Psychological Test and Assessment Modeling, Volume 60, 2018 (1), 81-100

Comparison of human rater and automated scoring of test takers' speaking ability and classification using Item Response Theory

Zhen Wang[1] *& Yu Sun*[2]

Abstract

Automated scoring has been developed and has the potential to provide solutions to some of the obvious shortcomings in human scoring. In this study, we investigated whether SpeechRater^{SM} and a series of combined SpeechRater^{SM} and human scores were comparable to human scores for an English language assessment speaking test. We found that there were some systematic patterns in the five tested scenarios based on item response theory.

Keywords: SpeechRater^{SM}, human scoring, item response theory, ability estimation and classification

[1] *Correspondence concerning this article should be addressed to:* Zhen (Jane) Wang, Educational Testing Service, Senior Psychometrician, Psychometrics, Statistics and Data Sciences (PSDS), Rosedale-Anrig, Princeton, NJ, U.S.A; email: jwang@ets.org.

[2] Senior Psychometric Analyst, Psychometrics, Statistics and Data Sciences (PSDS), Rosedale-Anrig, Princeton, NJ,

Language testing organizations in the United States routinely deal with large test taker samples, especially for certain Asian, European, and mid–eastern countries. For example, it is not unusual to have more than a million test takers worldwide taking a certain test in a given year, with each test taker responding to six items producing a total of more than six million responses. While having large samples is certainly not exclusive to language testing, constructed response item scoring, including essay scoring and spoken response scoring, is definitely an added complication for scoring.

Human scoring must be closely monitored within each administration and across administrations to ensure the quality and consistency of the human ratings. The effects of differences between human raters may substantially increase the bias in test takers' final scores without careful monitoring (Wang & Yao, 2013). This makes human scoring very labor intensive, time consuming and expensive (Zhang, 2013). The importance of these language tests for relatively high stake decisions places a lot of pressure on the entire system to ensure accurate scoring and consistent ratings on demand.

Automated scoring has been developed and has the potential to provide solutions to some of the obvious shortcomings in human scoring (e.g., rater inconsistency; rater drift; 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). Recently, more and more operational scoring programs have started using automated scoring to augment human scorers. In the literature, we found some studies on the comparison of human and automated scoring conducted in recent years (Attali, Bridgeman & Trapani, 2010; Attali & Burstein, 2006; Chodorow & Burstein, 2004; Laundauer, Laham, & Foltz, 2003; Nichols, 2005; Ramineni et al., 2012; Streeter, Bernstein, Foltz, & Deland, 2011; Wang & von Davier, 2014; Wang, Zechner & Sun, 2016; Williamson, Xi, & Breyer, 2012). Wang et al. (2016) conducted a comprehensive analysis of SpeechRater^{SM} and human scoring at item, rater and form levels based on the speaking data from 10 administrations. For automated scoring, there is no need to examine changes in rater severity. Automated scoring is not prone to these types of changes. However, if an engine update occurs, the scores need to be investigated for comparability against human rater scores. As Attali (2013) pointed out, the relation of automated scores to human scores provides important information about the validity of automated scores.

In terms of operational scoring, several researchers have called for the need to study how to combine human and machine scores (Attali, 2013; Zhang, Breyer, & Lorenz, 2013), which is lacking in the current research literature. Due to the shortcomings of human rater and automated scoring, the argument of combining the two is appealing. The shift to the combination of the two allows us to move beyond using human scoring as the default "gold standard" and become less concerned about the construct underrepresentation of the automated scoring.

The primary purpose of the present research is to investigate whether human raters and SpeechRater^{SM} scores and different combinations of the two are comparable for test takers' final scoring, ability estimation, and classification.

Research questions

The major research question in this study is whether SpeechRater^{SM} and a series of combined SpeechRater^{SM} and human scores are comparable to human scores for an English language assessment speaking test. The current study targeted the following specific research questions:

Do the scores from SpeechRater^{SM} only, Human Raters only, or different combinations of the two result in parallel[1], tau–equivalent[2], or congeneric[3] test scores for the final scoring of the speaking test?

Do the scores from SpeechRater^{SM} only, Human Raters only, or the different combinations of the two result in similar IRT parameter estimates (item difficulty and discrimination) and test takers' ability estimates of the population values and classifications?

Method

Data

The speaking section of the English language assessment used in this study elicits 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 in which reading and/or listening stimuli are presented first, and then the test taker has one minute to respond to each prompt 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 (fluency and pronunciation), language use (vocabulary and grammar aspects), and topic development (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 1100 test takers per administration. The speaking section of the English language assessment consists of six items. This yields a total of 10 × 1100 × 6 = 66,000 responses that were scored by the SpeechRater^{SM} engine. We pulled the first human rater scores (H1-rater), including second human rater scores (H2-rater, if available), from a data repository (note that “H1” and “H2” are logical labels for human raters; in actuality, each of “H1” scores and “H2” scores comprise scores from a large number of 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^{SM}. The H2-rater was from the same rater pool as the H1-rater, so there should not be any systematic differences between the results from 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^{SM} (discussed in next section). A separate data set of 52,200 responses from the same assessment was used for training the parameters of the ASR system.

