Article

LANGUAGE TESTING

Monitoring the performance of human and automated scores for spoken responses

Language Testing
2018, Vol. 35(1) 101–120

DOI: 10.1177/0265532216679451
journals.sagepub.com/home/ltj
SAGE

Zhen Wang
Educational Testing Service, USA

Klaus Zechner
Educational Testing Service, USA

Yu Sun
Educational Testing Service, USA

Abstract

As automated scoring systems for spoken responses are increasingly used in language assessments, testing organizations need to analyze their performance, as compared to human raters, across several dimensions, for example, on individual items or based on subgroups of test takers. In addition, there is a need in testing organizations to establish rigorous procedures for monitoring the performance of both human and automated scoring processes during operational administrations. This paper provides an overview of the automated speech scoring system SpeechRater[SM] and how to use charts and evaluation statistics to monitor and evaluate automated scores and human rater scores of spoken constructed responses.

Keywords

Automated speech scoring, language assessment, score monitoring, Shewhart control chart, human and machine scoring, reliability

Language testing organizations in the United States must routinely deal with large populations, especially for certain Asian, European, and Middle-Eastern countries. While having large populations is certainly not exclusive to language testing, constructed response (CR) item scoring, including essay scoring, and spoken response scoring, is

Corresponding author:
Zhen Wang, Educational Testing Service, 660 Rosedale Rd., Princeton, New Jersey, 08541, USA.
Email: jwang@ets.org

definitely an added complication for scoring. Human scoring has its limitations, such as severity/leniency, scale shrinkage, inconsistency, halo effect, and rater drift (Engelhard, 1994, 2002). Without careful monitoring (Wang & Yao, 2013), the human rater effects may substantially increase the bias in students' final scores. Human scoring is very labor intensive, time consuming, and expensive (Zhang, 2013). The importance of these language tests for relatively high-stakes decisions places a lot of pressure on the entire system to ensure accurate scoring and consistent ratings.

Automated scoring capabilities such as e-rater® and SpeechRater℠ have been developed and have the potential to provide solutions to some of the obvious shortcomings in human scoring (e.g., rater inconsistency, rater drift, and inefficiency). Bennett and Bejar (1998) indicated that automated scoring procedures allow for the scoring rules to be applied consistently. Automated scoring has some advantages including "fast scoring, constant availability of scoring, lower per unit costs, greater score consistency, reduced coordination efforts for human raters, and potential for a degree of performance specific feedback" (Ramineni, Trapani, Williamson, Davey, & Bridgeman, 2012). Therefore, some operational programs have already started using automated scoring to be used in combination with human scorers. The decision of these programs to use automated scoring was based on some research studies (Attali, 2007; Attali, Bridgeman, & Trapani, 2010; Attali & Burstein, 2006; Burstein & Chodorow, 1999; Chodorow & Burstein, 2004; Ramineni et al., 2012; Wang & von Davier, 2014).

SpeechRater is an automated scoring engine developed at Educational Testing Service (ETS) that has been used for a practice program since 2006. It consists of an automatic speech recognition (ASR) system, feature computation modules, and a multiple regression scoring model to predict scores for each spoken response (Zechner, Higgins, Xi, & Williamson, 2009).

The speaking construct that SpeechRater intends to measure is related to the notion of "communicative competence," as described by Bachman (1990) and Bachman and Palmer (1996). The construct is operationalized as a set of rubric dimensions (used for human scoring of spoken responses) that cover various aspects of spoken proficiency, including fluency; pronunciation; prosody; vocabulary range and sophistication; grammatical accuracy and complexity; content; and aspects of discourse.

In recent years, the coverage of the speaking construct has been substantially extended from its original focus on fluency, pronunciation, and prosody by adding features related to vocabulary, grammar, and content, among others (Chen & Zechner, 2011; Xie, Evanini, & Zechner, 2012; Yoon & Bhat, 2012; Yoon, Bhat, & Zechner, 2012).

There are a few studies on effective quality control procedures in these types of language testing settings (Bejar, 2011; Wang & von Davier, 2010; Williamson, Xi, & Breyer, 2012), including one related to human ratings and automated essay evaluations (Bridgeman, 2013). Wang and von Davier (2010) have proposed a set of statistics and a framework (examinee, test, prompt, and rater level) to monitor the quality of CR scoring. Bejar (2011) has provided a quality control and assurance framework for automated scoring. Williamson et al. (2012) have provided a framework for automated scoring evaluation and use of automated scoring and guidelines for implementation and maintenance in the context of constantly evolving technologies. Bridgeman (2013) summarized some of the procedures for monitoring and evaluating the quality of essay ratings using both

human and automatic scoring. Some researchers (Lee & von Davier, 2013; Luecht, 2010) have proposed using quality control techniques to monitor scoring, equating, and reporting of test scores. Lee and von Davier (2013) and Bejar (2011) also recommended using quality checking methods from other disciplines to monitor data routinely.

The main focus of the present research is to investigate whether human rater scores and the upgraded SpeechRater engine scores are comparable for each speaking item and administration. We used multiple methods including some quality control procedures such as traditional item analysis, agreement statistics, and graphical techniques to address this main research question. We also looked at scoring differences across different groups of test takers' native languages since fairness across subgroups of testing population is an important consideration when deploying any assessment. The deployment of automated scoring technology needs to be done with fairness in mind, that is, treating test takers of different subgroups in the same way.

Research questions

The major research question in this validity study is whether the newly upgraded SpeechRater engine produces scores that are comparable to human raters. The current study targeted the following three specific research questions:

1. Are the ratings from human raters and SpeechRater consistent in severity and variability?
2. Can SpeechRater be used to identify human raters who are very strict or very lenient?
3. Do SpeechRater scores of different language groups differ in the same way as the scores assigned by human raters?

General considerations on automated speech scoring

This section provides some general background for the research presented in our paper. We will first compare and contrast automated scoring with human scoring of constructed responses in general, and of spoken responses in particular, and then provide a brief overview on the history of automated speech scoring.

