Listening. Learning. Leading.®

Research Report
ETS RR–15-09

Automated Scoring for the *TOEFL Junior*® Comprehensive Writing and Speaking Test

Keelan Evanini

Michael Heilman

Xinhao Wang

Daniel Blanchard

June 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 for the *TOEFL Junior*® Comprehensive Writing and Speaking Test

Keelan Evanini, Michael Heilman, Xinhao Wang, & Daniel Blanchard

Educational Testing Service, Princeton, NJ

This report describes the initial automated scoring results that were obtained using the constructed responses from the Writing and Speaking sections of the pilot forms of the *TOEFL Junior*® Comprehensive test administered in late 2011. For all of the items except one (the edit item in the Writing section), existing automated scoring capabilities were used with only minor modifications to obtain a baseline benchmark for automated scoring performance on the TOEFL Junior task types; for the edit item in the Writing section, a new automated scoring capability based on string matching was developed. A generic scoring model from the *e-rater*® automated essay scoring engine was used to score the email, opinion, and listen-write items in the Writing section, and the form-level results based on the five responses in the Writing section from each test taker showed a human – machine correlation of $r = .83$ (compared to a human – human correlation of $r = .90$). For scoring the Speaking section, new automated speech recognition models were first trained, and then item-specific scoring models were built for the read-aloud picture narration, and listen-speak items using preexisting features from the *SpeechRater*SM automated speech scoring engine (with the addition of a new content feature for the listen-speak items). The form-level results based on the five items in the Speaking section from each test taker showed a human – machine correlation of $r = .81$ (compared to a human – human correlation of $r = .89$).

Keywords e-rater; SpeechRater; TOEFL Junior Comprehensive test; automated scoring

doi:10.1002/ets2.12052

The *TOEFL Junior*® Comprehensive test is a computer-based test that assesses English communication skills, including reading comprehension, listening comprehension, speaking, and writing. The test is targeted at students aged 11 years and older who are learning English as a foreign language, and it elicits a range of written and spoken constructed responses. The purpose of this study was to investigate how well existing automated scoring capabilities developed for other assessments can score these constructed responses with minimal system development effort. These results provide a baseline level of performance that can be improved on in the future with research focused on developing new automated scoring features and capabilities that specifically address unique aspects of the tasks included in the TOEFL Junior Comprehensive test.

Automated scoring capabilities have been developed at Educational Testing Service (ETS) for the Writing and Speaking sections of the *TOEFL iBT*® test, and the purpose of this research is to investigate how well these capabilities transfer both to the different task types included in the TOEFL Junior Comprehensive test and to the different population of test takers (namely, middle school students). Specifically, the *e-rater*® engine, ETS's capability for automated essay scoring, has a long history of development and evaluation using TOEFL iBT essays, and it is currently used as a contributory score for operational TOEFL iBT essays. Ramineni, Trapani, Williamson, Davey, and Bridgeman (2012) provided detailed statistics about e-rater's performance across a range of *TOEFL*® independent and integrated prompts. The *SpeechRater*SM engine, ETS's capability for automated speech scoring, is currently used for scoring the TOEFL Practice Online (TPO™), which contains retired TOEFL iBT items. Zechner, Higgins, Xi, and Williamson (2009) and Higgins, Xi, Zechner, and Williamson (2011) provided detailed information about SpeechRater's architecture and its performance on TPO data. In this study, these two automated scoring capabilities were used with minor modifications to score the constructed responses obtained from a pilot version of the TOEFL Junior Comprehensive assessment, and a new automated scoring capability was developed for the edit item in the Writing section, because it elicits written responses that cannot be scored by e-rater.

This report is organized as follows. First, the Data section of this report briefly introduces the item types included in the TOEFL Junior Comprehensive test and describes the data that were used in the experiments. The Automated Scoring of Written Responses section in this report presents the approach that was used for scoring the responses to the Writing

Corresponding author: K. Evanini, E-mail: kevanini@ets.org

section of the TOEFL Junior Comprehensive test as well as the results that were obtained. The Automated Scoring of Spoken Responses section of this report presents the details of the speech recognition system and the SpeechRater models that were used for scoring the responses to the Speaking section and the results that were obtained for that section. The Discussion section of this report covers the performance gap between the baseline automated scoring systems presented in this report and human–human agreement in relation to the TOEFL Junior Comprehensive scoring rubrics. Finally, the Conclusion section of this report summarizes the main results from this study.

Data

The data used in this study were collected in a pilot version of the TOEFL Junior Comprehensive assessment administered in late 2011. Six test forms were used in the pilot, and each contained six items in the Writing section and five items in the Speaking section. In total, the dataset contains constructed responses from 3,385 test takers. The average age of the participants was 13.1 years ($SD = 2.3$), and there were 1,847 females (54.6%) and 1,538 males (45.4%). The following native language backgrounds were represented among the test takers: Arabic, Chinese, French, German, Indonesian, Japanese, Javanese, Korean, Madurese, Polish, Portuguese, Spanish, Thai, and Vietnamese.