SpeechRater^{SM}

SpeechRater^{SM} is an automated scoring engine developed at ETS that has been used in the TOEFL Practice Online program since 2006. It consists of an automatic speech recognizer (ASR system), feature computation modules, filtering model, a multiple regression scoring model to predict scores produced by human scorers for each spoken response (Zechner, Higgins, Xi, & Williamson, 2009). The filtering model, used to filter out non-scorable responses, is an important component of SpeechRater^{SM} (Higgins, Xi, Zechner, & Williamson, 2011). 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).

Data analyses

Classical analyses

Score comparisons using classical analysis were made based on different scoring scenarios. Analyses were conducted to answer the question of whether scoring based on the SpeechRater^{SM} only, human raters only, or combinations of the two would result in parallel, tau-equivalent, or congeneric test scores for the final scoring of the speaking test.

Different scoring scenarios tested include the following five scores:

S1. Human rater (H1-rater) only;

S2. SpeechRater^{SM} only;

S3. Total of SpeechRater^{SM} and human rater;

S4. Lower weighting (1/3) on human and higher weighting (2/3) on SpeechRater^{SM};

S5. Higher weighting (2/3) on human and lower weighting (1/3) on SpeechRater^{SM}.

The above scenarios can also be expressed as different weights in the weighted sum of the two ratings (human rater and SpeechRater^{SM}, respectively):

S1. (1, 0);

S2. (0, 1);

S3. (1/2, 1/2);

S4. (1/3, 2/3);

S5. (2/3, 1/3).

The one-factor measurement model was initially tested within the linear structural equation framework of Jöreskog and Sörbom (1984). A single-factor measurement model was fit to the data, in which all the five speaking scores were set to load on the same common latent variable representing the English speaking ability construct.

The five different scoring scenarios (S1 – S5 above) for the final speaking scores, which are the aggregated score across all six speaking items, were considered as five measures (S1 – S5), each measure was based on the same speaking items in the test. Reliability was calculated for each of the scoring scenarios. Correlations were also calculated among the five measures based on the different scoring scenarios. Parallel measures are the most restrictive model, assuming equal true score variances and equal error variances. Tau-equivalent measures have equal true score variances, but possibly different error variances whereas congeneric models allow true score variances as well as error variances to differ. Covariance matrices were analyzed and weighted least squares estimates of the parameters were obtained. If the fit of the model to the data becomes worse as the model is made more restrictive, then the constraints are not plausible for the data (Maruyama, 1998).

Three fit indices were used to evaluate the model-data fit, root mean squared error approximation (RMSEA), comparative fit index (CFI), and the normed fit index (NFI). Some empirical guidelines were followed when evaluating these fit indices: a model with an RMSEA value below .08, a CFI value above .90, a NFI value above .90, were considered to be an acceptable fit; a model with an RMSEA value below .05 and a CFI (and NFI) value above .95 were considered a good fit (Browne & Cudeck, 1993; Hooper, Coughlan, & Mullen, 2008; & Hu & Bentler, 1999).

IRT analyses

Generalized partial credit model.

PARSCALE (Muraki & Bock, 1997) was used to analyze the speaking tests based on the Generalized Partial Credit Model (GPCM, Muraki, 1992). The normal distribution was used as the prior distribution of ability. The method employed in the calibration phase of PARSCALE is that of random-effects estimation through marginal maximum likelihood. The random-effects solution employs the EM method (estimation and maximization) of solving the marginal likelihood equations. We set the number of EM cycles to be 50, and the number of quadrature points as 41 in the EM and Newton estimation. The scale parameter was set at 1.7 in the PARSCALE run.

Each speaking item was analyzed with IRT models with the following scoring designs: a) with human rater only; b) with SpeechRater^{SM} only; c) with the sum of human and SpeechRater^{SM}; d) with different weighting: 1/3 of human and 2/3 of SpeechRater^{SM}; and e) with different weighting: 2/3 of human and 1/3 of SpeechRater^{SM}. A key goal of this study was to compare test takers' ability estimates, and IRT parameters from different scoring designs. To compare test takers' ability estimates and IRT parameters, we used squared bias and the mean squared error.

Root mean squared error (RMSE), absolute mean bias (ABSBIAS), and bias were computed for all parameters and were used to examine the IRT parameter estimate differences, such as ability, difficulty, and discrimination. For example, let f_{ref} be the ability parameter based on the scoring with human rater 1 only (reference) and let f_j be the SpeechRater^{SM} only or the combination of the two (see S1 to S5 above), then

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(f_j - f_{ref})^2} \tag{1}$$

where n is the number of test takers. Here f represents ability parameter estimates.

The bias and absolute mean bias are defined as the following:

$$BIAS = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(f_j - f_{ref})} \tag{2}$$

$$ABSBIAS = \frac{1}{n}\sum_{j=1}^{n}\left|(f_j - f_{ref})\right| \tag{3}$$

Bias and absolute mean bias were used to assess the parameter estimate differences. The squared differences and squared bias were also calculated and plotted to show results of parameter estimate differences and comparison among different scoring scenarios.

Results

Research question 1

Do the scores obtained from SpeechRater^{SM} only, Human Raters only, or the different combinations of the two result in parallel, tau-equivalent, or congeneric test scores for the final scoring of the speaking test?

Correlations.

Correlations were calculated (see Table 1) among the five scores (S1 – S5) based on different scoring scenarios (H1-rater only; SpeechRater^{SM} only; sum of SpeechRater^{SM} and H1-rater; lower weighting on human and higher weighting on SpeechRater^{SM}; higher weighting on human and lower weighting on SpeechRater^{SM}). The lowest correlation (.759) was found between the scores from H1-rater only and SpeechRater^{SM} only, and the other correlations range from .866 to .994.

