.

This research report describes the development and evaluation of scoring models for automated scoring of 21 speaking items contained in a pilot form of the Test of English-for-Teaching (*TEFT*™) assessment, a part of the *ELTeach*™ program. ELTeach is an online professional development program, consisting of two courses: English-for-Teaching and Professional Knowledge for ELT. Each course includes a coordinated assessment leading to a score report and certificate of completion for individual teachers. The learning materials and assessment are offered as an integrated program; neither component is independent. Developed with reference to international and national teaching standards and drawing on the resources of various national English-language curricula and teaching materials, the program is designed to ensure that teachers have the functional classroom English language and the professional knowledge to support the implementation of the English-language curricula they are expected to teach. This alignment bounds the classroom language known as English-for-teaching and defines the domain and the concepts known as professional knowledge for English-language teaching (ELT), which are presented in the learning materials and tested in the assessments.

The language taught in the English-for-Teaching course is a bounded set of functional words, phrases, and language skills teachers use to carry out essential classroom activities in English. The language in the course is organized into three areas: managing the classroom, understanding and communicating lesson content, and providing feedback. The language is anchored in the regular, predictable tasks that teachers perform in the course of teaching English rather than in a general proficiency framework (see Freeman, Katz, Garcia Gomez, & Burns, 2015, for an in-depth discussion of the

Corresponding author: K. Zechner, E-mail: kzechner@ets.org

concept of English-for-teaching). The specific language exemplars were gleaned, with direction from a global panel of experts from 10 countries, from national teaching materials, curricula, and local classroom data. The learning materials are designed to be self-access, thus allowing teachers to focus on areas of need and determine individually what they practice and how they progress. The TEFT assessment has been developed based on the same framework as the learning materials; it is designed to measure how test takers perform on typical classroom tasks as represented in the course materials.

The TEFT assessment consists of items related to the four language modalities: (a) multiple-choice (MC) reading items, (b) MC listening items, (c) constructed-response (CR) writing items, and (d) CR speaking items. The TEFT is designed to assess the English-language skills that are needed to complete basic tasks in the English-language teaching classroom, supported by English-language instructional materials. In this research report, we are concerned mainly with 21 speaking items in a pilot form of the TEFT.

Several significant challenges needed to be addressed during the course of building the automated speech scoring system, including, but not limited to the following:

- The 21 speaking items belong to eight different task types with different characteristics; therefore, we selected features[1] and built scoring models for each task type separately.
- The test takers speak a variety of native languages and thus have very different accents in their spoken English. Furthermore, the test takers also exhibit a range of speaking proficiency levels, which contributes to the diversity of their spoken responses.
- Since content accuracy is very important for the types of items contained in the test, even small error rates by the automatic speech recognition (ASR) system can lead to a noticeable impact on feature performance.
- About 9% of spoken responses are considered nonscorable by human raters, because, for example, they do not contain any spoken content, are off topic, or exhibit a level of noise that is too high.

The score distribution is highly skewed toward the higher end of the scale, which makes the building of scoring models more challenging than for more normally distributed score data. A highly skewed distribution is, however, what is expected for this test, given the purpose of the test (achievement rather than proficiency evaluation) and that test takers are examined on known materials.

Related Work

Automated speech processing and scoring technology has been applied to a variety of domains over the course of the past two decades, including evaluation and tutoring of children's literacy skills (Mostow, Roth, Hauptmann, & Kane, 1994), preparation for high-stakes English proficiency tests for institutions of higher education (*TOEFL Practice Online TPO*™; Zechner, Higgins, Xi, & Williamson, 2009), and evaluation of English skills of foreign-based call center agents (Chandel et al., 2007), to name a few (for a comprehensive overview, see Eskenazi, 2009).

These applications generally elicit restricted speech from the participants, and the most common item type by far is the read aloud, in which the speaker reads a sentence or collection of sentences out loud. Owing to the constrained nature of this task, it is possible to develop ASR systems that are relatively accurate, even with heavily accented nonnative speech. Several types of features related to a nonnative speaker's ability to accurately produce English sounds and speech patterns effectively have been extracted from these types of responses. Some of the best performing of these types of features include:

- pronunciation features, such as a phone's spectral match to native speaker acoustic models and a phone's duration compared to native speaker models;
- fluency features, such as the rate of speech, mean pause length, and number of disfluencies;
- prosody features, such as pitch and intensity contour; and
- reading accuracy features, such as the percentage of words incorrectly read.

In addition to the large majority of applications that elicit restricted speech, a small number of applications have also investigated automated scoring of nonnative spontaneous speech in order to more fully evaluate a speaker's communicative competence (e.g., Cucchiarini, Strik, & Boves, 2002; Zechner et al., 2009). In these systems, the same types of pronunciation, fluency, and prosody features can be extracted; in addition, features related to additional aspects of a

speaker's proficiency in the nonnative language can be extracted, such as vocabulary usage, syntactic complexity, and topical content.

As described previously, the domain for the automated speaking assessment investigated in this report is teachers of EFL around the world. Based on the fact that many of the item types are designed to assess the test taker's ability to productively use English constructions and linguistic units that commonly occur in English teaching environments, several of the item types elicit semirestricted speech. These types of responses fall somewhere between the heavily restricted speech elicited by a read aloud task and unconstrained spontaneous speech. In these semirestricted responses, the test taker may be provided with a set of lexical items that should be used to form a sentence; in addition, the test taker is often asked to make the sentence conform to a given grammatical template. Thus, the responses provided for a given question of this type by multiple different speakers will often overlap with each other; however, it is not possible to specify a complete list of all possible responses. These types of questions have only infrequently been examined in the context of automated speech scoring, which means that new types of features will need to be explored in order to evaluate them fully. Some related item types that have been explored previously include the sentence build and short item types described by Bernstein, Van Moere, and Cheng (2010); however, those item types typically elicited a much narrower range of responses than the semirestricted ones discussed in this report.

In order to address these semirestricted speaking test items in the TEFT assessment, Zechner and Wang (2013) developed a set of content accuracy features based on regular expressions, string matching, and n-gram statistics. As for delivery and language use, for the most part, previously developed features of *SpeechRater*^{SM} were used for TEFT scoring (Chen & Yoon, 2012; Jeon & Yoon, 2012; Yoon & Bhat, 2012; Yoon, Bhat, & Zechner, 2012; Yoon, Evanini, & Zechner, 2011; Zechner et al., 2009).

Overview of the Report

In the next section, we describe the TEFT assessment in detail and then provide a description of the TEFT Speaking items and a split-half reliability study. Next, we describe the TEFT assessment pilot data and the components of the automated scoring system: the ASR system, features used for scoring, linear regression scoring models, and the filtering models for identifying nonscorable responses. We then report results for the preliminary and final systems and follow with a discussion and conclusions. Several appendices contain additional materials related to this research.