The Writing section consisted of 6 items from five distinct task types: edit, email, opinion, dictation, and listen-write (each pilot administration included two edit items); however, the dictation items are not included in this study and will not be discussed further. The Speaking section consisted of 5 items from four distinct task types: read aloud, picture narration, listen-speak nonacademic, and listen-speak academic (each pilot test form included two listen-speak academic items). The six test forms used in the pilot test each contained a unique set of items. The dataset available for analysis in this report thus includes a total of 16,925 spoken responses and 16,925 written responses. All of the responses were double scored on a scale ranging from 1 to 4 by trained raters. Table 1 contains a brief description of the item types in the Writing and Speaking sections.[1]

The email, opinion, and listen-write items in the Writing section all elicit extended written responses that are approximately one paragraph in length. Although these responses can be scored by existing e-rater models, they are much shorter than the essays for which the existing version of e-rater was designed (the consequences of this fact will be discussed in more detail in the e-rater Engine section of this report). The responses to the edit item, conversely, consist of single words or short phrases and thus are not appropriate for e-rater. Therefore a new capability utilizing a string-matching approach was developed to score these items.

The read-aloud picture narration, listen-speak nonacademic, and listen-speak academic items in the Speaking section all elicit spoken responses that are 60 seconds in duration. These are similar in length to the TOEFL iBT spoken responses for which SpeechRater was designed; however, the distribution of lexical items in the responses is different than for TOEFL iBT responses because of the different content in the items. Furthermore, the read-aloud items elicit restricted speech from the test takers; in contrast, the other three items in the TOEFL Junior Speaking section elicit spontaneous speech (as do the items in the TOEFL iBT Speaking section). Because of these differences, new automated speech recognition models were trained before applying SpeechRater to the constructed responses in the Speaking section. Because the listen-speak nonacademic and listen-speak academic items elicited similar types of responses (for both types of items, the test takers were asked to summarize content related to three or four key concepts presented in the listening passage), they were combined for this study as the listen-speak item.

Automated Scoring of Written Responses

e-rater Results

In this section, we describe our approach to scoring the email, opinion, and listen-write items using e-rater. Because these items elicit multisentence written responses that assess some constructs similar to those assessed by existing TOEFL iBT writing tasks, we conducted a preliminary investigation into whether e-rater could be used to score these items accurately. The e-rater engine uses natural language processing technologies to automatically estimate writing proficiency, and it extracts 10 high-level variables, or features, to represent different aspects of writing proficiency, including grammar, word usage, mechanics, and so on. For a more detailed discussion of the features used in e-rater, see Attali and Burstein (2006) and Burstein, Tetreault, and Madnani (2013).

Table 1 Item Types Included in the TOEFL Junior Comprehensive Pilot Assessment

Section	Item type	Brief description
Writing	Edit	Test taker corrects four grammatical or mechanical errors in a paragraph
	Email	Test taker reads an e-mail message and writes a response
	Opinion	Test taker writes a paragraph presenting his or her opinion on a topic in the prompt
	Listen-write	Test taker listens to a presentation about an academic topic and writes a paragraph to answer a question about the content of the presentation
Speaking	Read aloud	Test taker reads a paragraph out loud
	Picture narration	Test taker narrates a story that is contained in a sequence of six pictures
	Listen-speak nonacademic	Test taker listens to a conversation about a school-related topic and provides a spoken response to answer a question about the content of the conversation
	Listen-speak academic	Test taker listens to a presentation about an academic topic and provides a spoken response to answer a question about the content of the presentation

Table 2 Sample Sizes and Mean Scores From the First Human Rater (*H1*) and e-rater for the Email, Opinion, and Listen-Write Items in the Writing Section

Item type	Sample size	*H1* mean	e-rater mean
Email	2,802	2.37	0.55
Opinion	2,970	2.29	1.75
Listen-write	3,122	2.09	0.93

The e-rater engine then uses a linear model, estimated from human-scored essays by a procedure based on least squares regression, to weight the various features depending on the task and produce a holistic score. For example, mechanics errors such as misspellings might be weighted differently in models for the *GRE*® test issue essays than in models for TOEFL independent essays. The e-rater engine provides various models that weight the various aspects of writing proficiency differently. For this experiment, we used a generic TOEFL model for expository text, which we felt was most appropriate for the TOEFL Junior task types.[2] However, we note that for a real application, it would be best to estimate new models from human-scored TOEFL Junior essays.