Table 1:
Correlations among Five Scoring Scenarios (S1 – S5)

	H1 only	SR only	H1+SR	1/3H1+2/3SR	2/3H1+1/3SR	Mean	*SD*
S1: H1 only	1					16.11	3.40
S2: SR only	0.759	1				16.01	2.52
S3:1/2H1+1/2SR	0.956	0.917	1			32.12	5.56
S4:1/3H1+2/3SR	0.911	0.960	0.992	1		16.04	2.64
S5:2/3H1+1/3SR	0.983	0.866	0.994	0.971	1	16.07	2.96

Reliability

Reliability was calculated for the five different scoring scenarios for each form. The reliability for H1 was the lowest, ranging from .85–.89 across all the speaking items, the reliability for SR only was the highest, ranging from .94–.96; the other combinations had reliabilities above .89 for all the speaking items, see Table 2 below.

Table 2:
Reliability Range of Different Scoring Scenarios across Ten Forms

Scenario	Mean	*SD*	Reliability
H1 only	2.47-2.90	0.64-0.72	0.85-0.89
SR only	2.45-2.76	0.41-0.52	0.94-0.96
H1+SR	4.92-5.64	0.90-1.17	0.93-0.94
1/3H1+2/3SR	2.46-2.80	0.42-0.55	0.93-0.96
2/3H1+1/3SR	2.46-2.85	0.50-0.63	0.89-0.93

Reliability was calculated for the five different scoring scenarios for the combined 10 forms (see Table 3). The reliability for H1 was the lowest, .87, the reliability for SR only was the highest, .95; the reliability for other combinations were above .91 for all the speaking items. All of the reliabilities are within an acceptable range. It should be noted that combining a higher reliability score with one of lower reliability will typically result in a combined score having reliability somewhere between the higher and lower ones.

Table 3:
Reliability of Different Scoring Scenarios for Ten Forms

Scores	No. Items	Mean	*SD*	Cronbach's Alpha
H1 only	6	2.72	0.71	0.87
SR only	6	2.63	0.48	0.95
H1+SR	6	5.35	1.06	0.93
1/3H1+2/3SR	6	2.66	0.49	0.95
2/3H1+1/3SR	6	2.69	0.58	0.91

One-factor measurement model testing.

As was mentioned earlier, a single factor measurement model was fit to the data of each of the six scenarios. None of the three measurement models (congeneric, tau–equivalent, & parallel, see definition in Appendix B) fit the data well (see Table 4). The root mean square error approximation (RMSEA) was found to be in the range of .67 to .82, and the NFI and CFI were in the range of .53 to .67, all of which failed to meet the good model fit criteria. A single-factor measurement model that was proposed cannot be confirmed, which indicated that the different scenarios such as SpeechRater^{SM} only, Human Raters only, or the different combinations of the two cannot be considered as measuring the same ability dimension. This might be due to the relatively low correlation between SpeechRater^{SM} and human raters (.759) because this correlation is related to the regression used to derive the SpeechRater^{SM} model. The overall correlations between human rater 1 and 2 were .706. The case count of human rater 2 results was only 10% of the human rater 1 case count. The overall human rater 1, human rater 2 and SpeechRater^{SM} score comparison box plot can be found in Appendix A.

Table 4:

Results of Congeneric, Tau-equivalent and Parallel Hypothesis Testing

Factor Model	*N*	NFI	CFI	RMSEA
Congeneric	10,246	0.67	0.67	0.82
Tau-equivalent	10,246	0.65	0.65	0.67
Parallel	10,246	0.53	0.53	0.68

Research question 2

Do the scores obtained from SpeechRater^{SM} only, Human Raters only, or the different combinations of the two result in similar IRT parameter estimates (item difficulty and discrimination) and test takers' ability estimates of the population values and classifications?

Parameter analyses

Item difficulty.

In terms of item difficulty parameter differences for the six items (1-6), we found similar pattern among the six speaking items for the five scoring scenarios (S1 – S5) except for S2 (see Figure 1). S5 had the lowest estimated item difficulty parameters across the six items, followed by S1. There were not many differences when comparing S3 and S4. Item 2 had the lowest estimated item difficulty parameter values while item 6 had the highest item difficulty parameter estimates, but the differences between them were relatively small. S2 was close to S1 for items 1 and 4 and was very different for the other 4 items.

Item discrimination.

In terms of estimated item discrimination parameter differences for the six items (1-6), we found similar patterns across the five scenarios (see Figure 2). S1 and S5 had the lowest discrimination values, followed by S3, S4, and S2. Item 1 and 2 had the lowest discrimination values and other items had slightly higher values.

