English Language Teaching; Vol. 13, No. 1; 2020
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

Comparison of Automatic and Expert Teachers' Rating of Computerized English Listening-Speaking Test

Cao Linlin[1]

[1] Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou, China

Correspondence: Cao Linlin, Guangdong University of Foreign Studies, No. 2 North Baiyun Avenue, Baiyun District, Guangzhou, China.

Received: September 10, 2019 Accepted: November 24, 2019 Online Published: December 4, 2019

doi: 10.5539/elt.v13n1p18 URL: https://doi.org/10.5539/elt.v13n1p18

Abstract

Through Many-Facet Rasch analysis, this study explores the rating differences between 1 computer automatic rater and 5 expert teacher raters on scoring 119 students in a computerized English listening-speaking test. Results indicate that both automatic and the teacher raters demonstrate good inter-rater reliability, though the automatic rater indicates less intra-rater reliability than college teacher and high school teacher raters under the stringent infit limits. There's no central tendency and random effect for both automatic and human raters. This research provides evidence for the automatic rating reform of the computerized English listening-speaking test (CELST) in Guangdong NMET and encourages the application of MFRM in actual score monitoring.

Keywords: English listening-speaking test, rater effects, automatic rating, MFRM

1. Introduction

In recent years, the National Matriculation English Test (NMET) in China pays increasing attention to the measurement of communicative ability and productive skills, one good embodiment is the Computerized English Listening and Speaking Test (CELST) issued in 2011 in Guangdong province. Consisting of three individual tasks, the test aims to examine English pronunciation, listening proficiency, interactional competence, short-term memory, ability to make summary, etc. CELST meets the requirements of a good oral test, such as focusing on information exchange, creating contextualized situation and authenticity to incorporate interactiveness into language communication (Yang, 1999). As a subjective test, rating quality is of paramount importance to guarantee test validity. However, rater effects are inevitable, especially the arduous rating process due to the large number of examinees and the need to realize quick turnover of test results. Therefore, computer automatic rating is advocated, hoping to save time and energy, raise the level of fairness and eliminate subjective human rating errors and to obtain test results more efficiently. Thus there is an urgent need to examine the appropriateness and practicality of automatic rating in replacing human rating. Consequently, this research is designed with the purpose to compare the rating difference between computer automatic rater and expert teacher raters' rating of CELST.

2. Literature Review

Automatic rating has been investigated frequently by researchers in scoring of writing, reading and speaking (Ishi et al., 2008; Huang et al., 2009). A series of automatic rating methods for the optimal solution have been explored (Coniam & David, 2009; Ge, 2010) and automatic rating tools and programs have been developed for various purposes (Li & Liu, 2013; Zhou et al., 2019). Generally, previous researchers are in favor of automatic rating. In particular, Zou & Chen (2010) claim that computer-assisted rating would be an efficient way to guarantee validity. However, previous validation research of automatic rating focus on the correspondence between human rating and automatic scoring (Xi, 2010; Xi et al., 2008). However, high correlation with human rating doesn't permit the validity of automatic rating since human raters are subject to rater effects (Yang, 2002). A plethora of studies have probed into instability or errors of human raters concerning their severity/leniency, accuracy/randomness, halo effect, central tendency, bias and restriction of range (Saal et al., 1980; Wolfe & Chiu, 1997; Xi & Mollaun, 2009; Dai, 2011), the validity of automatic rating will be impaired if it takes human rating as its exclusive quality standard.

Though many researchers have investigated rater effects and rating behaviors in spoken tests (Bachman, Lynch

& Mason, 1995; McNamara & Lumley, 1997; Meiron & Schick, 2000; Caban, 2003; Brown, 2007, Dai & You, 2010; Duan, 2011) and writing tests (Goodwin, 2016), the test types in these studies are mainly traditional single-skill test items, few studies concern integrated tests. What's more, previous research is restricted to the validation of human raters, rather than automatic scorer. Only Zhang (2013) contrasts automated and human scoring of essays comprehensively and describe methods to leverage the two scoring approaches to meet particular goals for assessment.

Therefore, the validity of automatic rating should be investigated synthetically. Since Many-facet Rasch model analysis (Linacre, 1994 & 1999; Bond, 2007), and its computer program FACETS have been widely adopted in rater effects analysis, especially in spoken test rating quality control (Wolfe et al., 2001; Tian, 2006; Bonk, 2007; Zhang, 2008; He & Zhang, 2008), rater effects comparison (Liu, 2010), and rater training (Weigle, 1998), it fits the current study as a rigorous tool. In addition, Wang (2015), Zhou and Zeng (2016) employ MFRM model analysis to investigate the validity of automatic scoring in CELST in Guangdong and found that automatic rater demonstrates higher rater reliability and rating accuracy, which act as the starting point of this study.