Automated versus human scoring of constructed response items

Constructed response (CR) items are typically elicited from test takers in order to provide evidence of a certain proficiency, (e.g., being able to write a concise essay on a given topic, or to summarize a video lecture by using speech). When human raters assign scores to such CRs, they usually follow a pre-defined rubric, a set of band descriptors for each score level, indicating the typical characteristics of a CR for a particular score along several dimensions of a construct. For example, a spoken response to a prompt asking the test taker to describe their last summer vacation may be evaluated based on dimensions of speech such as fluency, pronunciation, prosody, vocabulary usage, grammatical expression, and content. Human raters need to be trained and calibrated to ensure they

apply the rubric consistently across a large set of CRs in an assessment. However, in practice human ratings are not perfectly consistent, as discussed above.

Automated scoring systems usually identify various objectively measureable aspects of CRs, such as the rate of speech or the correctness of grammatical expressions in a spoken response, and then compute a score for a given CR by means of a weighted combination of these features, for example, by using a linear regression model that is trained on human-scored data. To the extent that the features computed by an automated scoring system are good representations of the dimensions of the construct that is associated with the response, such a scoring system is able to generate substantively meaningful scores. However, there has also been criticism that human raters are used as a "gold standard" to train automated scoring systems, even though it is known that they are far from perfect (Bennet & Bejar, 1988).

Obvious advantages of automated scoring systems are their perfect consistency, usually lower cost and faster scoring time, as well as the fact that it can be explained in detail what such a system is measuring, whereas this is much less obvious when using human raters for CR scoring.

Brief overview on the history of automated speech scoring

As speech recognition technology made substantial advances in the 1980s, researchers started to consider whether and to what extent this new technology could be used to evaluate the English proficiency of non-native speakers. The earliest systems focused on aspects of pronunciation and fluency (e.g., Bernstein et al., 1990; Cucchiarini et al., 1997a, 1997b, 2000a, 2000b; Franco et al., 2000b), and currently, the detection of pronunciation errors (e.g., for the purpose of language learning and tutoring systems), is still predominant in the field (e.g., EduSpeak, Franco et al., 2000a, 2010). Subsequently, this technology was used in various spoken language assessments with the emphasis on low-level aspects of speech, such as speaking rate, flow, hesitation, and pronunciation (Bernstein, 1999; Bernstein et al., 2000; Cucchiarini et al., 2002).

However, using only these types of low-level features that address delivery aspects of speech (such as fluency and pronunciation) does not capture the entire range of linguistic expression and representation that human raters will expect in spontaneous speech, as in the item responses in this study. These other dimensions include vocabulary range and sophistication, grammar accuracy and complexity, content appropriateness, progression and flow of ideas (discourse), and so on. The extraction of such higher-level language features has posed a challenge because (1) errors can be generated by the ASR system, and (2) it is difficult to devise natural language processing technologies to extract meaningful and accurate features, given the spontaneous nature and brevity of speech samples and the errors test takers may make in grammar or vocabulary choice.

Since the early 2000s, research and development work has been undertaken to automatically score not only predictable speech, but also more open-ended, spontaneous speech (Zechner et al., 2009). The automated speech scoring engine SpeechRater computes features in many diverse areas of the speaking construct, including fluency, pronunciation, prosody, vocabulary diversity, grammatical accuracy, and complexity, as well as content.

Method

Description of the data

The speaking section of the English language assessment used in this study elicit a total of 5.5 minutes of speech for a candidate: two independent items that ask test takers to talk for 45 seconds on a familiar topic (e.g., "Describe a person that you admire.") and four integrated items where reading and/or listening stimuli are presented first, and then the test taker has one minute each to respond to a prompt that is based on these stimuli.

Each response to a speaking item is scored holistically by a single trained human rater on a 4-point discrete scale of 1–4, with "4" indicating the highest proficiency level and "1" the lowest. The scores are assigned based on rubrics, one each for independent and integrated items. The rubrics describe the aspects of the speaking construct that are deemed most relevant for determining the speaking proficiency of test takers and thus guide human raters in their scoring decisions. Each score level has a description of prototypical observed speaking behavior in three main areas of spoken language: delivery (e.g., fluency and pronunciation), language use (vocabulary and grammar aspects), and topic development (e.g., progression of ideas and content relevance). Human raters usually get "batches" of responses for a particular prompt (rather than scoring, e.g., all the responses of one candidate). In addition, a random sample of about 10% of responses in each administration is scored by a second human rater for reliability control purposes. If the two scores disagree by more than one point, a third rater is asked to adjudicate the score. Finally, the six-item scores are aggregated and scaled for score reporting purposes.

Data were drawn from 10 administrations involving 110 countries in 2012–2013. Among the 10 administrations, half of them were mainly from the Western hemisphere and the other half were mainly from the Eastern hemisphere. We randomly sampled 1,100 test takers per administration. The speaking section of the English language assessment consists of six items. This yields a total of $10 \times 1{,}100 \times 6 = 66{,}000$ responses that were scored by the SpeechRater engine. We pulled the first human rater scores (H1-rater), including second human rater scores (H2-rater, if available), from a data repository.[1] (Note that "H1" and "H2" are logical labels for human raters; in actuality, "H1" scores and "H2" scores comprise scores from a large number of physical human raters.) As stated above, H2-rater scores were only available for 10% of the data, which is a random sample from the administrations selected for reliability purposes.

During the operational cycle, all human raters (both H1-rater and H2-rater) participated in a standardized training process before they were allowed to rate the speaking items. In this study, we focused on the comparison of the item scores between the H1-rater and SpeechRater. The H2-rater was from the same rater pool as the H1-rater, so there should not be any systematic differences between the H1-rater and the H2-rater. We also made comparisons between the scores assigned by the H1-rater and the H2-rater for the 10% reliability sample.

In addition to the main data set used for this study (66,000 spoken responses), we used 10,000 spoken responses to items in other forms of the same assessment to estimate the parameters of the linear regression model used by SpeechRater. A separate data set of 52,200 responses from the same assessment was used for training the parameters of the ASR system.

The SpeechRater system for scoring spoken responses

The SpeechRater system consists of the following four major components: (1) an ASR system that converts the test taker's response into a sequence of hypothesized words; (2) a component that computes a set of features related to the Delivery and Language Use constructs based on the ASR output and the speech signal; (3) a filtering model that identifies responses that should not be scored due to construct irrelevance or technical issues; and (4) a linear regression scoring model trained on a set of human-scored spoken responses.