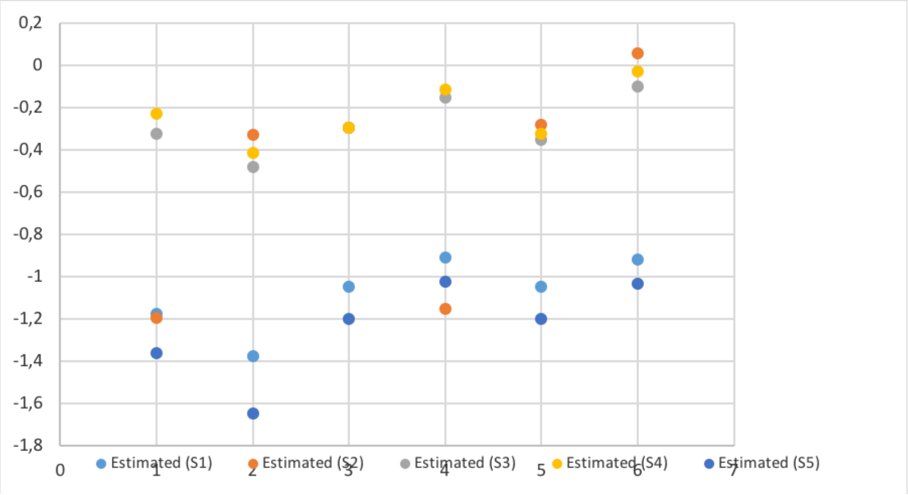


Figure 1:

Comparison of estimated difficulty parameters across five different scoring scenarios for the six speaking items.

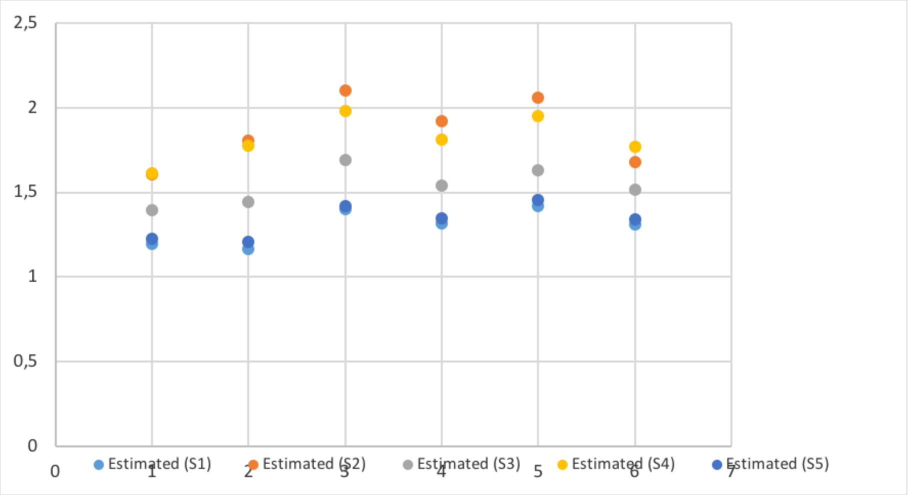


Figure 2:

Comparison of estimated discrimination parameters across five different scoring scenarios for the six speaking items.

Bias analyses

Ability estimates

In terms of test takers' estimated ability differences (see Figure 3), S1 and S2 had the largest absolute bias, RMSE and bias when combining all six items, followed by S1 and S4, S1 and S3, and S1 and S5.

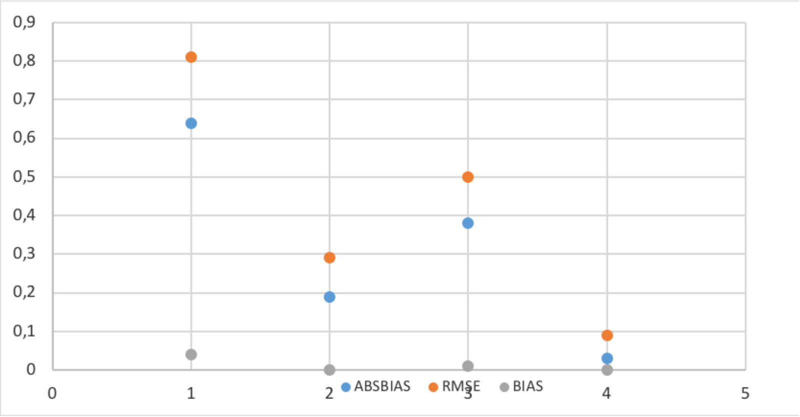


Figure 3:

Comparison of test takers' estimated ability values: comparison of four scoring scenarios against human rater (S1):1=S2- S1; 2=S3-S1; 3=S4-S1; 4=S5-S1.

Location estimates.

In terms of test takers' estimated location differences, S1 and S5 had the smallest absolute bias, RMSE and bias when combining all six items, followed by S1 and S2, S1 and S3, and S1 and S4 (Figure 4).

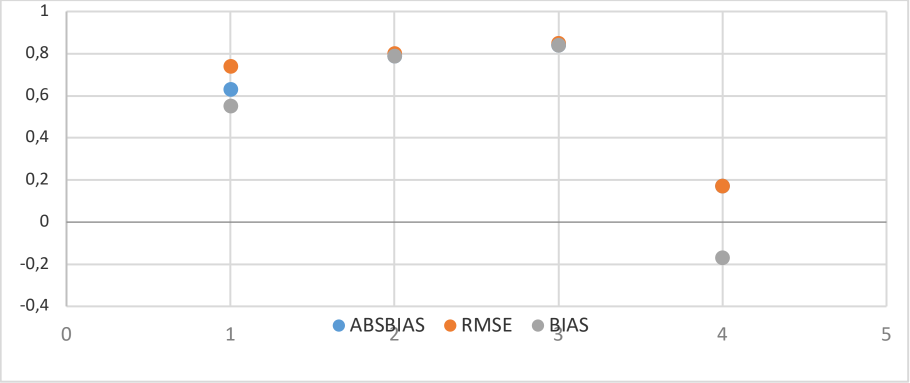


Figure 4:

Comparison of estimated location values: comparison of four scoring scenarios against human rater (S1):1=S2-S1; 2=S3-S1; 3=S4-S1; 4=S5-S1.