The TEFT Pilot Assessment

Two forms of the TEFT pilot assessment were administered in the fall of 2012. Listening and reading items are MC, whereas speaking and writing items elicit CR.

The TEFT assessment is divided into two sections: A and B. Section A contains three parts: reading, writing, and listening. Section B is divided into four simulated lessons based on a unified set of tasks addressing a single teaching goal. Each of these lessons contains listening, speaking, and writing items, and the pilot form contains five unique lessons, two of which were split between different sets of test takers.

Each lesson contains seven speaking items; thus, each test taker has 28 speaking items total in a TEFT pilot form. Since only 21 items were administered to all test takers and we wanted to maximize the data available for system development, we are using only the 21 items from the lessons taken by all pilot candidates in this report for system development and evaluation. Each response was scored by two human raters (H1 and H2) using the following scores: 0, 1, 2, 3, and TD (technical difficulty). For the pilot, the final item score was computed as either

- the rounded average of H1 and H2 if the absolute value of the difference between H1 and H2 was less than 2 and neither H1 nor H2 had a score of 0 or TD; or
- an adjudicated score provided by a third rater (a scoring leader) if the absolute value of the difference between H1 and H2 was greater than 1, or if either H1 or H2 was 0 or TD.[2] In the event that the adjudicator's score was either 0 or TD, a code indicating the reason for the 0 or TD score was provided by the adjudicator.

The scoring rubrics for the speaking items addressed three main constructs: delivery, language use, and content. The rubrics are included in Appendix A and descriptions of the 0 and TD subscores are presented in Appendix B.

Table 1 Types of TEFT Speaking Items Investigated for Automated Scoring

Item type	Description
Group 1: Low-entropy items	
MC	The test taker is presented with a classroom scenario and then is asked to read one of several choices aloud that best fit the scenario.
RA	The test taker reads aloud a set of sentences presented on the screen.
RP	The test taker listens to a short utterance and then repeats it aloud.
Group 2: Medium-entropy items	
IS	The test taker is given a sentence fragment and completes the sentence according to the instructions.
KW	The test taker uses the key words provided to formulate an utterance according to instructions.
CH	The test taker uses an example from a chart and then formulates a similar sentence using a given grammatical pattern.
KC	The test taker produces an utterance using given keywords and information in a chart.
VI	The test taker is asked to produce an utterance using information presented in two visuals.

Note. MC = multiple-choice read aloud; RA = read aloud; RP = repeat aloud; IS = incomplete sentence; KW = key words; CH = chart; KC = keyword chart; VI = two visuals.

Table 2 Descriptive Statistics for the Half Test and Total Test Scores and Split-Half Reliability Information for Speaking Scores

	Half test 1		Half test 2		All items		*r* Half 1 with	Number of items			Reliability	
Language skill	Mean	*SD*	Mean	*SD*	Mean	*SD*	Half 2	Half 1	Half 2	Total test	SBP	*SEM*
Writing	24.4	5.6	22.7	5.0	47.0	10.1	.83	11	10	21	.90	3.12
Reading	14.8	2.1	13.0	2.4	27.9	4.2	.75	17	16	33	.85	1.60
Listening	14.7	1.4	13.1	1.3	27.7	2.4	.62	16	14	30	.75	1.20
Speaking	21.4	3.9	30.6	4.3	52.0	7.8	.78	9	12	21	.86	2.91
Speaking (system)	21.6	4.0	30.2	4.7	51.8	8.4	.84	9	12	21	.90	2.65
Written language	39.2	7.2	35.7	6.8	74.9	13.5	.87	28	26	54	.93	3.64
Oral language	36.0	4.7	43.7	5.2	79.7	9.5	.82	25	26	51	.90	2.97
Oral language (system)	36.2	4.8	43.3	5.4	79.5	9.8	.85	25	26	51	.92	2.82
Total	75.2	11.1	79.4	11.1	154.6	21.7	.91	53	52	105	.95	4.82
Total (system)	75.4	11.0	79.0	11.1	154.4	21.5	.90	53	52	105	.95	4.82

Note. SBP = Spearman-Brown prophecy.

Speaking Item Types on Pilot Form 1

Table 1 provides a brief description of the eight CR speaking item types; the corresponding scoring rubrics are available in Appendix A (Tables A1 and A2). For scoring purposes, items were divided into two groups: Group 1, in which the test taker was asked to read a short text aloud or repeat the contents of a short audio stimulus, and Group 2, in which test takers were asked to use specific words, a grammatical structure, or visuals to construct a response. Accordingly, the content of test-taker responses was tightly controlled for Group 1 items (i.e., low-entropy or highly predictable responses were expected), whereas for Group 2 items the test taker had somewhat more latitude in making a response (i.e., medium-entropy or semipredictable responses were expected). As shown in Tables A1 and A2, two different scoring rubrics were used for these two different types of items.

Table 2 shows descriptive statistics and split-half reliability information for the TEFT assessment scored speaking items, where all CR scores are from the first human rater. We distributed the CR item types between the first and second half tests for the writing and speaking half tests, with intact lessons placed in the same half test for all skills. The fact that lessons were kept intact meant that in some cases one half of a test section contained more items than the other half (see Table 2) because each lesson contained a slightly different number of items for each modality (reading, listening, etc.). The data shaded in gray provide similar information, where the first human rater is replaced by the automated scoring system for speaking. Comparison of human and automated scores shows little degradation in score reliability. Further, standardized mean score differences between human and automated scores (all items) are 0.09 for speaking, 0.09 for oral language (speaking and listening), and 0.06 for the total scores.

Table 3 Observed Score Intercorrelation Matrix With and Without Correction for Attenuation

Language skill	Writing	Reading	Listening	Speaking	Speaking (system)
Writing	**.90**	*.83*	*.76*	*.79*	*.68*
Reading	.73	**.85**	*.83*	*.78*	*.68*
Listening	.63	.67	**.75**	*.77*	*.60*
Speaking	.70	.67	.62	**.86**	*.82*
Speaking (system)	.59	.56	.49	.72	**.90**

Note. Correlations above the diagonal are corrected for attenuation (in italic font). Split-half reliabilities calculated using the Spearman-Brown prophecy formula are provided in the diagonals (bold font).

Table 4 Size of the Partitions Used for Training and Evaluating SpeechRater for Form 1

Partition	Test takers	Speaking responses
ASR training	1,658	27,604
ASR development	25	700
ASR evaluation	25	700
Scoring model training	300	8,400
Scoring model evaluation	300	8,400

Note. ASR = automatic speech recognition.

Table 3 shows the intercorrelations among the different test scores. Correlations below the diagonal are observed score correlations, and those above the diagonal are corrected for attenuation (in italic font). Split-half reliabilities calculated using the Spearman-Brown prophecy formula are provided in the diagonals (bold font).