To build the ASR component of SpeechRater, we used a large data set of 52,200 English language assessment responses, consisting of more than 800 hours of speech. This data set was used to train the acoustic model and language model used by the state-of-the-art ASR system licensed from an external vendor. This ASR system achieved a word error rate of around 30% on an independent English language assessment test set. For building the scoring model, we used 10,000 spoken responses (all double scored by human raters) from an English language assessment ("training set"). For evaluating the scoring model, we used 66,000 responses from the same English language assessment, but used different administrations and different test takers ("test set"). This corresponds to the data set described above.

Based on the training set, we selected a set of 13 features representing the English language assessment speaking construct to a large extent, with the exception of the sub-construct of "topic development." Feature selection criteria included the correlation with human rater scores, normality[2] (determined via Q–Q plots), inter-correlation between selected features, and construct representation. The feature set includes features measuring fluency (e.g., rate of speech; presence of filled pauses, repetitions, and repairs; distribution of pauses), pronunciation accuracy, prosody (distribution of stressed syllables), grammar accuracy, and diversity of vocabulary.

Item analyses

Certain standards have been used to guide the analyses of data for the building and evaluation of automated scoring models (Williamson, Xi, & Breyer, 2012). Classical test theory item analyses statistics and some graphics such as box plots and Shewhart charts were applied as part of the monitoring procedures of both human raters and SpeechRater. These item statistics provide general indications of item quality and possible item development problems and can further be compared across raters or administrations to help indicate potential scoring discrepancies. This study consists of the following statistics to address the research questions: (1) mean differences and SD ratio; (2) standardized mean difference; (3) quadratic weighted kappa; (4) Pearson correlation; and (5) human rater bias. For computing kappa statistics, the raw SpeechRater scores were first truncated and rounded for comparison against the integer human scores. For other statistics, truncated SpeechRater scores without rounding were used for comparison with human scores.

Table 1 displays a summary of the flagging criteria and conditions for evaluating SpeechRater model performance. The criteria used were similar to those recommended

Table 1. Flagging criterion and conditions for SpeechRater evaluation.

Flagging criterion	Flagging condition
Mean differences between human score and SpeechRater	Mean differences greater than 0.15 in absolute value
Standardized mean difference between human score and SpeechRater	Standardized mean difference greater than 0.15 in absolute value
Quadratic weighted kappa between human score and SpeechRater	Quadratic weighted kappa less than 0.70
Pearson correlation between human score and SpeechRater	Correlation less than 0.70
Bias between the human scores and SpeechRater score	Human raters whose bias is greater than 0.30[a] are considered to be lenient raters; those whose bias is less than −0.30 are considered to be strict raters

[a]See the bias formula in the "Human rater bias" subsection. The 0.30 rule was proposed and used in Wang and von Davier (2014).

by Williamson et al. (2012) and Wang and von Davier (2014). Williamson et al. (2012) recommended that the quadratic weighted kappa and product–moment correlation between automated and human scoring must be at least 0.70 (rounded normally) with the underlying rationale that approximately half of the variance in human scores is accounted for by e-rater. They also recommended that the standardized mean score difference between the human scores and the automated scores cannot exceed 0.15. This criterion is applied to avoid differential scaling between the automated scoring and human scoring. By following the same logic, we also used this criterion to flag items in terms of the mean difference between human rater scores and SpeechRater. In terms of human rater bias, we applied 0.30 as the cut-off value by following Wang and von Davier's (2014) recommendation in their e-rater and human rater comparison study. In their study, they used both Shewhart chart and 0.30 rule to identify outlier human raters against e-rater. Therefore this rule can be used in our study to detect outlier human raters against SpeechRater as well.

Results

Research question 1

To address the first research question, we compared human raters and SpeechRater^{SM} to see if they differ in terms of severity and variability using the following analyses including box plots, mean difference/SD ratio, standardized mean difference, correlation, kappa, and Shewhart charts.

Box plots. Figure 1 shows the box plots for the overall mean scores of the H1-rater, H2-rater, and *SpeechRater* for the six speaking items across 10 administrations. The box plot provides a useful depiction of a moderate to large distribution of numbers. The box or "fence" captures the "interquartile range" representing the center-most 50% of the

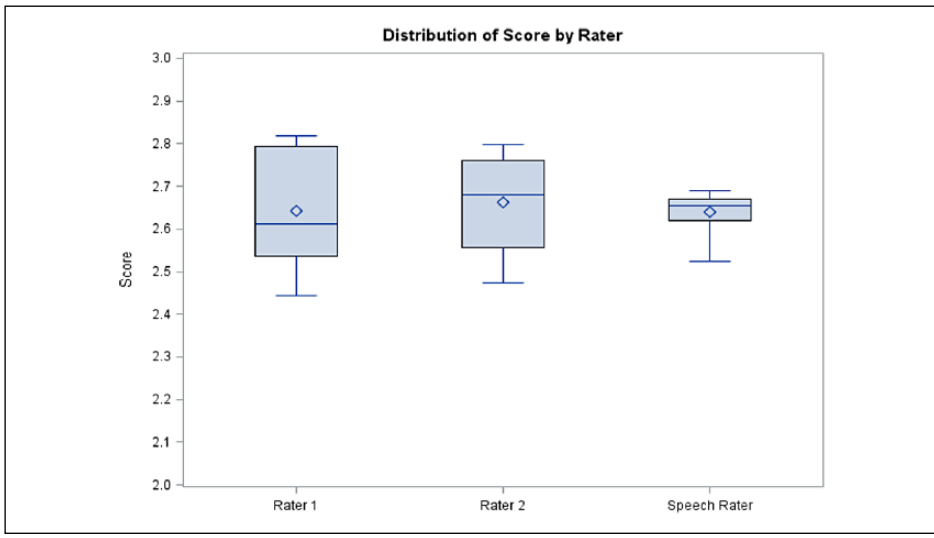

Figure 1. Box plots of overall mean ssssscores of the H1-rater, H2-rater and SpeechRater across administrations.

distribution of values (i.e., the scores ranging from the 25th to the 75th percentiles). The centerline in the box represents the median value and the diamond represents the mean. The "whiskers" extending from the box denote possible skewness in one or both tails of the distribution of values. In general, the overall means of the H1-rater, H2-rater, and SpeechRater scores across all administrations are close to each other.