Slope estimates.

In terms of test takers' estimated slope differences, S1 and S2 had the largest absolute bias, RMSE and bias when combining all six items, followed by S1 and S4, S1 and S3, and S1 and S5 (Figure 5).

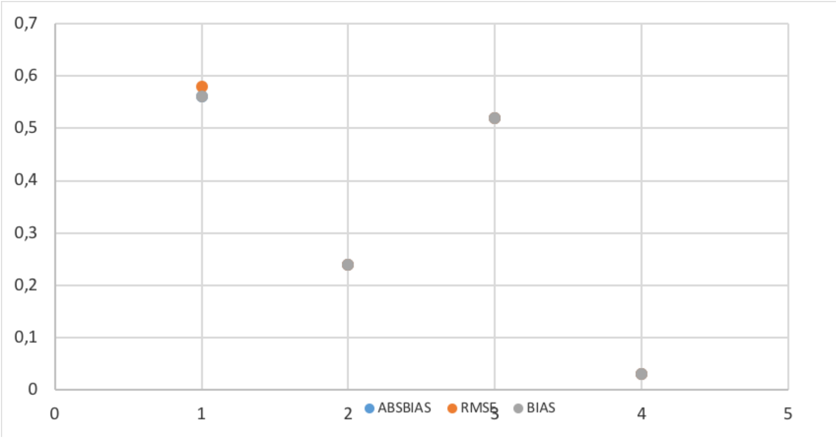


Figure 5:

Comparison of estimated slope values: comparison of four scoring scenarios against human rater (S1): 1=S2-S1; 2=S3-S1; 3=S4-S1; 4=S5-S1.

Classification

We used 3 hypothetical estimated ability values (theta) cut scores (0.0, 0.6 and 1.6) to classify test takers. We found that S5 was close to S1 at all the cut scores, followed by S3, S4 and S2 (Table 5).

Table 5:

Comparison of Test Takers' Passing Rate across Different Scoring Scenarios

	Theta=0.0	Diff vs S1	Theta=0.6	Diff vs S1	Theta=1.6	Diff vs S1
S1	46.06	–	32.19	–	5.43	–
S2	53.56	7.5	40.61	8.42	4.64	-0.79
S3	47.79	1.73	33.79	1.60	4.40	-1.03
S4	52.84	6.78	37.94	5.75	3.87	-1.56
S5	45.76	-0.3	32.19	0.00	6.19	0.76

Summary

In this paper, we summarized some findings regarding whether the scores obtained from SpeechRater^{SM} only, Human Raters only, or the different combinations of the two result in similar IRT parameters (item difficulty and discrimination) and test takers' ability estimate and classifications. Below, we summarize the findings from our research and suggest recommendations for operational practice.

For Research Question 1, we found that although the overall correlations between H1-rater scores and SpeechRater^{SM} scores were the lowest among all the different scoring designs, H1-rater had very high correlations with S5 and S3. The reliability for H1 was the lowest while the reliability for SR only was the highest (as would be expected because SR uses exactly the same criteria for scoring each time). A single-factor measurement model in the structural equation framework results indicated that all the different scoring scenarios such as SpeechRater^{SM} only, Human Raters only, or the different combinations of the two cannot be considered as measuring the same ability dimension.

For Research Question 2, in terms of test takers' estimated ability differences, S1 and S5 had the smallest absolute bias, RMSE and bias when combining all six items. For the classification passing rate based on IRT ability estimates, the closest to H1-rater were S5 and S3. Speechrater^{SM} yielded different results when compared with H1-rater at 2 out of 3 cuts. We also investigated the raw score distribution of both H1-rater and Speechrater^{SM} scores and found that their distributions were different to some extent (see Figure 6 in the Appendix). The fact that the score distributions differ would mean that different cut scores might result in different findings with respect to the classifications. The test information curves and standard error curves of the Speechrater^{SM} (S2) did not look similar to that of human rater (S1) (see Figures in Appendix C). Both test information curves and standard error curves indicated that S5 and S3 are closer to S1 than other scoring scenarios.

Generally speaking, some of the main findings indicate that the SR-only approach is most different from the H1-only approach in comparison to the other approaches (for example, it has the lowest correlation of 0.759 and a completely different pattern of difficulty parameter estimates across the six items, as shown in Figure 1). The explanation for these results would seem to be the fact that S3 - S6 all contain H1 in them (in combination with the SpeechRater^{SM} score), so it is a given that they will be more similar to S1 than S2.

Discussion

In this study, we used different approaches, such as classical and IRT modeling techniques to compare human rater and Speechrater^{SM} scores. We created different scoring scenarios based on arbitrary weights and investigated their differences. It seems that these approaches are effective in detecting the differences between human and automated scoring; the research results can help practitioners make decisions in terms of how to combine human and machine scores in the field. Identifying issues and differences between Speechrater^{SM} and human rater can help improve Speechrater^{SM}. However, the real data used for the human rater might not be perfect (Wang & Yao, 2013), which may have prevented us from finding the real issue. Simulation studies are needed to compare Speechrater^{SM} and human rater in a controlled manner.