Development of the Automated Speech Scoring System

Data Partitions

The data from the TEFT pilot administration was partitioned into the following five sets for the purpose of developing an automated scoring capability for responses to TEFT pilot Form 1 items:

- ASR training: responses used to train the speech recognizer's language model (responses from Pilot Form 2 for 14 additional items that overlapped with Form 1 were also included in this set)
- ASRdevelopment[3]: responses used to tune the parameters of the speech recognizer
- ASRevaluation: responses used to obtain the final performance statistics for the speech recognizer
- Scoing model training: responses used to train the scoring model
- Scoing model evaluation: responses used to evaluate the scoring model

The number of test takers and speaking responses contained in each of these partitions is presented in Table 4. All responses were further transcribed verbatim by human transcribers in order to train and evaluate the ASR system and to build features that depend on sets of human responses.

Participants from 10 different countries took part in the TEFT pilot administration. Table 5 presents the number of participants from each of these 10 countries that took part in the ASR training, scoring model training, and scoring model evaluation partitions.

Table 6 presents the number of responses in the ASR training, scoring model evaluation, and scoring partitions that received each of the five human scores (0, 1, 2, 3, and TD) based on the H1 score. As the table shows, Score Point 3 accounts for approximately 55% of the responses in all three partitions. Also, Score Point 1 is relatively infrequent, accounting for approximately 10% of the responses in each partition.

Automated System Development

The SpeechRater system used for automated assessment of the TEFT responses consists of the following four major components: (a) an ASR system that converts the test taker's response into a sequence of words; (b) a component that computes

Table 5 Number of Test Takers From Each Country

	ASR training		Scoring model training		Scoring model evaluation	
Country	Frequency	Percent	Frequency	Percent	Frequency	Percent
Brazil	40	2.4	11	3.7	8	2.7
Chile	32	1.9	3	1.0	3	1.0
China	910	54.9	116	38.7	120	40.0
Dominican Republic	97	5.9	14	4.7	11	3.7
Italy	190	11.5	58	19.3	87	29.0
Korea	53	3.2	14	4.7	7	2.3
Mexico	209	12.6	33	11.0	23	7.7
Mongolia	7	0.4	5	1.7	2	0.7
Peru	62	3.7	11	3.7	10	3.3
Vietnam	58	3.5	35	11.7	29	9.7
Total	1,658	100.0	300	100.0	300	100.0

Note. ASR = automatic speech recognition.

Table 6 Human Score Distributions for the Speaking Responses in Three Partitions

	ASR training		Scoring model training		Scoring model evaluation	
Score	Frequency	Percent	Frequency	Percent	Frequency	Percent
0	469	1.7	95	1.1	142	1.7
1	2,339	8.5	858	10.2	781	9.3
2	7,235	26.2	2,376	28.3	2,237	26.6
3	15,142	54.9	4,723	56.2	4,671	55.6
TD	2,419	8.8	348	4.1	569	6.8
Total	27,604	100.0	8,400	100.0	8,400	100.0

Note. ASR = automatic speech recognition; TD = technical difficulty.

Table 7 Word Error Rate (WER) of the Speech Recognition System for Low-Entropy Items

Item type	RA	MC	RP	Microaverage
WER	11.4	17.1	21.8	12.2

Note. RA = read aloud; MC = multiple-choice read aloud; RP = repeat aloud. Microaverage is the mean of all WERs of all responses.

approximately 100 different features based on the ASR output and the speech signal (these features are related to different aspects of the speaking construct, but mainly cover the following three areas: delivery, language use, and content accuracy); (c) linear regression scoring models trained on the 8,400 responses in the scoring model training partition (one model for each of the eight task types); and (d) filtering models that flag responses that cannot or should not be scored automatically (these responses are routed to human raters in operation).

Speech Recognizer

A baseline ASR system was trained using a corpus of about 800 hours of transcribed *TOEFL iBT*® spoken responses.[4] An item-independent TEFT recognizer (i.e., containing a single language model and acoustic model for all items) was then built using responses from all of the 35 items (21 scored and 14 additional) in order to adapt its language model (the interpolation weight was 0.9 for the TEFT corpus used here). This ASR system was then used to generate word hypotheses. Tables 7 and 8 present the performance of the recognition system (in terms of word error rate [WER]) on the 8,400 responses in the scoring model evaluation partition; the performance is presented separately for each item type, although a single generic recognizer was used across all items. The overall WER of the ASR system on this data set is 15.6%.

The WER for low-entropy items ranges from 11.4 to 21.8 (with a microaverage of 12.2), and the WER for medium-entropy items ranges from 26.3 to 41.4 (with a microaverage of 27.8). The WER for read-aloud (RA) items of 11.4 is

Table 8 Word Error Rate (WER) of the Speech Recognition System for Medium-Entropy Items

Item type	CH	KW	IS	VI	KC	Microaverage
WER	26.3	28.7	41.4	30.4	28.8	27.8

Note. CH = chart; KW = key words; IS = incomplete sentence; VI = two visuals; KC = keyword chart. Microaverage is the mean of all WERs of all responses.

comparable to error rates we have observed in other assessments such as *TOEIC®*. The WER for medium-entropy items (27.8) is very similar to typical WERs we have observed for TOEFL iBT data that are high entropy in nature. This indicates that ASR for TEFT items is somewhat more challenging than for other assessments, given its low- to medium-entropy characteristics. We conjecture that part of the reason is the relatively high number of nonscorable responses (9% were rated as 0 or TD) in TEFT, which causes problems for the ASR system.

Feature Generation

SpeechRater generated a total of 106 features covering the three major construct areas of TEFT as specified in the rubrics (see Appendix A): delivery, language use, and content accuracy (language use is only relevant for medium-entropy items and content accuracy is largely reduced to string matching for low-entropy items). The delivery group is composed of features that assess a nonnative speaker's fluency, pronunciation, and prosody. The language use group includes features that assess the speaker's vocabulary diversity and grammatical accuracy. Finally, the content accuracy group measures the correctness of the content contained in the speaker's response, with the features based on string matching, n-grams, edit distance, and regular expressions (Zechner & Wang, 2013). While most of the features used in the TEFT scoring models were originally developed to measure spontaneous speech (in particular, delivery and language use features), the content accuracy features for medium-entropy item types were specifically designed for the TEFT assessment definitions, and detailed explanations of each feature used in the scoring models are presented in Appendix C. More detail regarding the content features can be found in Zechner and Wang (2013).