Mean differences and SD ratio. Table 2 shows the results for the means, mean differences, and SD ratio of H1-rater/H2-rater and H1-rater/SpeechRater for each administration. No administration was found to have a H1-rater/H2-rater difference larger than 0.15; also, no administration was found to have a H1-rater/SpeechRater difference larger than 0.15. SD ratios for H1-rater/H2-rater range from 0.99 to 1.11, whereas SD ratios for H1-rater/SpeechRater range from 1.47 to 1.66, indicating that SpeechRater scores have a much smaller variance (about 0.24) than that of human raters (about 0.57). The closer the SD ratio to 1.00, the more similar are the variances of the two scores.

Standardized mean difference. Table 3 shows the results for the standardized mean differences between H1-rater/H2-rater and H1-rater/SpeechRater at the item level. Three items were found to have large effect sizes between the H1-rater and H2-rater; a total of 16 items were found to have large effect sizes (> 0.15) between the H1-rater and SpeechRater. For some items, SpeechRater gave higher scores than human raters while for other items, human raters gave higher scores.

Correlations. Table 4 shows the results for the correlations between the H1-rater and H2-rater, and the H1-rater/SpeechRater at the Administration level. The correlations

Table 2. Means and standard deviations for human rater scores and SpeechRater scores for each administration.

Administration ID	Count	H1-rater Mean (SD)	H2-rater Mean (SD)	SpeechRater Mean (SD)	H1 – H2 rater differences Mean (SD Ratio)	H1 – SpeechRater differences Mean (SD Ratio)
A	1070	2.44 (0.80)	2.48 (0.76)	2.53 (0.54)	−0.04 (1.05)	−0.09 (1.47)
B	1066	2.54 (0.77)	2.60 (0.74)	2.62 (0.53)	−0.06 (1.05)	−0.08 (1.47)
C	1071	2.71 (0.76)	2.74 (0.71)	2.67 (0.48)	−0.03 (1.07)	0.05 (1.60)
D	1073	2.68 (0.81)	2.75 (0.75)	2.64 (0.49)	−0.07 (1.07)	0.04 (1.64)
E	1079	2.81 (0.74)	2.80 (0.69)	2.69 (0.45)	0.01 (1.07)	0.12 (1.66)
F	1086	2.53 (0.75)	2.55 (0.67)	2.67 (0.49)	−0.01 (1.11)	−0.14 (1.51)
G	1094	2.82 (0.73)	2.79 (0.68)	2.67 (0.46)	0.03 (1.07)	0.15 (1.59)
H	1094	2.55 (0.73)	2.60 (0.68)	2.63 (0.48)	−0.09 (1.07)	−0.07 (1.51)
I	1093	2.80 (0.73)	2.77 (0.68)	2.69 (0.44)	0.03 (1.07)	0.11 (1.65)
J	1084	2.55 (0.73)	2.56 (0.74)	2.65 (0.50)	−0.01 (0.99)	−0.10 (1.56)

Note: The numbers in the column under the heading "Count" refer to data for the H1-rater and data for SpeechRater; the H2-rater is only 10% of the H1 data.

Table 3. Comparison of human rater and SpeechRater^{SM} scores for each speaking item using standardized mean difference test.

Administration ID	Item	N	H1- rater	H2-rater	SpeechRater^{SM}	Effect	
						Size 1	Size 2
A	3	1070	2.47	2.52	2.62	−0.05	**−0.16**
	6		2.38	2.46	2.57	−0.07	**−0.20**
B	6	1066	2.42	2.63	2.49	**−0.20**	−0.07
C	1	1071	2.86	2.96	2.58	−0.10	**0.31**
E	2	1079	2.87	2.86	2.73	0.01	**0.17**
	6		2.78	2.81	2.61	−0.03	**0.20**
F	4	1086	2.46	2.67	2.69	**−0.21**	**−0.26**
	5		2.56	2.64	2.73	−0.08	**−0.19**
	6		2.31	2.23	2.48	0.07	**−0.17**
G	1	1094	2.87	2.75	2.57	0.12	**0.36**
	2		2.97	2.95	2.68	0.02	**0.36**
	4		2.61	2.67	2.47	−0.06	**0.16**
H	3	1094	2.59	2.82	2.67	**−0.24**	−0.10
	5		2.45	2.6	2.61	−0.15	**−0.18**
I	1	1093	2.88	2.82	2.71	0.06	**0.22**
	2		2.93	2.86	2.67	0.07	**0.32**
J	3	1084	2.58	2.63	2.77	−0.05	**−0.22**
	4		2.45	2.46	2.66	−0.01	**−0.24**

Note: The numbers in the column under the heading "N", refer to data for the H1-rater and data for SpeechRater^{SM}. The H2-Rater is only 10% of the H1 data.

between the H1-rater and SpeechRater scores are slightly higher than those between the H1-rater and H2-rater, ranging from 0.69 to 0.81, with all of them being higher than 0.70

Table 4. Correlations between human rater and SpeechRater scores and two human rater scores.

Administration ID	H1–SP	H1–H2
A	0.81	0.76
B	0.81	0.76
C	0.78	0.73
D	0.78	0.75
E	0.76	0.71
F	0.80	0.70
G	0.76	0.71
H	0.76	0.69
I	0.69	0.69
J	0.81	0.73

Note: H2 is only 10% of the H1 data (1100). For the aggregate correlation, each student's mean scores of the six items for the H1-rater, H2-rater (if available) and SpeechRater are used for the calculation.

Table 5. Quadratic weighted kappa for the H1-rater and SpeechRater.

Administration ID	N	Quadratic weighted kappa[a]
A	1056	0.80
B	1053	0.80
C	1056	0.75
D	1057	0.78
E	1065	0.72
F	1076	0.80
G	1084	0.72
H	1078	0.77
I	1081	0.70
J	1074	0.79

[a]Quadratic weighted kappa is the mean of the six speaking items' kappa values for each administration.

except for one administration (0.69). Note importantly that the H2-rater was randomly selected and represents about 10% of the H1-rater/SpeechRater cases.

Kappa. The "quadratic weighted kappa" statistics for the H1-rater/SpeechRater were calculated (see Table 5). The quadratic weighted kappa by Administration ranged from 0.70 to 0.80, which met the 0.70 requirement.

Monitoring rating consistency using Shewhart charts. In order to ensure a high quality of scoring, a Shewhart chart was used as a technique for ensuring quality in a measurement process. A Shewhart chart is one of the statistical process control (SPC) techniques that have been widely used in industrial settings as tools for maintaining product quality (Vani, 1995). SPC charts have also been applied to educational measurement (Meijer,

2002; Omar, 2010; Veerkamp & Glas, 2000). This study illustrates how we applied the Shewhart SPC chart to monitor human rater and SpeechRater performance differences over time, an application which differs from all the other applications of Shewhart control charts in assessments (Gao, 2009; Lee & von Davier, 2013; Omar, 2010).