In this study we used real data and investigated the differences between different scoring methods. We found that there are some systematic patterns in the combined scenarios based on both classical and IRT approaches, such as their raw score distributions, test information curves (see Appendix C), standard error curves (see Appendix C), and percentage of passing rate. As pointed out above the percentage of passing rate results may well differ for different cut scores. The fact that S5 and S3 are closer to S1 than other scoring scenarios is within our expectation because they account for 1/2 and 2/3 of the human score.

In this study, we compared the differences among different combinations of human and machine scores, and more such studies are needed (Attali, 2013; Zhang et al., 2013). We believe that the use of the statistical analyses (both classical and IRT) in this study is a useful way to advance the study of automated scoring in the evaluation of speech responses. Our study can help our clients make decisions related to machine and human scoring. A future study is needed to provide guidelines about how to establish the set of weights that will generate optimal reliability to combine human rater and SpeechRater^{SM} scores. Reliability is just one criteria and there might be other criteria, such as validity that are also important to investigate. A simulation study should be conducted to test different scoring combination scenarios, and different cut scores, under different conditions.

Limitations

Since human raters' scores are based on three areas (delivery, language use, and topic development) while SpeechRater^{SM} scores are based only on delivery and language use, differences between the two types of scores will probably always exist (and these differences relate to validity issues). There are issues and questions that need to be investigated and answered before an automated speech scoring system such as SpeechRater^{SM} can be implemented for the operational scoring of English language assessment speaking items.

We want to stress that the automated speech scoring engine SpeechRater^{SM} is still evolving, and that additional features, covering a more extended subset of the Speaking construct, are currently under development and are planned for use in future scoring models. Moreover, we are also exploring alternative approaches to the standard linear regression scoring model currently used by SpeechRater^{SM}, which may lead to improvements in its scoring performance. The current Speechrater^{SM} distribution has high frequency at middle scores (see Figure 6 in the Appendix C), which may be due to the regression toward the mean effects from the linear regression scoring model.

Notes

1–3. See definitions in Appendix B.

References

Attali, Y. (2013). Validity and reliability of automated essay scoring. In M. D. Shermis & J. Burstein (Eds.) *Handbook of automated essay evaluation: Current applications and new directions* (pp. 181-198). New York, NY: Routledge.

Attali, Y., Bridgeman, B., & Trapani, C. (2010). Performance of a generic approach in automated scoring. *Journal of Technology, Learning, and Assessment*, *10*, 1-16. Retrieved from http://www.jtla.org.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e–rater®v.2. *Journal of Technology, Learning, and Assessment, 4*, 1-30. Retrieved from http://www.jtla.org.

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

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.). *Testing structural equation models* (pp. 136–162). Beverly Hills, CA: Sage.

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) (pp.722-731). Portland, OR: Association for Computational Linguistics.

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.

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.

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modeling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods, 6*, 3–60.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criterion versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Jöreskog, K., & Sörbom, D. (1984). *LISREL IV: Analyses of linear structural relationships by maximum likelihood, instrumental variables, and least squares methods*. Mooresville, IN: Scientific Software.

Laundauer, T. K., Laham, D. & Foltz, P. (2003). Automatic essay assessment. *Assessment in Education, 10*, 295–308.

Maruyama, G. (1998). *Basics of structural equation modeling*. Thousand Oaks CA: Sage.

Nichols, P. (2005). Evidence for the interpretation and use of scores from an automated essay scorer (White Paper). Iowa City, IA: Pearson.

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.

Streeter, L., Bernstein, J., Foltz, P., & Deland, D. (2011). *Pearson's automated scoring of writing, speaking, and mathematics* (White Paper). Iowa City, IA: Pearson.

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.

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

Wang, Z., Zechner, K., & Sun, Y. (2016). Monitoring the performance of human and automated scores for spoken responses. *Language Testing.* Advance online publication. doi: 10.1177/0265532216679451

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 North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL–HLT),* Montreal, Canada: Association for Computational Linguistics.

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

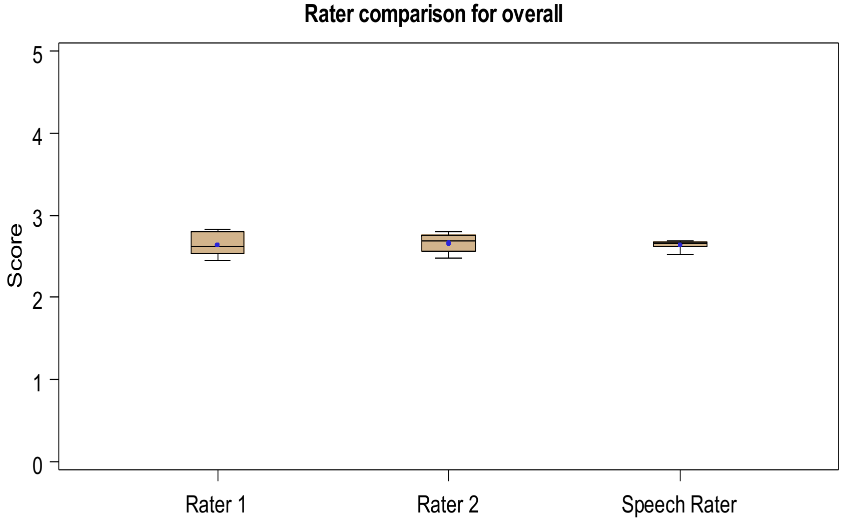
Yoon, S.-Y., Bhat, S., & Zechner, K. (2012). *Vocabulary profile as a measure of vocabulary sophistication.* Paper presented at the 7th Workshop on Innovative Use of NLP for Building Educational Applications, North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL–HLT). Montreal, Canada: Association for Computational Linguistics.