As mentioned in the Data section of this report, the email, opinion, and listen-write items are all scored on a scale ranging from 1 to 4. Responses that are not scorable receive either a 0 or a special code; examples of nonscorable responses include blank responses, responses in a language besides English, and off-topic responses. For this experiment, all of these nonscorable responses were filtered out, and only responses that received a valid score on the 4-point scale from all raters were retained.[3]

To measure agreement, we computed Pearson's r correlation coefficients and quadratic weighted kappa between e-rater scores and human scores from the first rater. For computing the Pearson correlations, we used the raw e-rater scores (i.e., the unrounded continuous scores that were not restricted to the range of scores for TOEFL essays). When computing quadratic weighted kappa, we restricted the range of the scores to 1 – 4 and rounded the scores to integers.

An important trend we observed in this experiment was that the e-rater scores were much lower on average than the human scores, as shown in Table 2. For example, whereas the mean human score for email items was 2.37, the mean e-rater score was only 0.55. We suspect that this difference is due primarily to the shorter length of the TOEFL Junior responses (which are usually only a few sentences long) because some e-rater features are closely related to essay length (e.g., features measuring the number and length of discourse elements).

To adjust for the difference in response length and thereby simulate an e-rater model built specifically for TOEFL Junior, we rescaled the e-rater scores to match the mean and standard deviation of the TOEFL Junior human scores. We standardized the e-rater scores by subtracting their mean and dividing them by their standard deviation. We then multiplied the standardized scores by the standard deviation of the human scores and added the mean of the human scores. We used twofold cross-validation to compute human – machine agreement for the rescaled scores (cross-validation was not used for the unscaled results). That is, the mean and standard deviations used for rescaling when evaluating on each half of the data were estimated from the other half of the data, and the results for the two halves were averaged.

Table 3 Agreement Statistics Comparing Scores From the First Human Rater (*H1*) to e-rater Scores (*M*), e-rater Scores Rescaled to Match the Means and Standard Deviations of Human Scores, and Scores From the Second Human Rater (*H2*)

	Pearson's *r*			Quadratic weighted kappa		
Item type	*H1*:*M*	*H1*:*M* (rescaled)	*H1*:*H2*	*H1*:*M*	*H1*:*M* (rescaled)	*H1*:*H2*
Email	.67	.67	.76	.22	.64	.76
Opinion	.65	.65	.72	.59	.62	.72
Listen-write	.66	.66	.79	.51	.65	.79

Table 3 presents the results of these evaluations.[4] In this table (and in subsequent tables), *H1* refers to the score from the first human rater, *H2* refers to the score from the second human rater, and *M* refers to the machine score. As the table shows, the automated scores have correlations of $r = .67$ (email), $r = .65$ (opinion), and $r = .66$ (listen-write) with human scores. Although this level of agreement is decent, it is lower than the human–human agreement levels for all three items; the amount of degradation from human–human agreement in terms of Pearson's *r* correlation is .09 for email, .07 for opinion, and .13 for listen-write. Furthermore, Table 3 shows that rescaling has a fairly strong positive effect on weighted kappa results; for example, the raw e-rater scores had a weighted kappa of only .22 with human scores for the email item type, but a weighted kappa of .64 after rescaling.

We believe that these results are encouraging and suggest that it would be feasible to score automatically these TOEFL Junior item types with e-rater. However, we believe that item-specific models, which would be trained from human-scored responses to match the means and standard deviations of human scores, would be more accurate than the generic model used in the current study.

String Matching Results

In this section, we describe an approach using a string-matching technique to score the responses to the edit item. In this item, the test taker is presented with a paragraph containing four highlighted errors and is asked to correct each of the errors without changing the meaning of the passage. The types of errors contained in this item include grammatical errors (such as verb tense errors and incorrect preposition usage) and mechanical errors (such as capitalization and punctuation errors).

During the process of creating each edit item, the test developers provided an answer key for each of the four errors in the passage specifying one or more correct responses. For example, in the following sentence, the words *were called* are highlighted as an error:

> On foggy days, crew members released large black birds *were called* ravens to see in which direction they should travel.

The answer key for this error might contain the following three possible corrections, which would all produce a grammatically correct sentence without changing the meaning of the passage: *that were called*, *which were called*, *called*.

To score this item, we adopted a simplistic string-matching approach: If the test taker's response exactly matched one of the correct answers provided in the key for that item, then it was scored as correct. Because each edit item contains four errors that need to be corrected, the automated score for the item is the number of errors that were corrected (on a scale ranging from 0 to 4)—this approach matches the human scoring rubrics, in which a score of 4 indicates that the test taker corrected all errors, a score of 3 indicates that three errors were corrected, and so on.