Review of related literature indicates that few studies are in regard to validity research of automatic rating in integrated listening to speaking tests, not to speak of an all-around probe of automatic rating. Besides, as a large-scale and high-stake test, the CELST is a new item format with relatively few studies concerning its rating. Therefore, following Wang (2015) and Zhou and Zeng's (2016) research, the present study aims to collect evidence in different contexts for the validity of automatic rating in CELST. Experiment is designed to compare the difference between computer automatic rater and expert teacher raters in a mock test of CELST in Guangdong NMET with many-facet Rasch model adopted to examine rater severity, reliability, central tendency and randomness. The research questions are:

(1) To what extent are automatic and expert teachers' rating of CELST different concerning severity and reliability?

(2) Do automatic rater and expert teacher raters demonstrate central tendency and random effect?

3. Method

3.1 Participants

Participants in this study include 119 test takers and 6 raters. The test takers are senior three students in a school in Guangzhou, who have already practiced for the CELST. As to the 6 raters, 3 (rater 1, 2 and 3) are college teachers, 2 (rater 4 and 5) high school teachers and 1 (rater 6) the automatic rater. To avoid influence of major or experience, the selected teacher raters have similar educational, teaching and rating experience which qualify them as expert raters.

3.2 Instruments

The 2013 CELST is adopted in this research. In this test, task one is the reading aloud task, which requires the test takers to watch a video first, and then read the passage sentence by sentence on the screen. (See Appendix A for the detailed information of Task one). While test takers are watching the video for the first time, the passage is presented as English subtitles at the bottom of the screen, and moves on in accordance with the content of the video. They should pay attention to the dubbing and try to imitate the pronunciation, intonation, stress, rhythm, etc. After watching, the video plays a second time with subtitles but without dubbing. This time examinees read the subtitles with their speech recorded. Task Two (See Appendix B) is role play, which requires test takers to listen to a short conversation between two persons and imagine that he/she is one of the person. After listening, there will be eight short-answer questions. Test takers need to give their answers, which also will be recorded. At last, Task Three (See Appendix C) is a retelling task. There will be hints for the complicated or new words for the students, and the story will be played twice. Students then try to retell the story by using proper words and sentences.

An analytical rating scale is adopted for this test. Test takers will be rated considering their pronunciation, intonation, rhythm, fluency, grammar and accuracy. They also need to cover the content, or rather, completely and accurately finish the three tasks according to the requirement, for grades are given on the number of information points that the candidates can provide. Table 1 is the rating scale for Task One (the rating scales of Task Two and Task Three are in the appendices).

Table 1. Rating scale of Read-aloud task in CELST in 2013 Guangdong NMET

	Pronunciation & Intonation		Speed & content	
Grade	Score	Standard	Score	Standard
A	8--12	Clear and accurate articulation; correct and natural intonation; coherent and fluent flow of speech	6--8	Read in accordance with the original speed with complete content (miss three words at most)
B	4--7	Basically accurate articulation; correct intonation on the whole; relatively fluent flow of speech	3--5	Basically read in accordance with the original speed with several words missed
C	0--3	Most of the phonemes are incorrectly pronounced; non-fluent flow of speech	0--2	Cannot read according to the original speed, missing an incomplete sentence or more than 10 words

3.3 Scoring Procedures

The teacher raters have received training before the formal scoring. Firstly, all the 5 raters were given 30 minutes to familiarize themselves with the rating scale and the test materials. Secondly, the assessment criteria was introduced in detail to the raters. A discussion was held about the scales and any unclear descriptions concerning the rating scale was made clear. Then 20 samples from the NMET representing a range of proficiency levels were given to the raters to rate individually. Their rating results were collected and compared with the original scores to see whether the teacher raters have achieved a common interpretation on the rating criteria. In the rater training process, the standard sample was determined through the consistency of ratings between teacher raters and the original rating. When inconsistency appeared, the recording of that examinee should be discussed to re-score the examinee's proficiency on each language trait. At this step, if most of the raters arrived at a consensus upon the rating of this specific recording, the rating would be adopted. The rater who still held a different opinion had to be re-introduced to the rating criteria. At the end of the scoring meeting, possibilities that would cause rating divergences were made clear.

In the formal rating stage, each rater rated the 119 samples independently in a computer lab with raters listening to the recordings and marked them on the computer screen with the scoring software. The rating process lasted for 5 hours, including the automatic rating. After rating, mean score given by college teacher raters (No. 1) and high school teacher raters (No. 2) for each individual task for each student were obtained.