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

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

Zhang, M., Breyer, J., & Lorenz, F. (2013). *Investigating the suitability of implementing e-rater in a large-scale English language testing program.* (Research Report RR-13–36). Princeton, NJ: Educational Testing Service.

Appendix A



Note. Human rater 2 is only 10% of the H1 data.

Figure A1. Rater comparison between rater 1, rater 2 and SpeechRater^{SM}.

Appendix B

Parallel measures

The measures (items) comprising a scale are parallel if the following two conditions hold:

1. $\tau_{ip} = \tau_{jp}$ for all *i* and *j*; τ = true scores; *p* = person;
2. $\sigma_{ei}^2 = \sigma_{ej}^2$ for all *i* and *j*: e = error:

This implies that the amount of variation in the item score that is determined by the true score is the same for all items. It also implies that the expected value of each of the items will be the same.

Tau-equivalent measures

When measures are tau-equivalent, $\tau_{ip} = \tau_{jp}$ for all *i* and *j*, as in the case of parallel measures, but we relax the assumption that $\sigma_{ei}^2 = \sigma_{ej}^2$ for all *i* and *j*.

Congeneric measures

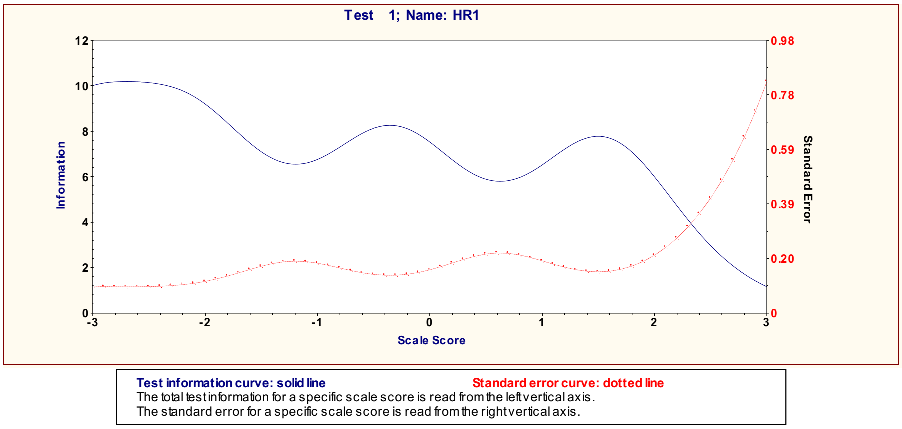
Congeneric measures relax both the assumption that $\tau_{ip} = \tau_{jp}$ for all *i* and *j*, and that $\sigma_{ei}^2 = \sigma_{ej}^2$ for all *i* and *j*.

Appendix C

Test Information Curves and Standard Error Curves

Figure C1:

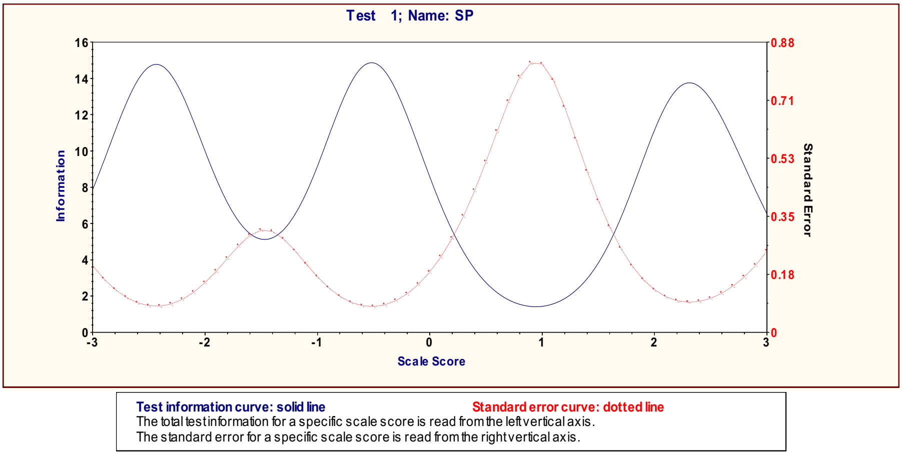
S1.



The total test information for a specific scale score is read from the left vertical axis.
The standard error for a specific scale score is read from the right vertical axis.

Figure C2:

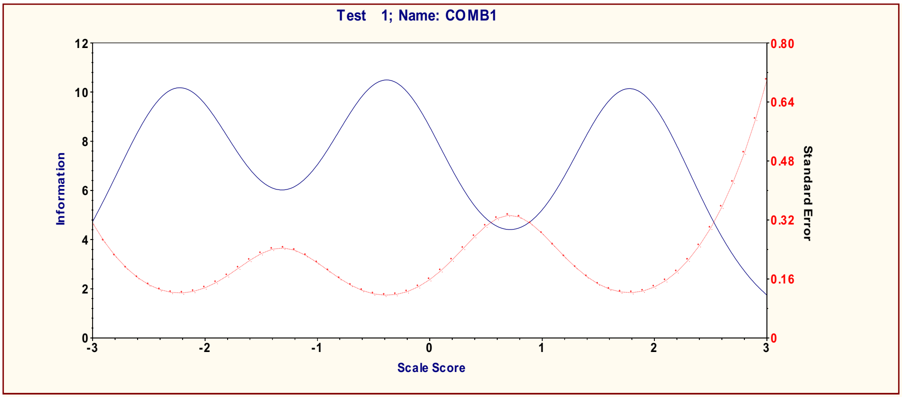
S2



The total test information for a specific scale score is read from the left vertical axis.
The standard error for a specific scale score is read from the right vertical axis.