The language use and content accuracy features were generated for each individual response separately, whereas the delivery features were generated based on a concatenated representation of the responses to the seven items contained in a single lesson in Section B of the assessment. This concatenation-based approach was implemented because the average duration of the spoken responses in the TEFT assessment is shorter than in other speaking assessments, such as the TOEFL iBT, and it is difficult to extract reliable delivery features from such short responses. Therefore, under the assumption that a speaker's delivery skill remains relatively constant across different utterances,[5] the seven responses in a lesson were concatenated into one long response. Delivery features were then generated from this concatenated response; thus, all seven responses in a single lesson received identical values for the delivery features.[6]

Feature Selection and Scoring Model Building

The construction of the scoring models started with the process of dividing the SpeechRater features into constructs that reflected the scoring criteria contained in the scoring rubrics for the speaking items, namely delivery, language use, and content. The delivery features were further divided into the subconstructs of fluency, segmental pronunciation, and prosody (rhythm and intonation), whereas the language use features were divided into the subconstructs of grammar and vocabulary. Feature selection was done using the scoring model training data set only.

Initially, features were selected from all subconstructs on the basis of zero-order correlations of raw (untransformed) features with human scores. (For low-entropy items, we did not select any grammar or vocabulary features because evidence of ability to use these aspects of language is not elicited by these items.) The selected features were transformed when appropriate to obtain normality and help ensure normal distribution of residuals as required by the regression model, as well as to better evaluate correlations with human scores. A set of features was chosen for each item type based on the following four criteria: (a) coverage of the subconstructs, (b) high correlation with human scores, (c) distributions that best approximated normality, and (d) no correlation greater than .90 with any other feature in the set. Ultimately, one feature set was chosen for each item type based on these four criteria; the feature sets for each item type are given in Appendix D.

Table 9 Features for the Filtering Model Baseline

Feature name	Description
Numwds	Number of words in the word hypothesis of spoken responses
Segdur	Total duration of utterance without disfluencies or pauses
Powmean	Mean energy of the entire audio file
Powmax	Global maximum energy of the entire audio file
Powvar	Variance of energy distribution
Nrprob	MFCC-based audio quality score

Note. MFCC = mel-frequency cepstral coefficient.

A variety of multiple regression scoring models was constructed, with SpeechRater features as predictors and human scores (the average of H1 and H2) as the dependent variable; features with negative zero-order correlation with human scores were multiplied by −1 so that negative beta weights would indicate disagreement in sign between the weight and the zero-order correlation. Regression models were computed for pooled items within each item type as well as separately for each individual item. Model performance was evaluated in terms of the ability of each model to predict single-rater (H1) scores in the scoring model evaluation partition. The following regression models were investigated:

- An empirical weight model with deleted weights (labeled empWtDrop). After obtaining an empirical weight multiple regression model, any feature whose beta weight was negative was assigned a weight of zero.
- An empirical weight model where negative beta weights were converted to small positive values (labeled empWtBalanced). After obtaining an empirical weight multiple regression model, positive standardized beta weights were multiplied by 95% (0.95). Then, 5% of the sum of all positive weights was evenly distributed among the features whose weights were negative, resulting in small positive values for these features.

Ultimately, it was found that the empWtBalanced model generally gave relatively good results; additionally, this model retains features for all subconstructs. While the empWtDrop model showed similar performance, the removal of features resulted in a loss of construct coverage. In addition, little difference was seen in model performance when regression parameters were calculated for individual items compared to when a common regression model was used for all items within an item type. Given that the use of the same regression model/parameters/features for all items within an item type simplifies implementation and produced similar results across individual items, the SpeechRater scores reported in this research report were generated using the empWtBalanced model with the same parameters used for all items within an item type.

Filtering Model Building

Filtering Model Baseline

Approximately 7% of responses in the TEFT pilot data had a TD that resulted in the human raters being unable to provide a valid score (e.g., equipment errors, loud background noise), and 1% to 2% of responses had a human score of 0 (e.g., empty responses, non-English responses, off-topic responses). The suboptimal characteristics of these two types of responses (hereafter, nonscorable responses) make automated scoring more difficult and result in a decreased score reliability for these responses. Therefore, in order to increase the reliability of the automated scoring system, we propose a two-step hybrid approach to operational scoring of spoken responses: first, nonscorable responses are filtered out by a filtering model and sent to scoring leaders for human scoring; then, only the remaining responses are scored using the scoring model from the automated system.

A decision tree-based filtering model was developed as a baseline using a combination of SpeechRater features and additional acoustic features (see Table 9). It was trained on the scoring model training data set and evaluated on the scoring model evaluation data set. The acoustic features were designed to improve the model's performance in detecting loud background noise, distorted speech, and nonspeech. The filtering model was tested on the scoring model evaluation data and obtained an accuracy rate (the exact agreement between the filtering model and H1 concerning the distinction between scorable and nonscorable) of 97%; it identified 90% of the nonscorable responses in the data set, with a false positive rate of 21%.

Extended Filtering Models

After the first (preliminary) evaluation of the TEFT Speaking scoring system (results reported below), our goal was to exclude a larger subset of problematic or nonscorable responses from the process of automated scoring and thereby improve the scoring accuracy of the overall system. We first identified different subsets of responses for which the system score was likely to be incorrect using the filtering model. From previous SpeechRater research on TPO data, we had demonstrated that the SpeechRater scores are less reliable when the responses include only a few words and when the mean confidence score[7] of the speech recognizer is low. Based on this previous research, we implemented two additional filters for TEFT:

- Short response filter: If a response contains two words or less, then the response is filtered out.
- Confidence score filter: If the average word-level confidence score of the recognized word hypothesis is lower than 0.4, then the response is filtered out.

Furthermore, a third additional filter was implemented for the low-entropy items: a repetitive response filter. In the TEFT assessment, some students repeated their response multiple times, possibly with minor modifications in the repetitions, for low-entropy items (in particular, MC and RP), and other students inserted creative sentences that were not contained in the stimulus materials. Human raters do not penalize the responses containing these additional sentences, as long as the content is consistent with the task. In contrast, the current automated system does not have a method to handle such additional material and penalizes these responses because their content is a poor match to the models of expected content. To address this issue, we implemented a filter to exclude responses that contain more than twice as many words as the expected response.

Results

The automated scoring system was evaluated using the scoring model evaluation partition. From the 8,400 responses, all nonscorable responses and all responses with system failures[8] were excluded. Also, speakers that had more than a set threshold of TD scores (final scores by human raters) were excluded as well. Only the remaining 7,776 responses[9] were scored by the automated system.

Responses for which the automated scoring engine could not produce a score were scored by a single human rater during operational deployment of the TEFT assessment. In order to simulate this operational scenario, the TD scores produced by the automated scoring system in the evaluation were replaced with the score of the second human rater (H2) when an H2 score was available. Otherwise, it was replaced by the average system score over the scorable responses for that speaker. The H1 score (score of the first human rater) was calculated based on the standard imputation rule: If H1 had a score of TD, it was replaced by the average H1 score over all scorable responses from that speaker.[10] Finally, the speaker-level score was calculated by summing up all 21 raw scores or imputed scores.