3.4 Data Analysis

Designed on the basis of Many-Facet Rasch Model, the statistical software FACETS can provide estimates of rater severity on a linear scale as well as fit statistics, making it the most widely used instrument in the analysis of rater Response Theory. In this study, Facets Version 3.58.0 (Linacre, 2005) for windows is employed. A Partial Credit Model that allows researchers to examine rater severity at an individual level provides more precise information for the purpose of the study. The following is the function expression of the Many-Facet Rasch Model:

$$Log(P_{nijmk}/P_{nijm(k-1)})=B_n - D_i - C_j - E_m - F_k$$

In the above function, P_{nijmk} represents the probability of the candidate *n* gets the score *k* on item *i* when rated by rater *j*; $P_{nijm(k-1)}$ stands for the probability that candidate *n* obtains *k-1* on item *i* when rated by rater *j*; B_n is the proficiency parameter of candidate *n* (n=1,2,.....N); D_i equals the difficulty parameter of item *i* (i=1,2,.....I); C_j is the severity of rater *j* (j=1,2.....J); E_m represents the relative difficulty of rating item *m* (the relative difficulty that a candidate can get high score on this item.); F_k equals the step difficulty in Partial Credit Model that a candidate obtains the score from *k-1* to *k*, each item shares the same *k* levels of rating. This function represents the barrier to being observed in category *k* relative to category *k-1*.

We have extracted the following information: infit mean square values (indicating consistency), category of scores (indicating central tendency if there are any) and logit values (indicating severity) measures for each type of raters on each task and on the whole test. Afterwards, we transferred the values to descriptions of rating behavior (such as Severe, inconsistent, etc.). All of the above information has been reported in this research.

4. Results and Discussion

4.1 Rater Severity

Figure 1 graphically displays automatic and human raters' relative severity, the examinees' proficiency levels and the difficulty levels of the three tasks.

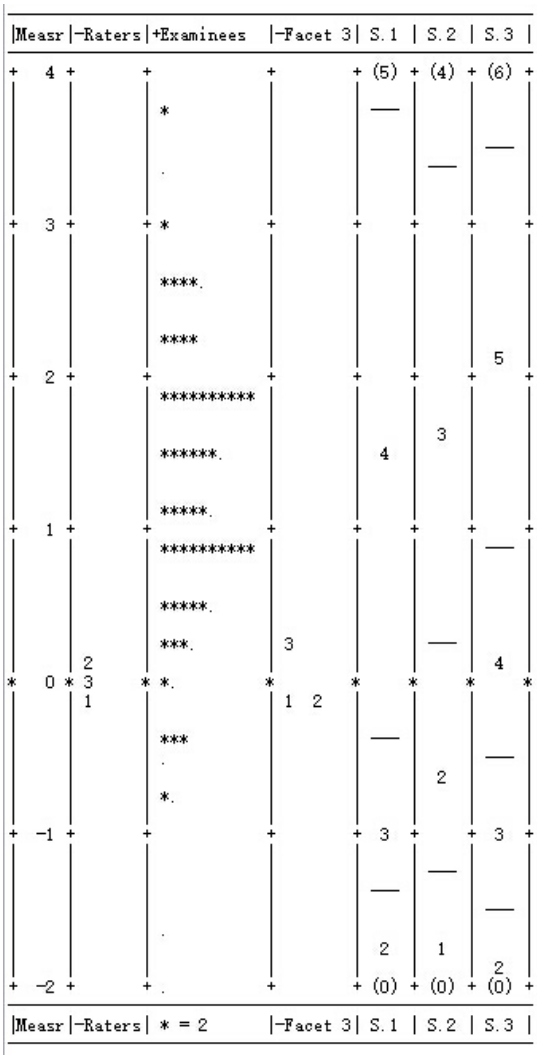

Figure 1. The facets map

Figure 1 indicates that the range between the most severe and the most lenient rater is from -0.10 to 0.16 (about 0.26 logit spread), automatic and human raters have similar severity levels without significant difference. The ability measures of the examinees spread from about -1 to 4 (about 5 logits spread). So the spreads of examinees' ability is about 25 times larger than the spreads of rater severity. The much larger range of examinee proficiency compared to the range of rater severity indicates that the impact of individual differences in rater severity on examinee scores is likely to be relatively small (Yang, 2010). As Myford and Wolfe (2000) claim, the influence of individual differences in rater severity on scores is considered quite big if the range of examinee proficiency is not bigger than twice as wide as the range of rater severity. Furthermore, from column 4 we know that task three (retelling) is more difficult than task one (reading aloud) and task two (short answer questions). Task one and two share the same difficulty level.