A Shewhart control chart has four elements (the last two elements are combined): points that represent a statistic (mean) of the measurement of a quality characteristic in samples taken from the process at different times, a center line that is drawn at the value of the mean of the statistic, and upper and lower control limits that indicate the threshold at which the process output is considered. Control limits are computed from the process standard deviation.

The upper control limits (UCL) and lower control limits (LCL) are

$$\text{UCL} = \text{mean of means} + k\left(\text{process standard deviation}\right) \quad (1)$$

$$\text{LCL} = \text{mean of means} - k\left(\text{process standard deviation}\right) \quad (2)$$

where k is the distance of the control limits from the baseline (mean of means), expressed in terms of the standard deviation unit.

In this study, control limits (lines) are drawn at six sigma (SDs) from the center line and represent the threshold where points above (or below) those lines are considered outliers. The Shewhart chart was plotted based on the rating difference; the H1-rater was subtracted from the SpeechRater score. If the human rating was more stringent, the rating difference resulted in a negative score and vice versa. The goal of perfect rating agreement between the human and SpeechRater is represented by a zero difference. Thus, a natural control difference of zero ($\mu_d = 0$) is used as the target for this SPC chart. Larger mean difference effects would be observed if human raters are either more or less lenient than the SpeechRater scoring.

Shewhart control charts were used for displaying the rating mean differences between the SpeechRater and the H1-raters based on a combination of 6 items in each of the 10 administrations. Shewhart control charts were also used to identify potential native language groups whose mean differences between human raters and SpeechRater are much larger than others across the 10 administrations.

We used the Shewhart control charts to identify potential "outlier" rating differences between the H1-rater and SpeechRater across 10 administrations. Figure 2 provides a Shewhart control chart of the overall mean rating differences by the H1-rater versus SpeechRater for the six speaking items by each administration. It is not difficult to see that some mean rating differences are above the six-sigma upper control limit while some others are below the six-sigma lower control limit. For example, in Figure 2, there were three administrations (Admin_E, Admin_G, Admin_I) where the mean rating differences were above the six-sigma upper control limit, and there were two administrations (Admin_F, Admin_J) where the mean rating differences were below the lower six-sigma control limit. These "out-of-control" mean differences indicate that the human raters rate differently from SpeechRater to some extent. One interesting finding is that those administrations that had mean differences close to or beyond the upper control limit are mainly from Western countries whereas those close to or below the lower limit are mainly from Eastern countries.

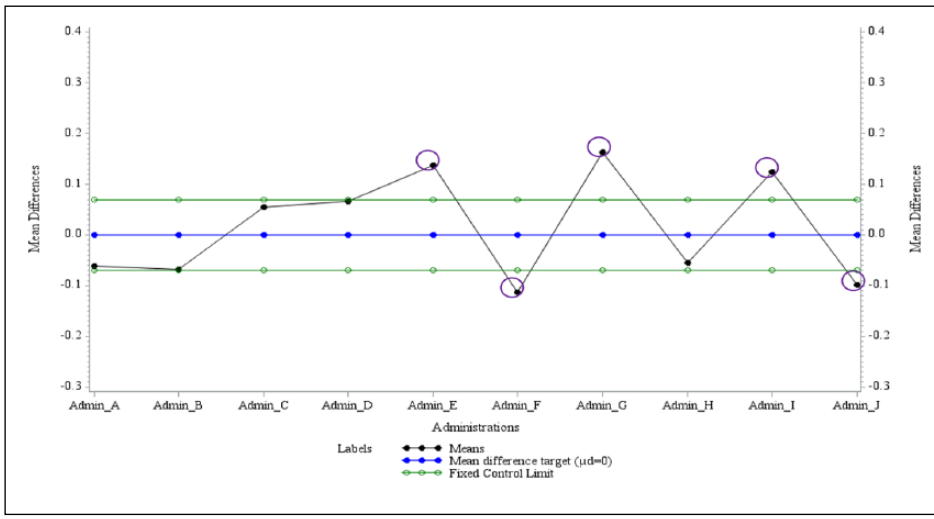

Figure 2. A Shewhart control chart for the difference between the H1-rater and SpeechRater in the mean score of six items by administration.

Research question 2

To address the second research question, we investigated whether SpeechRater[SM] could be used to check human raters' severity and leniency using the following human rater bias analyses.

Human rater bias. In this study, we chose to examine further the data in order to understand better the trends of human raters by comparing them with SpeechRater scores. To identify those human raters who were stricter or more lenient than other raters, SpeechRater scores were used to help identify the "outlier" human raters by looking at the differences between the ratings of the human raters and their corresponding SpeechRater scores.

The mean of these differences was labeled "Bias."[3]

$$\text{Bias} = \frac{1}{N_p}\sum_{i=1}^{N_p}(D_i), \tag{3}$$

where D_i is the difference between the H1-rater score and the SpeechRater score for item i, and N_p is the total number of items scored by a given rater. Raters for whom the bias was equal to or above an absolute value of 0.30 were labeled as potential "outlier" raters.

In order to address research question 2, we used SpeechRater scores to help identify the "outlier" human raters by looking at the differences between the ratings of the human raters and their corresponding SpeechRater scores. The bias between the ratings of the

Table 6. List of human rater outliers versus SpeechRater[SM] for each item.

	N of bias > 0.3 or bias < −0.3	Rater severity
Item 1	3	Harsh
	22	Lenient
Item 2	4	Harsh
	14	Lenient
Item 3	4	Harsh
	1	Lenient
Item 4	6	Harsh
	2	Lenient
Item 5	4	Harsh
	4	Lenient
Item 6	8	Harsh
	5	Lenient

human raters and their corresponding SpeechRater scores was calculated for each human rater. Raters who scored a reasonable number of spoken responses (N > 20) and whose absolute bias value was greater than or equal to 0.30 (an arbitrary cut-off value from our previous e-rater study; see Wang & von Davier, 2014) were listed as potential outlier raters in the study. For the first speaking items (the first items in each administration were combined across administrations, see Table 6) we found 3 harsh and 22 lenient H1-raters out of 210 who scored at least 20 items when comparing their ratings to SpeechRater scores. For the second items across administrations, we found 4 harsh and 14 lenient human raters out of a total of 211 raters; for the third items, we found 4 harsh and only 1 lenient human rater out of a total of 204 raters; for the fourth items, we found 6 harsh and 2 lenient human raters out of 213 raters across administrations; for the fifth items, we found 4 harsh and 4 lenient raters out of 206 raters; for the sixth items, we found 8 harsh and 5 lenient raters out of 203 raters.