Preliminary Results

The preliminary results presented in this subsection were obtained using the baseline filtering model settings as described previously. After this first evaluation, additional filtering models were added to the system in order to flag additional responses that should not be scored by automated scoring models (and would thus be rerouted to human scoring leaders during operation). The final evaluation results with the extended set of filtering models in place are reported below.

Table 10 presents the correlations between the automated and human scores, after score imputation, computed between sums over all responses for each candidate (28 speakers in the evaluation partition were excluded from this analysis because they had too many TD responses). This table shows the following three correlations: system with H1, system with H2, and H1 with H2.

Table 11 shows human and machine score mean and standard deviations for each item type, and Tables 12 through 14 provide comparisons between human – system (H1-S) and human – human (H1-H2) correlations and quadratic kappas at the item-type level (Table 12) and item level (Tables 13 and 14).

In all of these analyses, the scoring model evaluation partition, which excludes all nonscorable responses, was used. At the item level, correlations between H1 and the system range between .36 and .57 for medium-entropy items and between .25 and .80 for low-entropy items, respectively (Tables 13 and 14). Quadratic weighted kappa ranges between .34 and .56 for medium-entropy items and between .25 and .76 for low-entropy items.

Table 10 Correlations Between Automated Scoring System and Human (H1/H2) Rater Scores

Score	H1	H2
Automated system	.72	.74
H1	-	.93

Note. Correlation computed between sums over the responses from each candidate, with imputation, $N = 272$.

Table 11 Mean Human (H1) and Machine Scores and Standardized Mean Score Differences by Item Type

Item type	Number of operational items (Pilot Form 1)	*n*	H1		SR		Std. mean score diff. H1-SR
			Mean	*SD*	Mean	*SD*	
CH	5	1,372	2.42	0.64	2.51	0.65	−0.09
IS	1	260	1.95	0.76	1.93	0.88	0.02
KC	1	274	2.26	0.65	2.26	0.72	0.00
KW	1	275	2.43	0.69	2.46	0.71	−0.03
MC	4	1,036	2.50	0.73	2.51	0.77	−0.01
RA	6	1,653	2.58	0.55	2.56	0.57	0.01
RP	2	543	2.45	0.61	2.57	0.66	−0.12
VI	1	272	2.54	0.62	2.72	0.62	−0.18

Note. SR = SpeechRater; CH = chart; IS = incomplete sentence; KC = keyword chart; KW = key words; MC = multiple-choice read aloud; RA = read aloud; RP = repeat aloud; VI = two visuals.

Table 12 Results by Item Type

Item type	*n*	Correlation			Quadratic kappa		
		H1-SR	H1-H2	Degradation	H1-SR	H1-H2	Degradation
CH	1,372	.44	.67	−.23	.44	.67	−.23
IS	260	.46	.69	−.23	.46	.68	−.23
KC	274	.57	.74	−.17	.56	.74	−.17
KW	275	.44	.67	−.23	.44	.67	−.22
MC	1,036	.67	.83	−.16	.67	.83	−.16
RA	1,653	.34	.51	−.18	.34	.51	−.17
RP	543	.41	.73	−.32	.40	.73	−.33
VI	272	.43	.80	−.36	.42	.79	−.38

Note. H1 = first human rater; H2 = second human rater; SR = SpeechRater; CH = chart; IS = incomplete sentence; KC = keyword chart; KW = key words; MC = multiple-choice read aloud; RA = read aloud; RP = repeat aloud; VI = two visuals. Degradation is the differences between H1-H2 and H1-SR measures. Correlation is Pearson *r*. Values computed at the response level, without aggregation by item type or candidate.

As shown in the preceding tables, the automated system achieved better performance in keyword chart/MC/read aloud items than other items; the degradations in both correlation and kappa for these items were smaller than .20. The system obtained low performance in scoring repeat aloud/two visuals items, and the degradations were over .30.

Table 15 presents the correlations between the speaking scores and the skill scores obtained from the 21 items in the other sections of the TEFT assessment. These skill scores include listening, reading, and writing. In addition, we also report correlations with the total scores (the sum of the listening, reading, writing, and speaking scores).

Final Results

After the preliminary evaluation of the automated scoring system reported in the previous subsection, we extended the filtering model substantially with the three additional components described above. Table 16 provides the results of a final evaluation, in which the problematic responses were first identified by the filtering model. Then, the SpeechRater scores for these responses were replaced with the H2 score if it was available. If H2 was unavailable, then H2 was imputed.

As Table 16 shows, with the addition of three filtering models we were able to achieve a substantially higher correlation with human scores, as compared to the results of the preliminary system that included only a simple baseline filtering

Table 13 Item-Specific Results for Medium-Entropy Items

			Correlation			Quadratic kappa		
Item type	Item	*n*	H1-SR	H1-H2	Degradation	H1-SR	H1-H2	Degradation
CH	1	272	.57	.71	−.15	.56	.71	−.15
	2	277	.41	.62	−.21	.41	.62	−.21
	3	275	.41	.61	−.20	.40	.61	−.20
	4	276	.39	.55	−.16	.34	.55	−.20
	5	272	.36	.76	−.41	.35	.76	−.41
IS	6	260	.46	.69	−.23	.46	.68	−.23
KC	7	274	.57	.74	−.17	.56	.74	−.17
KW	8	275	.44	.67	−.23	.44	.67	−.22
VI	9	272	.43	.80	−.36	.42	.79	−.38

Note. H1 = first human rater; H2 = second human rater; SR = SpeechRater; CH = chart; IS = incomplete sentence; KC = keyword chart; KW = key words; VI = two visuals. Degradation is the differences between H1-H2 and H1-SR measures.

Table 14 Item-Specific Results for Low-Entropy Items

			Correlation			Quadratic kappa		
Item type	Item	*n*	H1-SR	H1-H2	Degradation	H1-SR	H1-H2	Degradation
MC	1	257	.80	.88	−.08	.78	.88	−.10
	2	256	.76	.87	−.12	.73	.87	−.14
	3	260	.63	.75	−.12	.62	.75	−.13
	4	263	.43	.65	−.22	.40	.65	−.25
RA	5	274	.42	.53	−.12	.41	.53	−.12
	6	274	.41	.55	−.14	.40	.55	−.14
	7	275	.35	.53	−.18	.34	.52	−.18
	8	277	.32	.58	−.26	.32	.58	−.26
	9	279	.30	.35	−.05	.28	.34	−.06
	10	274	.25	.49	−.24	.25	.49	−.24
RP	11	268	.41	.72	−.31	.40	.72	−.32
	12	275	.40	.74	−.34	.39	.74	−.34

Note. H1 = first human rater; H2 = second human rater; SR = SpeechRater; MC = multiple-choice read aloud; RA = read aloud; RP = repeat aloud. Degradation is the differences between H1-H2 and H1-SR measures.