To evaluate the performance of the automated scoring procedure for the edit items, all responses that received a human or automated score of 0 were excluded. This was done because a human score of 0 can also indicate an invalid response, such as an off-topic response, no response, or a foreign language response (in addition to indicating a response in which the test taker did not successfully correct any of the errors). Table 4 presents the agreement between human scores and the automated scores based on the string-matching approach for the edit item in terms of Pearson's *r* and quadratic weighted kappa.

Table 5 presents the confusion matrix comparing the scores from the first human rater and the automated scores.

Table 4 Agreement Statistics Comparing Scores From the First Human Rater (*H1*) to Machine Scores (*M*) and Scores From the Second Human Rater (*H2*)

	Pearson's *r*		Quadratic weighted kappa	
Item type	*H1*:*M*	*H1*:*H2*	*H1*:*M*	*H1*:*H2*
Edit	.82	.91	.78	.91

Note. $N = 2{,}802$.

Table 5 Confusion Matrix for the Scores From the First Human Rater (*H1*) and the Machine Scores (*M*) for the Edit Item

	M			
H1	1	2	3	4
1	**931**[a]	13	0	0
2	197	**636**[a]	8	0
3	34	204	**351**[a]	3
4	5	25	86	**92**[a]

[a]Values in boldface indicate exact agreement between human and automated scores.

As the confusion matrix in Table 5 shows, nearly all of the instances of disagreement between human and machine scores have a higher human score than machine score; that is, the human raters were more likely than the automated scoring system to decide that a higher number of the four errors in the reading passage were corrected successfully by the test taker. This is likely because test takers sometimes provided a response that the raters deemed to be an appropriate correction of the error, even though it wasn't explicitly listed in the key for that item. The automated scoring system, however, necessarily scored these types of responses as incorrect, because of the nature of the approach based on exact string matching. To address this, test taker responses that received a score of 4 but were not included in the original key can be added to the automated scoring model. In addition, future approaches to automated scoring for the edit item could be made more robust by incorporating paraphrase matching technology (Heilman & Madnani, 2012).

Form-Level Results

To compute the form-level results for the Writing section, the scores from each of the five Writing items were summed, and then correlations were calculated between the summed human and machine scores. For this analysis, only test takers who had valid scores (1–4) for all five Writing items were included ($N = 1{,}244$). The results of this analysis show a correlation between human and machine scores (*H1*:*M*) of $r = .83$ and a corresponding human–human agreement (*H1*:*H2*) of $r = .90$.

Automated Scoring of Spoken Responses

In this section, we describe the approach taken to scoring the responses from the Speaking section of the TOEFL Junior Comprehensive pilot form.[5] The first necessary step in this process was to train a speech recognition system for TOEFL Junior data, and this work is described in the Automatic Speech Recognition System Development subsection of this report. Next, the SpeechRater Results subsection presents the SpeechRater features that were used to build scoring models for each of the item types as well as the performance of the scoring models. Finally, the Form-Level Results subsection presents the aggregated form-level results for the TOEFL Junior Speaking section.

Automatic Speech Recognition System Development

One of the major challenges for automated assessment of children's speech is the difficulty of building accurate automatic speech recognition (ASR) systems for children's speech. Owing to the differences in vocal tract length between children and adults, acoustic models trained on adult speech will produce worse results on children's speech. In addition, children may have different speech patterns in linguistic areas such as pronunciation, prosody, lexical choice, and syntax. To overcome these problems, several corpora containing only children's speech have been collected (CSLU, 2008; Hagen, Pellom,

Table 6 Data Partitions Used for the Automatic Speech Recognition (ASR) System and the Speaking Scoring Models

			Score			
Partition	No. speakers	No. responses	1	2	3	4
ASR training	1,625	7,300	1,434 (19.6)	3,065 (42.0)	2,088 (28.6)	713 (9.8)
ASR development	30	149	14 (9.4)	54 (36.2)	56 (37.6)	25 (16.8)
ASR evaluation	30	150	19 (12.7)	48 (32.0)	66 (44.0)	17 (11.3)
Model training	967	4,338	798 (18.4)	1,802 (41.5)	1,277 (29.4)	461 (10.6)
Model evaluation	733	3,297	664 (20.1)	1,368 (41.5)	918 (27.8)	347 (10.5)

Table 7 Word Error Rate Results of Two Different Language Model (LM) Adaptation Approaches With Several Different Interpolation Weights on the Automatic Speech Recognition (ASR) Development Partition

LM type	Interpolation weight	Read aloud	Picture narration	Listen-speak	Total
Generic LM	1.0	**12.3**[a]	31.6	35.2	29.5
	0.9	12.9	**31.4**[a]	**33.4**[a]	**28.6**[a]
	0.8	12.7	31.6	33.6	28.7
	0.7	13.6	32.3	33.9	29.2
	0.6	14.0	33.1	34.2	29.6
Item-type-specific LM	1.0	**9.7**[a]	30.8	33.1	27.6
	0.9	10.6	31.2	32.9	27.7
	0.8	10.6	**29.6**[a]	**32.6**[a]	**27.3**[a]
	0.7	11.5	29.8	32.7	27.6
	0.6	11.7	30.3	33.3	28.0

Note. $N = 149$.