Research question 3

To address the third research question, we examined weather SpeechRater[SM] scores of different language groups differ in the same way as the scores assigned by human raters using the Shewhart chart and kappa statistics.

Shewhart chart. The performance differences between the H1-rater and SpeechRater in the mean scores of the 6 items across all the 10 administrations were plotted for the 16 largest language groups[4] with more than 100 students within each group (see Figure 3). Fluctuating control limit was also plotted based on the different standard deviation of each operational administration. Outliers of mean score differences were observed for the Chinese language group (with higher SpeechRater scores) and for the German, English, French, Italian, Portuguese, and Russian language groups (with lower SpeechRater scores). It seems that SpeechRater ratings were close to human rater's ratings for some

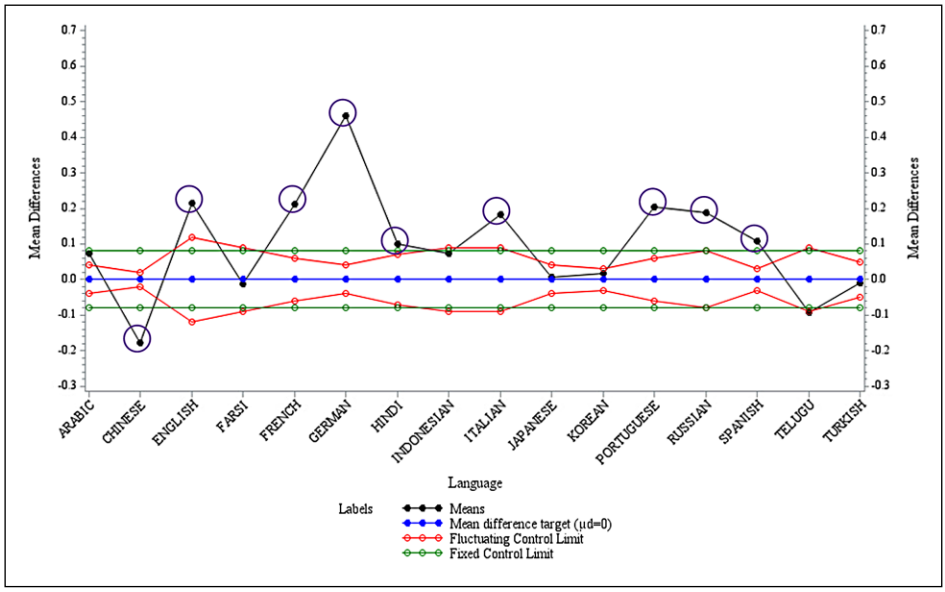

Figure 3. A Shewhart control chart for the difference between the H1-rater and SpeechRater in the mean score of six items by native language group.

of the native language groups, but not for the others (e.g., Chinese, German, English, French, Italian, Portuguese, and Russian). For German, the SpeechRater scores seem to be much lower than the corresponding human rater scores when they are compared to the other language groups.

Kappa. The mean "quadratic weighted kappa" of H1-rater/SpeechRater was calculated for each of the largest 16 native language groups. Overall, the Japanese group had the highest kappa values (0.86) and the German group had the lowest kappa values (0.39).

Summary

The quality control of human rater and SpeechRater scores is essential. This paper summarizes some important considerations and illustrates the application of specific statistical and graphical procedures that can help monitor both human and SpeechRater performances. Below, we summarize the findings from our research and suggest recommendations for operational practice.

As for the first research question, we found that although the overall correlations between the H1-rater scores and SpeechRater scores are similar to the correlations between the H1-rater and H2-rater scores, at the item level, we observed more differences between scores from the H1-rater and SpeechRater than between scores from the H1-rater and H2-rater. Items with large discrepancies between the H1-rater and SpeechRater need further investigation. For example, we can use the H2-rater to identify

Table 7. Mean quadratic kappa of H1-rater/SpeechRater for each language group.

Language group	N[a]	Quadratic weighted kappa
Arabic	81	0.80
Chinese	335	0.72
English	14	0.49
Farsi	12	0.59
French	65	0.67
German	79	0.39
Hindi	14	0.52
Indonesian	12	0.59
Italian	18	0.52
Japanese	115	0.86
Korean	124	0.80
Portuguese	54	0.74
Russian	14	0.61
Spanish	207	0.70
Telugu	13	0.47
Turkish	79	0.81

[a]N is the mean sample size per administration for each language group. Quadratic weighted kappa is the mean of 10 administrations' kappa values for each language group.

whether a SpeechRater score is an outlier if the H1-rater and H2-rater scores are close to each other, or we can use the H2-rater to identify whether the H1-rater score is an outlier if its SpeechRater score and H2-rater score are close to each other.

At the operational administration level, the H1-rater and SpeechRater agree well if we look at the means and correlations, indicating that the scores from the two scoring methods are comparable. The results from the kappa statistics (quadratic weighted kappa) met our expectation (>0.70). Additionally, it appears that SpeechRater produced slightly higher scores than human raters for Eastern countries and gave slightly lower scores for Western countries, whereas human raters (both the H1-rater and H2-rater) did not have such a pattern.

Some statistics and charts such as means, SDs, box plots, correlations, and standardized mean differences are very effective in detecting outlier speaking items, human raters, and SpeechRater scores. We recommend they be used jointly while monitoring human rater and SpeechRater scores at the examinee level, item level, and test score level throughout the operational cycle.

As for research question 2, we found a few lenient and a few harsh human raters by using SpeechRater scores since automated scoring is more consistent than human scoring. Such bias analyses can help identify human raters who tend to give harsh or lenient scores. In terms of items 1 and 2 (the two independent items in the test), we identified more lenient human raters than for items 3–6 (integrated items). This might indicate that human raters give more credit to content aspects of these independent items compared to SpeechRater, whose features are only related to the other two construct dimensions, that

is, delivery and language use. Due to the limited sample sizes, the results of these rater bias analyses need to be replicated and confirmed by using larger samples (e.g., $N > 100$).