Table 15 Correlations Between System/Human (H1) Speaking Scores and Skill Scores for 21 Items

Items	Listening	Reading	Writing	Total score
Automated speech system scores – 21 speaking items	.49	.56	.59	.70
H1 scores – 21 speaking items	.62	.67	.70	.87

Note. Correlation computed between sums over the responses from each candidate, with imputation, $N = 272$.

model (.81 vs. .73). This performance increase, however, comes with the additional expense of a larger subset of total responses that need to be scored by human raters during an operational administration (hybrid scoring approach).

Discussion

This report addresses developing automated scoring models for speaking items of the TEFT assessment in the ELTeach program. The TEFT assessment is a language achievement test for English teachers who are not native speakers of English and is designed to measure language skills for performing essential functions in an English-speaking English-language learner classroom, after teachers have studied the related curriculum materials.

Although for some areas of the item type constructs we could use features previously developed in SpeechRater (e.g., fluency, pronunciation, and grammar), for other areas, in particular for content accuracy for semipredictable item types, new features had to be developed (Zechner & Wang, 2013). While we believe that the final feature sets chosen for the eight scoring models for each item type constitute a reasonable representation of their constructs, there is certainly still room

Table 16 Effect of Replacing SpeechRater (SR) Scores With Human (H1/H2) Scores for Filtered Responses

Filter	Number of responses filtered	H1-SR (corr.)	H1-SR (kappa)	H2-SR (corr.)	H2-SR (kappa)
ShortResponseFilter	259	.77	.77	.78	.77
ConfidenceScoreFilter	236	.76	.76	.78	.76
RepetitiveResponseFilter	351	.76	.76	.78	.76
Combination of three filters	690	.81	.81	.84	.83

Note. Evaluations are on the speaker level with 21 items per speaker.

for improvement in this regard, for example, adding features measuring more aspects of prosody or grammar than what is measured currently.

The feature correlations with human rater scores vary widely across item types and subconstructs, but we note that overall, content accuracy related features perform fairly well and robustly with correlations of .40 to .50 for most item types, while delivery-related features achieve a comparatively lower performance (around $r = .20$; see Appendix E).

As we analyzed response lengths and delivery feature performance, it became clear that in order to obtain more reliable and stable estimates for such delivery features (e.g., fluency and pronunciation) we needed to measure them on samples longer than a single response. Therefore, we implemented a concatenation approach whereby seven speaking items of each lesson are concatenated and delivery features are computed based on this concatenation, while content accuracy features are computed on each individual spoken response. As a result of this concatenation, we found that delivery feature correlations with human rater scores increased overall by around .13, averaged over all eight item types and six features evaluated (from fluency, pronunciation, and prosody). Results were mixed, however, for different item types, with five item types seeing an overall improvement in correlation when using concatenation (averaged over all six features) and three item types showing a decrease. When looking at the individual features, we observed that one feature did not see changes due to concatenation and two features had mixed results; however, three features exhibited a clear improvement, with six or more item type correlations improving for feature concatenation. Moreover, in a separate effort of human analytic scoring for content accuracy and delivery, we found that content scores vary more than delivery scores for a test taker (i.e., content accuracy scores have a standard deviation that is about 20% higher than that of delivery scores), further supporting the argument for using the concatenation approach.

Another observation we made during this work was the high rate of nonscorable responses (about 9% based on human ratings) for reasons such as nonresponses by the candidates, high-noise levels in recordings, and so forth. Although we used an initial baseline filtering model to handle the most frequent classes of these responses, we substantially extended this filtering model after the initial system evaluation to ensure that only responses with a high likelihood of being able to be scored by the automated system are actually routed to the system, whereas the remainder of responses (nonscorable) would be scored by human raters in an operational setting. This led to a hybrid approach whereby, for the pilot data we are using, about 80% of responses are scored automatically and 20% are scored by human raters. As reported previously, this hybrid approach was able to substantially improve the overall correlation between final system scores and human rater scores from .73 to .81, due to this mixing in of human raters for a subset of about 20% of responses that are deemed nonscorable by the extended filtering model. In the final evaluation, the correlation between the hybrid system and human raters, aggregating 21 speaking items, was .81, compared to an interrater correlation for the same data set of .93.

When looking at item-type-specific performance, we noted that the correlations and quadratic weighted kappas vary quite substantially, from as low as .34 for read-aloud items to .67 for multiple-choice read-aloud items. For some items, even the reference interhuman agreement can be quite low; for example, only .51 for read-aloud items, which may be a factor that contributes to these results. Another likely factor is the highly skewed distribution of candidate scores across many items. In an ideal (normal) distribution, the mean score for a 3-point item spoken response set with a scale of 1 – 3 would be 2.0, but we observed mean scores of 2.50 on average and as high as 2.72 for the VI items. Since the TEFT assessment is not a high-stakes language proficiency test (such as TOEFL or TOEIC) but rather a low-stakes achievement test meant to indicate the extent to which a candidate is able to master the materials presented in the preceding curriculum, this skewness of score distribution can be considered a matter of design, rather than an issue or a problem. However, it is still the case that building automated scoring models for a population whose scores are skewed to the higher end of the score scale is substantially more challenging than it would be for a population with a more normal score distribution.

Conclusions and Future Work

In this research report, we have described the process of building scoring models for automated scoring of TEFT Speaking items using ASR technology, a set of diverse features covering all main areas of the test construct, several filtering models that flag nonscorable responses and route them to human raters, and item-type-specific linear regression models for score prediction. We had to address numerous challenges in this work, including but not limited to a highly skewed score distribution of the data, sometimes low interrater agreement, a high percentage of nonscorable responses, a mix of different native languages of test takers, and a large number of different item types. The overall end-to-end system performance, where nonscorable responses are receiving a score by a human rater, was measured as $r = .81$ between the system and an independent human rater, when aggregating 21 speaking items per test taker.

We were able to demonstrate some significant accomplishments with this work, including the following:

- TEFT represents an innovative hybrid of both human and machine scoring. It is also innovative in the way delivery features are generated, such that relatively sophisticated delivery features can be obtained from short responses via response concatenation.
- The scoring model approach used optimizes both quality of prediction and construct coverage. Overall, the system represents a balance between scoring efficiency, scoring quality, and validity of scores.

In future work, we plan to extend and expand the set of SpeechRater features in order to cover the construct areas more completely, improve scoring models and filtering models, as well as the ASR system that operates as the first stage in the SpeechRater pipeline.