Figure C3:

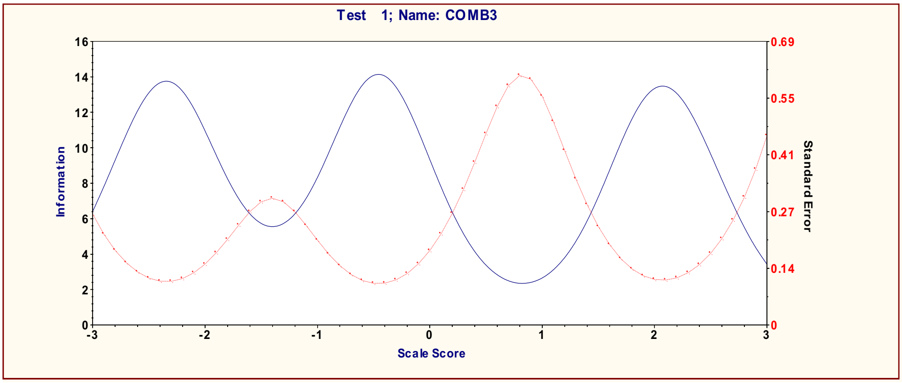
S3.



Test information curve: solid line Standard error curve: dotted line
The total test information for a specific scale score is read from the left vertical axis.
The standard error for a specific scale score is read from the right vertical axis.

Figure C4:

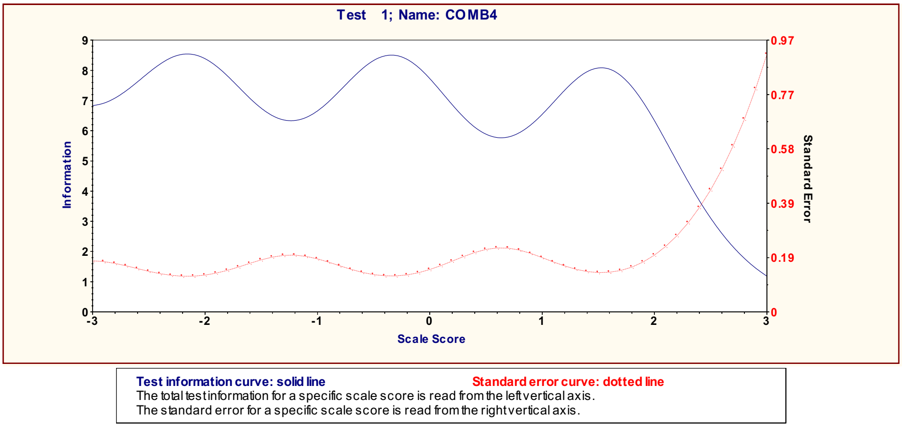
S4



Test information curve: solid line Standard error curve: dotted line
The total test information for a specific scale score is read from the left vertical axis.
The standard error for a specific scale score is read from the right vertical axis.

Figure C5:

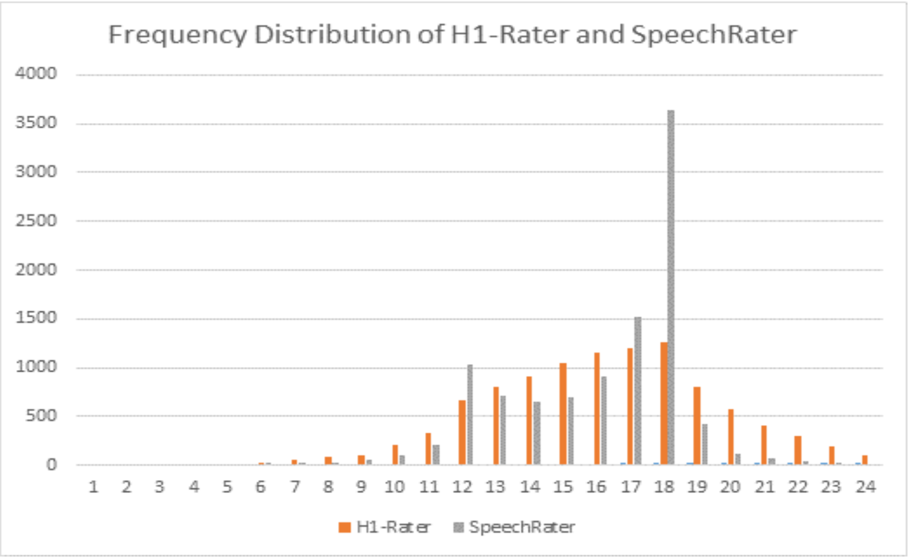
S5



Test information curve: solid line Standard error curve: dotted line
The total test information for a specific scale score is read from the left vertical axis.
The standard error for a specific scale score is read from the right vertical axis.

Figure C6:

Frequency distribution of H1-Rater and SpeechRater^{SM}.


Frequency Distribution of H1-Rater and SpeechRater
H1-Rater
SpeechRater