As for the third research question, by looking at the H1-rater and SpeechRater score mean differences for each of the 16 largest native language groups, the largest differences occurred in the Chinese and German groups. SpeechRater mean scores were higher than the H1-rater scores for Chinese and vice versa for German. One reason for these discrepancies could be the difference in overall mean scores for different native languages. For example, the H1 mean score for German test takers in our data is 3.31, whereas the overall mean score of all test takers is 2.64. In addition, since the scoring model used by SpeechRater only evaluates a subset of the speaking construct, and in particular does not use features related to content and discourse, it may have a positive bias for first language test takers whose responses lack in these dimensions of topic development and vice versa.

Overall, scores for Japanese and Korean speakers have higher kappa values than for other native language groups such as German. More investigations need to be conducted to identify the major reasons regarding why some native language groups have lower kappa values.

In general, we further found that some statistics and charts used in the study, such as standardized mean differences, correlations, kappa values, box plots and Shewhart charts are very effective in detecting the differences between human and automated scoring at both item and administration level.

Discussion and limitations

Generally speaking, this study addresses several important research questions that need to be investigated before an automated speech scoring system such as SpeechRater can be implemented for the operational scoring of English language assessment speaking items. There are some systematic patterns in the score differences between human raters and SpeechRater. Figure 2 seems to suggest a mixture in the mean differences across administrations, which is likely a result of regional effects (East or West). SpeechRater tended to produce comparable mean scores across operational administrations regardless of the region, while the human raters' scores varied substantially across administrations in different regions. These patterns may relate to how the administrations were scored. Thus, for future analyses, it would be worth exploring the region of individual administrations when introducing the data and examining the results by region. How different item types (independent vs. integrated) are related to the differences and correlations between scores is also worth exploring.

When the standard deviations of SpeechRater were compared with those of human raters, the former were significantly lower, likely related to the central tendency of the multiple regression scoring approach used by SpeechRater. Also, the score distributions of human raters and SpeechRater appear to be somewhat different. A generalizability study may need to be conducted to investigate different sources of error between the two scoring modes. Furthermore, SpeechRater scoring seems to exhibit a small bias against several specific language groups (e.g., German), and in favor of the Chinese group. Reasons could include the following: (1) effects of the central tendency of the linear

regression scoring model used by SpeechRater, which reduces the number of predicted scores at the extremes of the scoring scale (having a stronger effect on languages with a higher overall mean score, such as German); (2) differences in considered construct aspects between SpeechRater and human raters in conjunction with the extent to which speakers with different first languages exhibit differential speaking profiles; (3) effects related to the score and language distribution in the training sample (e.g., bias towards certain L1 score distributions); and (4) effects related to differential functioning of certain SpeechRater features for different native languages.

When using automated scoring operationally, language bias can be reduced by using a contributory scoring approach, where both automated and human scores contribute jointly to a final item score.

Biased human raters were also identified by using SpeechRater scores. Since human raters' scores are based on three areas (delivery, language use, and topic development), whereas SpeechRater scores are based only on delivery and language use, differences between the two types of scores would probably exist even if there were no rater effects.

Another limitation of automated speech scoring is related to imperfect automatic speech recognition: the system used in the current study has a word error rate of around 30%, which is quite substantial. ASR systems for native speech, in contrast, can obtain much smaller word error rates (much less than 5%), but recognizing non-native speech from a large variety of first language backgrounds and speaking proficiency levels remains a challenge for current ASR technology. Still, in recent years, the availability of new algorithms (in particular, Deep Neural Networks), more powerful hardware (Graphics Processing Units), in combination with substantially larger data sets for ASR training, has led to a noticeable reduction in word error rate; for example, Cheng, Chen, and Metallinou (2015) report a word error rate of 23% on open-ended non-native speech.

In this study, we only applied one statistical approach to identify biased human raters owing to the lack of a human rater study related to our speech data. We therefore recommend additional studies and analyses for future work in order to confirm the findings related to the human rater bias in this study.

In this context, we also want to stress that the automated speech scoring engine SpeechRater is still evolving in terms of improved speech recognition, and that additional features, covering a more extended subset of the Speaking construct, are currently under development and are planned to be used in future scoring models. Moreover, we are also exploring alternative approaches to the standard linear regression scoring model currently used by SpeechRater, which may lead to improvements in its scoring performance.

Finally, as a result of this study, we recommend using multiple procedures (statistics and plots) to identify outlier items/administrations and human raters. Note that these recommendations are in some sense preliminary, given that our operational experience with SpeechRater is relatively limited. This study responds to the need for statistical analyses to provide a consistent and standardized approach for monitoring the quality of automated speech scoring over time and across programs (Bejar, 2011; Bridgeman, 2013; Ramineni et al., 2012; Wang & von Davier, 2014). Since studies on effective quality control procedures in these types of analyses along with automated scoring including SpeechRater are lacking, we believe that the use of the statistical analyses in this study is

a useful way to identify issues in both human and automated scoring. Graphics (e.g., Shewhart control charts, and box plots) should be an integral part of the quality control system because they present informative multivariate views of the data and can highlight general trends and correspondence with expectations.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Spoken responses with special exception scores not in the score range above were excluded from this study.
2. For linear regression models it is desirable to use normally distributed independent variables.
3. Also used in Way, Vickers, and Nichols' (2008) study to investigate the effects of different training and scoring approaches to human CR scoring.
4. Chinese, $N = 3{,}253$; Korean, $N = 1162$, Spanish, $N = 1003$; Japanese, $N = 666$; Arabic, $N = 610$; German, $N = 394$; Turkish, $N = 343$; French, $N = 312$; Portuguese, $N = 204$; Hindi, $N = 139$; English, $N = 136$; Russian, $N = 134$; Farsi. $N = 121$; Telugu, $N = 117$; Italian, $N = 106$; Indonesian, $N = 101$.

References

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

Attali, Y. (2007). *Construct validity of e-rater® in scoring L essays* (Research Report No. RR−07–21). Princeton, NJ: Educational Testing Service.

Attali, Y., Bridgeman, B., & Trapani, C. (2010). Performance of a generic approach in automated scoring. *Journal of Technology, Learning, and Assessment*, *10*(3). Retrieved from www.jtla.org

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Bejar, I. I. (2011). A validity-based approach to quality control and assurance of automated scoring. *Assessment in Education: Principles, Policy & Practice*, *18*(3), 319–341.