Notes

1 The term *feature* in this report refers to automatically computed linguistic measures based on speech recognition output, for example, speaking rate, pronunciation accuracy, and so forth.
2 If H1 was either 0 or TD, then the response went directly to adjudication, and no H2 was provided.
3 The ASR training set was further used to build and train the content features (e.g., n-gram and regular expression based features).
4 Using this large corpus resulted in a better acoustic model than using the comparatively smaller TEFT ASR training corpus.
5 We obtained confirmation for this assumption in a study where responses were rated analytically on two dimensions (delivery and content accuracy) by human raters. Standard deviations of human delivery scores were on average smaller than those for human content scores (0.10 vs. 0.12).
6 Initial experiments using delivery features extracted from the individual responses were also conducted, and the concatenation-based approach was shown to provide slightly better overall performance.
7 The confidence score is a type of self-diagnosis of a speech recognizer's performance. The speech recognizer may generate a word hypothesis even though the input speech does not match the pretrained models well. In this case, it outputs low confidence scores for the words.
8 These are cases in which the automated scoring system cannot generate a score due to a technical failure such as speech recognition failure or pitch generation failure.
9 The number of responses to scored items in this data set was 5,865, while the number of responses to additional items was 1,911.
10 When a given speaker has more than a set threshold of responses that received a score of TD, the speaker is excluded here, since he or she will not receive a score in an operational TEFT administration.

References

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests, *Language Testing*, *27*, 355–377.

Chandel, A., Parate, A., Madathingal, M., Pant, H., Rajput, N., Ikbal, S., . . . Verma, A. (2007). Sensei: Spoken language assessment for call center agents. In *2007 IEEE Workshop on Automatic Speech Recognition & Understanding* (pp. 711–716). Abstract retrieved from http://dx.doi.org/10.1109/ASRU.2007.4430199

Chen, L., & Yoon, S. (2012, September). *Application of structural events detected on ASR outputs for automated speaking assessment*. Paper presented at the 13th annual conference of INTERSPEECH, Portland, OR. Retrieved from http://www.researchgate.net/profile/Lei_Chen32/publication/260593349_Application_of_Structural_Events_Detected_on_ASR_Outputs_for_Automated_Speaking_Assessment/links/00463531a99f2d568d000000.pdf

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America*, *111*, 2862 – 2873.

Eskenazi, M. (2009). An overview of spoken language technology for education. *Speech Communication*, *51*(10), 832 – 844.

Freeman, D., Katz, A., Garcia Gomez, P., & Burns, A. (2015). English-for-Teaching: Rethinking teacher proficiency in the classroom. *ELT Journal.* Advance online publication. http://dx.doi.org/10.1093/elt/ccu074

Jeon, J. H. & Yoon, S-Y. (2012, September). *Acoustic feature-based non-scorable response detection for an automated speaking proficiency assessment.* Paper presented at the 13th annual INTERSPEECH conference, Portland, OR. Retrieved from http://www.isca-speech.org/archive/archive_papers/interspeech_2012/i12_1275.pdf

Mostow, J., Roth, S., Hauptmann, A., & Kane, M. (1994). A prototype reading coach that listens. In *Proceedings of the Twelfth National Conference on Artificial Intelligence* (Vol. 1, pp. 785 – 792). Menlo Park, CA: American Association of Artificial Intelligence.

Yoon, S., & Bhat, S. (2012). Assessment of ESL learners' syntactic competence based on similarity measures. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 600 – 608), Stroudsburg, PA: Association for Computational Linguistics.

Yoon, S., Bhat, S., & Zechner, K. (2012, June). *Vocabulary profile as a measure of vocabulary sophistication*. Paper presented at the 7th Workshop on Innovative Use of NLP for Building Educational Applications, NAACL-HLT, Montreal, Canada. Retrieved from http://aclweb.org/anthology-new/W/W12/W12-2021.pdf

Yoon, S., Evanini, K., & Zechner, K. (2011). Non-scorable response detection for automated speaking proficiency assessment. In *Proceedings of the 6th NAACL-HLT Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 152 – 160). Stroudsburg, PA: Association for Computational Linguistics.

Zechner, K., Higgins, D., Xi, X., & Williamson, D. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, *51*(10), 883 – 895.

Zechner, K., & Wang, X. (2013, June). *Automated content scoring of spoken responses in an assessment for teachers of English*. Paper presented at the 8th Workshop on Innovative Use of NLP for Building Educational Applications, Atlanta, GA. Retrieved from http://aclweb.org/anthology/W/W13/W13-1709.pdf

Appendix A

TEFT Scoring Rubrics

Table A1 Low-Entropy Items: Multiple-Choice Read Aloud, Read Aloud, and Repeat Aloud

Score	Response description
3 (Successful)	• The text presented (or correct option) is read aloud or repeated aloud, AND • Delivery is highly intelligible but may require a small amount of listener effort. Other language influence may be present but does not affect overall intelligibility. • Variations of the content (additions, omissions) may exist but they do not affect overall meaning.
2 (Partially successful)	• The text presented (or correct option) is read aloud or repeated aloud, BUT • The text or correct option may be only partially complete or accurate. • Delivery is generally intelligible but some listener effort is required. • Variations (additions, omissions) of the content may obscure meaning at times.
1 (Unsuccessful)	The response may be marked by one or more of the following: • The text presented (or correct option) is read or repeated aloud but is substantially incomplete or inaccurate, OR • A text different from the one presented is read aloud or repeated aloud. • An incorrect option/s is read aloud or all options are read aloud. • Delivery is mostly unintelligible and significant listener effort is required.
0	No response, or only a language other than English is used.

Table A2 Medium-Entropy Items: Chart, Incomplete Sentence, Keyword Chart, Key Words, Two Visuals

Score	Response description
3 (Successful)	**The test taker completes the task. The content of the response is accurate and appropriate to the task. The delivery is highly intelligible.** • The language used fulfills the demands of the task, AND • Delivery requires little listener effort. Other language influence may be present but does not affect overall intelligibility.
2 (Partially successful)	**The test taker partially completes the task. The content of the response is not fully accurate, or not fully appropriate to the task. The delivery is generally intelligible.** • The language used contains inaccuracies (grammar, vocabulary, omissions, additions) that may substantially affect meaning at times. • Delivery is generally intelligible but some listener effort may be required.
1 (Unsuccessful)	**The test taker does not complete the task. The content of the response is inaccurate or inappropriate to the task, is substantially incomplete, or delivery is mostly unintelligible.** • The language used contains inaccuracies that obscure meaning or does not address the task. • Delivery is mostly unintelligible and significant listener effort is required.
0	No response, or only a language other than English is used.