[a]The values in boldface are the lowest word error rate for each item type.

& Cole, 2003; Kazemzadeh et al., 2005; Kantor et al., 2012; LDC, 1997) and have been used to train or adapt ASR systems so that they will perform better on children's speech.

The 2011 TOEFL Junior pilot administration collected a large number of responses from children from several different L1 backgrounds, so these data can be used for training or adapting an ASR system specifically for TOEFL Junior. To ensure an unbiased approach to ASR system training, complete sets of five responses from the participants in the pilot were partitioned into the following three sets: ASR training (for training the acoustic model [AM] and language model [LM] the ASR system uses), ASR development (for tuning parameters of the AM and LM), and ASR evaluation (for calculating the final ASR system performance results). In total, 7,300 pilot responses, totaling approximately 137 hours of audio data, were available from the TOEFL Junior pilot for ASR system training. In addition, participants from the pilot were also partitioned into two additional sets for developing and evaluating the linear regression scoring models—the model training and model evaluation partitions. Table 6 presents information about how many speakers and responses are contained in each of the five partitions. Furthermore, Table 6 presents the number of responses (and the corresponding percentages) at each score point for each of the partitions. Note that all spoken responses receiving a score of 0 or TD (technical difficulty) were excluded from the data partitions, as was done for the items in the Writing section.

To train the TOEFL Junior ASR system, we first started with an AM and LM that had been trained on a larger corpus (more than 800 hours) of TOEFL iBT data. Then, we adapted a new AM and LM based on these initial models using the data from the TOEFL Junior ASR training partition.[6] Two approaches to LM adaptation were investigated: (a) a single LM based on responses to all of the TOEFL Junior items was trained and interpolated into the existing background model and (b) three separate LMs for each of the three TOEFL Junior item types (read aloud, picture narration, and listen-speak) were trained and interpolated separately into the existing background model. For each of these two approaches, several different interpolation weights were tested on the ASR development set. Table 7 presents the results of these experiments.

As Table 7 shows, the item-type-specific adapted LM typically outperforms the generic adapted LM by about 1–2%, depending on the item type and LM interpolation weight. Because this difference is not very large, and because the generic LM reduces the complexity of configuring the SpeechRater automated scoring system, subsequent ASR and scoring model experiments were conducted with the generic adapted LM.

Table 8 Word Error Rate Results After Language Model (LM) and Acoustic Model (AM) Adaptation on the Automatic Speech Recognition (ASR) Development and ASR Evaluation Partitions

	Read aloud	Picture narration	Listen-speak
ASR development[a]	9.8	29.7	31.7
ASR evaluation[b]	9.7	26.5	29.4

[a] $N = 149$. [b] $N = 150$.

To obtain further improvements in the ASR performance, AM adaptation (using the maximum a posteriori adaptation procedure; Huang, Acero, & Hon, 2001) was conducted starting with the background TOEFL iBT AM and using the ASR training partition as adaptation data. Table 8 presents the WER results after AM adaptation (in combination with LM adaptation using the generic LM interpolated with a weight of .9) on the ASR development and evaluation partitions.

A comparison between the results on the ASR development partition in Table 8 and the results in Table 7 for the generic LM with an interpolation weight of .9 shows that AM adaptation resulted in an improvement of approximately 2 – 3%. This configuration with the generic adapted LM (with an interpolation weight of .9) and an adapted AM is the ASR configuration that was used for the SpeechRater model building experiments described in the SpeechRater Results subsection.

SpeechRater

Automated scoring of the spoken constructed responses from the TOEFL Junior Comprehensive assessment was conducted using SpeechRater (Zechner et al., 2009). First, ASR output for each response was obtained using the ASR configuration described in the ASR System Development subsection. Then, forced alignment was conducted using native-speaker acoustic models to obtain the word and phoneme boundaries. Subsequently, SpeechRater features were extracted using the speech signal, the ASR output, and the forced alignment output; the majority of these features assess aspects of the test taker's delivery, such as fluency, pronunciation, and prosody, but a few of the features also address aspects of language use, such as reading accuracy (for the read-aloud items).