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice*, *17*(4), 9–17.

Bernstein, J., Cohen, M., Murveit, H., Rtischev, D., & Weintraub, M. (1990). Automatic evaluation and training in English pronunciation. *Proceedings of the ICSLP-90: 1990 International Conference on Spoken Language Processing* (pp. 1185–1188). Kobe, Japan.

Bernstein, J., DeJong, J., Pisoni, D., & Townshend, B. (2000). Two experiments in automatic scoring of spoken language proficiency. *Proceedings of InSTILL2000*. Dundee, UK.

Bridgeman, B. (2013). Human ratings and automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.) *Handbook of automated essay evaluation: Current applications and new directions* (pp. 221–232). New York: Routledge.

Burstein, J., & Chodorow, M. (1999). Automated essay scoring for nonnative English speakers. In M. Broman Olsen (Ed.), *Computer mediated language assessment and evaluation in natural language processing* (pp. 68–75). Morristown, NJ: Association for Computational Linguistics.

Chen, M., & Zechner, K. (2011). Computing and evaluating syntactic complexity features for automated scoring of spontaneous non–native speech. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics and the Human Language Technologies Conference (ACL–HLT–2011)*, Portland, OR, June.

Cheng, J., Chen, X., & Metallinou, A. (2015). Deep neural network acoustic models for spoken assessment applications. *Speech Communication*, *73*, 14–27.

Chodorow, M., & Burstein, J. (2004). *Beyond essay length: Evaluating e-rater® 's performance on essays* (Research Report No. RR−04−04). Princeton, NJ: Educational Testing Service.

Cucchiarini, C., Strik, H., & Boves, L. (1997a). Automatic evaluation of Dutch pronunciation by using speech recognition technology. Paper presented at the meeting of IEEE Automatic Speech Recognition and Understanding Workshop, Santa Barbara, CA.

Cucchiarini, C., Strik, H., & Boves, L. (1997b). Using speech recognition technology to assess foreign speakers' pronunciation of Dutch. Paper presented at the Third International Symposium on the Acquisition of Second Language Speech: NEW SOUNDS 97, Klagenfurt, Austria.

Cucchiarini, C., Strik, H., & Boves, L. (2000a). Different aspects of expert pronunciation: Quality ratings and their relation to scores produced by speech recognition algorithms. *Speech Communication*, *30*(2–3), 109–119.

Cucchiarini, C., Strik, H., & Boves, L. (2000b). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *Journal of the Acoustical Society of America*, *107*, 989–999.

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*(6), 2862–2873.

Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*, 93–112.

Engelhard, G., Jr. (2002). Monitoring raters in performance assessments. *Large–scale assessment programs for all examinees: Validity, technical adequacy, and implementation* (pp. 261–287). Mahwah, NJ: Lawrence Erlbaum.

Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R., & Butzberger, J. (2000a). The SRI EduSpeak system: recognition and pronunciation scoring for language learning. *Proceedings of Intelligent Speech Technology in Language Learning, InSTiLL-2000*. Dundee, UK.

Franco, H., Bratt, H., Rossier, R., Gadde, V. R., Shriberg, E., Abrash, V., & Precoda, K. (2010). EduSpeak®: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing*, *27*(3), 401–418.

Franco, H., Neumeyer, L., Digalakis, V., & Ronen, O. (2000b). Combination of machine scores for automatic grading of pronunciation quality. *Speech Communication*, *30*, 121–130.

Gao, R. (2009). Detect cheating using statistical control methods for computer based CLEP examinations with item exposure risks. Unpublished manuscript.

Lee, Y.-H., & von Davier, A. A. (2013). Monitoring scale scores over time via quality control charts, model-based approaches, and time series techniques. *Psychometrika*, *78*(3), 557–575.

Luecht, R. M. (2010). Some small sample statistical quality control procedures for constructed response scoring in Language Testing. Paper presented at the National Council on Measurement in Education, Denver, CO.

Meijer, R. R. (2002). Outlier detection in high-stakes certification testing. *Journal of Educational Measurement*, *39*, 219–233.

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, *46*, 371–389.

Omar, M. H. (2010). Statistical process control charts for measuring and monitoring temporal consistency of ratings. *Journal of Educational Measurement*, *47*(1), 18–35.

Ramineni, C., Trapani, C., Williamson, D. M., Davey, T., & Bridgeman, B. (2012). *Evaluation of e-rater® for the GRE issue and argument prompts* (Research Report No. RR–12−06). Princeton, NJ: Educational Testing Service.

Vani, J. (1995). *McGraw-Hill's certified quality engineer examination guide*. NewYork: McGraw-Hill.

Veerkamp, W., & Glas, C. A. W. (2000). Detection of known items in adaptive testing with a statistical quality control method. *Journal of Behavioural and Educational Statistics*, *25*, 373–389.

Wang, Z., & von Davier, A. A. (2010). Proposed procedures to monitor the performance of the human- & electronic ratings for all programs. Unpublished manuscript.

Wang, Z., & Yao, L. (2013). *Investigation of the effects of scoring designs and rater severity on students' ability estimation using different rater models* (Research Report. No. RR–13–23). Princeton, NJ: Educational Testing Service.

Wang, Z., & von Davier, A. A. (2014). *Monitoring of scoring using the e-rater automated scoring system and human raters on a writing test* (Research Report No. RR–14−04). Princeton, NJ: Educational Testing Service.

Way, W. D., Vickers, D., & Nichols, P. (2008). Effects of different training and scoring approaches on human constructed response scoring. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13.

Xie, S., Evanini, K., & Zechner, K. (2012). Exploring content features for automated speech scoring. *Proceedings of NAACL–HLT 2012*. Montreal, Canada.

Yoon, S.-Y., & Bhat, S. (2012). Assessment of ESL learners' syntactic competence based on similarity measures. *Proceedings of EMNLP–CoNLL 2012*. Jeju, Korea.

Yoon, S.-Y., Bhat, S., & Zechner, K. (2012). Vocabulary profile as a measure of vocabulary sophistication. *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications, NAACL–HLT 2012*. Montreal, Canada.

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*(10), October.

Zhang, M. (2013, March). *Contrasting automated and human scoring of essays*. (R & D Connections, No. 21). Princeton, NJ: Educational Testing Service.