Appendix B

Alpha Codes for Speaking Responses Receiving Technical Difficulty (TD) and 0 Scores

Table B1 Alpha Codes Provided by Adjudicators for TEFT Speaking Responses Scored as Technical Difficulty (TD)

Alpha code	Description
MS	Missing samples: Parts of the audio are missing. Voice often cuts in and out.
NC	Noise constant: Constant noise in the recording obscures candidate response (buzzing, clicking, crackling, static, or other mechanical noise).
DR	Distorted recording: Speech recording is too distorted to fairly evaluate. Includes fast playback, slow playback, or overamplified and distorted speech recording without any noise constant.
SI	No response—Total Silence: The silence in this category is what you hear if the microphone was not plugged in. No background or room noise can be heard. (Note: This is not the same as the candidate simply not speaking but breathing, coughing, or the occasional background sound can be heard).
OT	Other: Any audio problem that obscures the candidate response not covered by the above categories.

Table B2 Alpha Codes Provided by Adjudicators for TEFT Speaking Responses Scored as 0

Alpha code	Description
XE	Non-English response: Response is spoken entirely in another language.
XU (only for low-entropy items)	Unintelligible speech.
XS	No speech: Speaker does not attempt to respond to the task, but background noise such as breathing, coughing, or room noise can be heard.
XT	Off topic: Entire response is inappropriate/not relevant to the task.
XO	Other.

Appendix C

List of Features Used in Scoring Models

Table C1 List of Features by Construct Dimension (Part 1)

Construct	Subconstruct	Feature	Definition
Delivery	Fluency	ipc	Number of interruption points (IP) of repair/repeat disfluencies per clause
Delivery	Fluency	longpfreq	Number of long silences per word
Delivery	Fluency	longsilratio	Proportion of long within-clause silences to all within-clause silences
Delivery	Fluency	repfreq	Number of repetitions per word
Delivery	Fluency	silpwd	Number of silences per word
Delivery	Fluency	wdpchk	Average chunk length in words (chunks are word sequences not interrupted by silence)
Delivery	Fluency	withinclausesilmean	Average duration of all within-clause silences
Delivery	Fluency	wpsecutt	Number of words per second
Delivery	Pronunciation	phn_shift	Average difference in normalized vowel durations compared to native speech
Delivery	Prosody	relstresspct	Relative frequency of stressed syllables in percent
Delivery	Prosody	stresyllmean	Mean distance between stressed syllables in syllables

Table C2 List of Features by Construct Dimension (Part 2)

Construct	Subconstruct	Feature	Definition
Language use	Grammar	lmscore	Language model score
Language use	Grammar	poscvamax	Score level of highest similarity between response part-of-speech sequences and sets of training responses
Language use	Vocabulary	tpsec	Number of word types per second
Language use	Vocabulary	tpsecutt	Number of word types per second (ignoring initial and final silence)
Content accuracy	Content_low	cwpm	Correctly read words per minute
Content accuracy	Content_low	lowentwer	WER between prompt and ASR hypothesis
Content accuracy	Content_medium	bleu_s3	BLEU score between response and set of high-level responses
Content accuracy	Content_medium	re_match	Regular expression based match of response
Content accuracy	Content_medium	wer_s3	WER between response and set of high-level responses

Appendix D

Feature Sets Used in Scoring Models

Table D1 Scoring Model Features for Low-Entropy Item Types

	Item type		
Construct dimension	Multiple-choice read aloud	Read aloud	Repeat aloud
Content	cwpm	cwpm	cwpm
	lowentwer	lowentwer	lowentwer
Delivery			
Fluency	withinclausesilmean	repfreq	repfreq
	wpsecutt	wpsecutt	ipc
		silpwd	longsilratio
Prosody	relstresspct	relstresspct	stretimemdev
Pronunciation	phn_shift	phn_shift	phn_shift
Language use			
Vocabulary	n/a	n/a	n/a
Grammar	n/a	n/a	n/a

Table D2 Scoring Model Features for Medium-Entropy Item Types

	Item type				
Construct dimension	Incomplete sentence	Keyword	Chart	Keyword/chart	Two visuals
Content	re_match wer_s3 bleu_s3	re_match wer_s3 bleu_s3	re_match wer_s3	re_match wer_s3	re_match wer_s3 bleu_s3
Delivery					
Fluency	wdpchk withinclausesilmean wpsecutt	ipc silpwd wpsecutt stresyllmean	longsilratio	repfreq	wdpchk longpfreq stresyllmean
Prosody	stretimemdev	phn_shift	stretimemdev	relstresspct	(no usable features)
Pronunciation	phn_shift		phn_shift	(no usable features)	phn_shift
Language use					
Vocabulary	tpsecutt	tpsecutt	tpsecutt	tpsecutt	tpsec
Grammar	poscvamax	poscvamax	poscvamax	lmscore	poscvamax

Appendix E
Correlations Between SpeechRater Features and Human Scores

Table E1 Correlations Between SpeechRater Features (After Transformations) and Human Scores for Low-Entropy Item Types

	Item type					
Construct dimension	Multiple choice read aloud	Correlation	Read aloud	Correlation	Repeat aloud	Correlation
Content	cwpm	.64	cwpm	.33	cwpm	.40
	lowentwer	.67	lowentwer	.38	lowentwer	.46
Delivery	withinclausesilmean	.26	repfreq	.20	repfreq	.20
	wpsecutt	.21	wpsecutt	.19	ipc	.25
	relstresspct	.24	silpwd	.22	longsilratio	.33
	phn_shift	.21	relstresspct	.22	stretimemdev	.18
			phn_shift	.22	phn_shift	.20

Table E2 Correlations Between SpeechRater Features (After Transformations) and Human Scores for Medium-Entropy Item Types

	Item type									
Construct dimension	Incomplete sentence	Corr.	Keyword	Corr.	Chart	Corr.	Keyword/ chart	Corr.	Two visuals	Corr.
Content	re_match	.56	re_match	.45	re_match	.34	re_match	.61	re_match	.37
	wer_s3	.54	wer_s3	.47	wer_s3	.47	wer_s3	.58	wer_s3	.37
	bleu_s3	.50	bleu_s3	.45					bleu_s3	.43
Delivery	wdpchk	.28	ipc	.25	longsilratio	.25	repfreq	.30	wdpchk	.21
	withinclausesilmean	.31	silpwd	.23	stretimemdev	.19	relstresspct	.12	longpfreq	.20
	wpsecutt	.28	wpsecutt	.21	phn_shift	.23			stresyllmean	.21
	stretimemdev	.27	stresyllmean	.26					phn_shift	.13
	phn_shift	.31	phn_shift	.22						
Language use	tpsecutt	.32	tpsecutt	.26	tpsecutt	.29	tpsecutt	.21	tpsec	.26
	poscvamax	.49	poscvamax	.40	poscvamax	.40	lmscore	.23	poscvamax	.34

Suggested citation:

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). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12080

Action Editor: Beata Beigman Klebanov

Reviewers: Jidong Tao and Swapna Somasundaran