Separate linear regression scoring models were trained for each of the three task types (read aloud, picture narration, and listen-speak) using the responses in the model training partition and a subset of SpeechRater features that were selected based on construct relevance and correlations with human scores on the model training partition. The subset of features that were used in this experiment are presented in Table 9 and are described in more detail in Zechner et al. (2009). Table 9 also indicates the Speaking subconstruct that each feature targets; as the table shows, most of the features target the delivery construct (represented by the fluency, pronunciation, and prosody subconstructs), but a few features also target the language use construct (represented by the grammar subconstruct) and the content construct (represented by the accuracy subconstruct for the read-aloud items and by the content subconstruct for the listen-speak items). Because the current set of SpeechRater features does not include metrics for assessing the content of spoken responses containing spontaneous speech, an additional set of features was developed to address this aspect for the listen-speak items.[7] One of these features, *factfreq_mean*, was also included in the scoring model and is described in more detail in Xiong, Evanini, Chen, and Zechner (2013).

Table 10 presents the correlations between the SpeechRater features and human scores for the responses to each of the three task types in the model training partition[8] (empty cells in the table indicate that the feature was not included in the scoring model for that task type).

The responses from the model evaluation partition were used to evaluate the performance of the scoring models that were trained for each task type. Table 11 presents the scoring model results in terms of Pearson correlations and quadratic weighted kappa (in this table, the correlations are computed based on the unrounded SpeechRater scores, and the kappa values are computed based on the rounded scores).

As Table 11 shows, the SpeechRater scores come closer to matching human – human agreement levels when the correlation values are considered as opposed to the weighted kappa values; this is because some information contained in the SpeechRater score is lost when it is rounded to the nearest integer. Table 11 also shows that the performance of the SpeechRater scores matches human – human agreement at $r = .70$ for the read-aloud items when it is evaluated using Pearson correlation and has a degradation in performance of .08 for the picture narration items and .06 for the listen-speak items when it is evaluated using Pearson correlation.

Table 9 Descriptions of SpeechRater Features Used in the Scoring Models

Subconstruct	Feature	Description
Fluency	*wpsecutt*	Number of words per second
	silpwd	Number of silences per word
	longpfreq	Number of long silences per word; long silence is a silence with duration ≤0.495 s
	tpsecutt	Number of lexical types per second
	wdpchk	Average chunk length in words (a chunk is defined as a segment whose boundaries are set by silences longer than 0.20 s)
Pronunciation	*conftimeavg*	Average speech recognizer word-level confidence score per second
	amscore	AM score; compares the pronunciation of nonnative speech to a reference pronunciation model trained on native speech
	L7	Average AM score density across all words, normalized by the rate of speech
	phn_shift	Average vowel duration shifts relative to a native-speaker model, normalized to account for differential durations of different vowel phones
Prosody	*stretimemean*	Mean distance between stressed syllables, in seconds
Grammar	*lmscore*	Global language model score (normalized by response length)
Accuracy	*cwpm*	Number of words read correctly per minute
Content	*factfreq_mean*	Mean number of correct facts about the target concept contained in each segment of a response

Table 10 Individual Feature Correlations Using Responses From the Model Training Partition for Three Task Types: Read Aloud, Picture Narration, and Listen-Speak

Subconstruct	Feature	Read aloud[a]	Picture narration[b]	Listen-speak[c]
Fluency	*wpsecutt*	0.59	0.58	0.59
	silpwd	0.53	–	–
	longpfreq	0.47	0.50	0.49
	tpsecutt	–	0.50	0.50
	wdpchk	–	0.47	0.47
Pronunciation	*conftimeavg*	0.53	–	–
	amscore	0.59	–	0.57
	L7	–	0.51	–
	phn_shift	–	0.44	0.33
Prosody	*stretimemean*	0.55	0.48	0.51
Grammar	*lmscore*	0.60	0.39	0.31
Accuracy	*cwpm*	0.66	–	–
Content	*factfreq_mean*	–	–	0.49

[a]$N = 882$. [b]$N = 875$. [c]$N = 2{,}048$.

Table 11 Agreement Statistics Comparing Scores From the First Human Rater (*H1*) to SpeechRater Scores (*M*) and Scores From the Second Human Rater (*H2*) Using Responses From the Model Evaluation Partition

	Pearson's *r*		Quadratic weighted kappa	
Item type	*H1*:*M*	*H1*:*H2*	*H1*:*M*	*H1*:*H2*
Read aloud[a]	.70	.70	.59	.70
Picture narration[b]	.62	.70	.52	.69
Listen-speak[c]	.66	.72	.58	.72

[a]$N = 684$. [b]$N = 668$. [c]$N = 1{,}568$.

Form-Level Results

The form-level results for the Speaking section were computed in a similar manner to the Writing section (as presented in the Form-Level Results subsection of this report): The scores from each of the five Speaking items were summed, and correlations were calculated between the summed human and machine scores. Again, only test takers who had valid scores (1–4) for all of the five Speaking items were included in the analysis ($N = 374$). This analysis resulted in a correlation of

Table 12 SpeechRater Features Included in the Scoring Model for the Read-Aloud Item That Address Each of the Rubric Descriptors for a High-Scoring Response in the Scoring Guide

Rubric descriptor	SpeechRater features
Reading is mostly fluid and intelligible	*conftimeavg*
Words are grouped in meaningful phrases with effective pauses; punctuation is marked appropriately throughout	*silpwd*, *longpfreq*
Intonation varies to match text provided	*stretimemean*
Speech is clear and distinct with only minor mispronunciations, substitutions, or omissions	*amscore*, *cwpm*
Rate of speech is mostly appropriate	*wpsecutt*

$r = .81$ between the human and machine form-level scores; the human – human correlation between the two sets of human scores on this dataset was $r = .89$.

Discussion

The experiments presented in this report indicate that current automated scoring capabilities achieve a level of performance that is lower than human – human agreement when applied to constructed responses from the TOEFL Junior Comprehensive assessment, but not by a large amount. Given that little additional system development was conducted to address the specific task types contained in the assessment, future improvements to the automated scoring features and models are likely to result in even higher levels of human – machine agreement. The form-level human – machine correlation for the four items in the Writing section was $r = .83$, compared to a human – human correlation of $r = .90$, and the form-level human – machine correlation for the five items in the Speaking section was $r = .81$, compared to a human – human correlation of $r = .89$.

The item that resulted in the best automated scoring performance across the Writing and Speaking sections, in terms of smallest degradation from human – human agreement, was the read-aloud item in the Speaking section. For this item, the human – machine correlation matched the human – human correlation of $r = .70$. For other items, however, the degradation in correlation ranged from .06 (for the listen-speak item in the Speaking section) to .13 (for the listen-write item in the Writing section). One explanation for this variability in performance by task type can be found by relating the features used in the scoring models to the constructs that are being assessed, as indicated in the scoring rubrics that are used by the human raters for each item type.[9] For example, the scoring rubrics for the read-aloud item in the Speaking section refer exclusively to a speaker's delivery and reading accuracy, and these are both areas that are represented by features used in the SpeechRater scoring model for this experiment. Table 12 lists the rubric descriptors that are used by the raters for a read-aloud response receiving the highest score of 4 and indicates the SpeechRater features from Table 10 included in the scoring model that address each of the rubric descriptors.

Conversely, the construct is covered less adequately for the other items, all of which exhibit degradations when the human – machine correlations are compared to the human – human correlations. For example, the rubric descriptors for a listen-write response receiving the highest score of 4 include the following content-related descriptions: "accurately provides all key points" and "provides support using relevant details from the talk." However, the generic e-rater models used to score the listen-write responses do not contain any features that assess the accuracy or appropriateness of the content of the response in relation to the stimulus materials; therefore these aspects of the scoring rubrics were not addressed at all by the e-rater scoring model used in these experiments, and it is not surprising that human – machine agreement using this set of features would not attain the level of human – human agreement. With the development of additional features to address the areas of the TOEFL Junior Writing and Speaking constructs that are not currently covered by the existing e-rater and SpeechRater features, it is expected that the gap between the human – machine and human – human agreement levels would decrease. This was demonstrated for the listen-speak items in the Speaking section by Xiong et al. (2013). That article described in detail a set of content-related features that were designed to address the rubric descriptors describing a high-scoring spoken response as "Content is full and appropriate to the task. Key information is conveyed coherently and accurately." When two of these content-related features were added to a baseline SpeechRater scoring model (consisting of the same features that were used in this experiment and that are listed in Table 10), the human – machine correlation

increased from $r = .62$ to $r = .66$. It is expected that similar improvements would be realized for the other item types with the addition of new features to address areas of the construct that are not currently covered by the existing e-rater and SpeechRater features.

Conclusion

In this report, we have presented the initial results of using existing automated scoring technology to score constructed responses from the Writing and Speaking sections of the TOEFL Junior Comprehensive assessment. For this work, two automated scoring capabilities, e-rater and SpeechRater, were used with only minor modifications to determine the current baseline performance that can be achieved for these task types. In addition, a new capability based on string matching was developed for the edit item in the Writing section, because existing automated scoring capabilities developed at ETS were not appropriate for this item. The results showed that human – machine correlations were lower than human – human correlations for all of the items (except for the read-aloud item in the Speaking section), but that the degradations were relatively small, ranging from .06 for the listen-speak item in the Speaking section to .13 for the listen-write item in the Writing section. Several additional steps could be taken in future work to improve on these baseline results, such as retraining the feature weights for the e-rater models for the email, opinion, and listen-write items and developing additional automated scoring features to address aspects of the construct that are not covered by the current feature sets in e-rater and SpeechRater.

Notes

1 Further information about the TOEFL Junior Comprehensive test as well as sample questions for each of the task types discussed in this report can be found in Rybinski (2012) and at http://toefljr.caltesting.org/sampletest/index.html.
2 The model ECXT0000 was used for feature extraction with e-rater v. 12.
3 Approximately 10 – 15% of the responses from the Writing and Speaking sections were excluded for this reason (except for the edit items in the Writing section, for which over 50% of the responses were excluded). In addition, because the data in this study were drawn from a pilot study, this percentage of nonscorable responses is substantially higher than in operational administrations.
4 Note that the Pearson correlations were mostly unaffected by rescaling because they already take into account differences in means and standard deviations. They would be identical if not for the twofold cross-validation procedure used for evaluating rescaled scores.
5 An earlier version of some of the contents of this section appeared in Evanini and Wang (2013).
6 This approach was also compared to the simpler approach of training Acoustic and Language models from scratch using only the TOEFL Junior data. While the performance of these models was better for the read-aloud items (by about 2 – 3%), their performance was consistently worse (by about 2 – 3%) for the picture narration and listen-speak item types. On the basis of these results, we decided to conduct subsequent experiments using the adapted models.
7 The development of content-based features for the picture narration items will be undertaken in a subsequent study.
8 Before calculating the feature correlations, outliers were truncated to a threshold value of 4 standard deviations from the mean, and some of the features were transformed by applying the inverse function, 1/x. In addition, all features with a negative correlation were multiplied by −1; Table 10 thus presents the absolute values of the feature correlations.
9 The scoring rubrics for the Speaking section are available at http://www.ets.org/s/toefl_junior/pdf/toefl_junior_comprehensive_speaking_scoring_guides.pdf and the scoring rubrics for the Writing section are available at http://www.ets.org/s/toefl_junior/pdf/toefl_junior_comprehensive_writing_scoring_guides.pdf.

References

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

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater automated essay scoring system. In M. D. Shermis & J. Burstein (Eds.), *Handbook for automated essay scoring* (pp. 55 – 67). New York, NY: Taylor and Francis.

CSLU. (2008). *Kids speech corpus*. Retrieved from http://www.cslu.ogi.edu/corpora/kids

Evanini, K., & Wang, X. (2013). Automated speech scoring for non-native middle school students with multiple task types. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*, Lyon, France, August 25 – 29 (pp. 2435 – 2439).

Hagen, A., Pellom, B., & Cole, R. (2003). Children's speech recognition with application to interactive books and tutors. In *Proceedings of the 2003 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, St. Thomas, US Virgin Islands (pp. 186 – 191).

Heilman, M., & Madnani, N. (2012). ETS: Discriminative edit models for paraphrase scoring. In *Proceedings of *SEM 2012: The first joint conference on lexical and computational semantics* (Vol. 1, pp. 529 – 535). Montreal, Canada: Association for Computational Linguistics.

Higgins, D., Xi, X., Zechner, K., & Williamson, D. M. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech and Language*, *25*, 282 – 306.

Huang, X., Acero, A., & Hon, H.-W. (2001). *Spoken language processing: A guide to theory, algorithm, and system development*. Mahwah, NJ: Prentice Hall.

Kantor, A., Cerňak, M., Havelka, J., Huber, S., Kleindienst, J., & Gonzalez, D. B. (2012). Reading companion: The technical and social design of an automated reading tutor. In *Proceedings of the 2012 Interspeech Workshop on Child, Computer, and Interaction*, Portland, OR (pp. 53 – 59).

Kazemzadeh, A., You, H., Iseli, M., Jones, B., Cui, X., Heritage, M., Price, P., Anderson, E., Narayanan, S., Alwan, A. (2005). TBALL data collection: The making of a young children's speech corpus. In P*roceedings of Interspeech 2005*, Lisbon, Portugal (pp. 1581 – 1584).

LDC. (1997). *The CMU kids corpus*. Retrieved from http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC97S63

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* (ETS Research Report No. RR-12-06). Princeton, NJ: Educational Testing Service.

Rybinski, P. (2012). *Introducing TOEFL Junior tests: Assessments for young learners of English*. Paper presented at the annual conference of the Association of Boarding Schools, Washington, DC. Retrieved from http://www.tabs.org/conference/2012/handouts/D13 Rybinski.pdf

Xiong, W., Evanini, K., Chen, L., & Zechner, K. (2013). Automated content scoring of spoken responses containing multiple parts with factual information. In *Proceedings of the 2013 Interspeech Workshop on Speech and Language Technology in Education (SLaTE)*, Grenoble, France (pp. 137 – 142).

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

Suggested citation:

Evanini, K., Heilman, M., Wang, X., & Blanchard, D. (2015). *Automated scoring for the TOEFL Junior® Comprehensive Writing and Speaking test* (Research Report No. RR-15-09). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12052

Action Editor: Beata Beigman Klebanov

Reviewers: Jidong Tao and Klaus Zechner