4.2 Rater Reliability

4.2.1 Intra-rater Reliability

Before reporting how consistent each type of rater is when he/she is applying the rating scales, it should be made clear that the "reliability" here refers to intra-rater reliability, that is, how each type of rater is consistent with the rating of themselves. According to Myford and Dobria (2006), more stringent limits (0.7-1.3) are required for

high-stakes tests rating. As shown in Table 2, the college and high school teacher raters (No. 2 and 3) indicate satisfying fit values within the quality-control limits. While the infit values for the automatic rater go slightly beyond the upper limit of 1.3, with an infit value of 1.37. The slightly larger infit value of the automatic rater implies that it does not perform the same consistency in using the rating scale as the other two types of raters.

Table 2. Raters measurement report

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Dicrm	N	Raters
1315	354	3.7	3.79	.10	.08	.85	-1.6	.88	-1.3	1.05	2	2
1321	354	3.7	3.80	.06	.08	1.37	3.5	1.20	1.9	.83	3	3
1351	354	3.8	3.88	-.16	.09	.78	-2.4	.81	-2.1	1.14	1	1
1329.0	354.0	3.8	3.82	.00	.09	1.00	-.2	.96	-.5		Mean(Count:3)	
15.7	.0	.0	.04	.12	.00	.26	2.6	.17	1.8		S.D.(Populn)	
19.3	.0	.1	.05	.14	.00	.32	3.2	.21	2.2		S.D.(Sample)	

Model, Populn: RMSE .09 Adj (True) S.D. .08 Separation .93 Reliability .46

Model, Sample: RMSE .09 Adj (True) S.D. .11 Separation 1.34 Reliability .46

Model, Fixed (all same) chi-square: 5.4 d.f.: 2 significance (probability): .06

Model, Random (normal) chi-square: 1.5 d.f.: 1 significance (probability): .23

4.2.2 Inter-rater Reliability

From Table 3 we can see that all the raters demonstrate great inter rater consistency. Rater 3, 4, 1 and the automatic rater are all beyond 0.93, which is relatively high. As to the average error, the automatic rater is the second best after human rater 3. The same for degree of identity, when the automatic rater is 0.73 while rater 3 is 0.76. At last, in adjacent consistency four raters among them, rater 1, 2, 4 and the automatic rater all get the value of 1, and rater 3 obtains 0.99, which indicates that the five raters' scorings are almost the same and one can be replaced by another. The great inter-rater reliability reveals that the ratings of automatic rater is comparable to that of the human raters.

Table 3. Rating consistency

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Automatic rater	Benchmark score
Relevancy	0.943	0.936	0.945	0.945	0.838	0.930	1
Average error	1.095	1.468	1.047	1.263	2.107	1.066	0
Degree of identity	0.714	0.563	0.756	0.681	0.613	0.731	1
Adjacent consistency	1	1	0.992	1	0.933	1	1

4.3 Central Tendency

Central tendency is the phenomenon that during rating, some raters may tend to use categories or scores in the middle of the rating scales all too often. This phenomenon indicates that raters cannot perfectly differentiate the examinees according to their proficiency, they can't well perceive and apply the rating scales' standards, so they choose to play it safe during rating, especially when they are rating middle-level examinees. In consequence, more examinees are placed in middle level. In large-scale high-stakes tests, the rating behaviors of raters, and the process and quality of their rating will be supervised closely with occasional feedback. So raters tend to have the possibility of playing it safe by more frequently using the middle scores or categories in the rating scales. Central Tendency can be interpreted through observing the category statistics in Table 4, which indicates the overall circumstance of all the raters' application of categories in the rating scales. The "Counts Used" in column two indicates the times that each category has been used by raters. While in column 3, the "Counts%" concludes the proportion of usage of each category by the raters. We can obviously see that category 4 is most frequently employed (40%), and the second comes category 3 (28%), with category 1 (nearly 1%) accounts for the least percentage. Therefore, through the analysis of the general situation in using the rating scale, it can be summarized that obvious central tendency does not exist.

Table 4. Category statistics

Data Category Score	Counts Used	%	Cum. %	Quality Control Avge Meas	Exp. Meas	Outfit MnSq	Step Calibrations Measure	S.E.	Expectation Measure at Category	-0.5	Most Probable from	.5 Cumul. Probabil.at	Cat Peak Prob
0	9	1%	1%	-2.54	-3.64	2.9			(-4.65)		low	low	100%
1	7	1%	2%	-2.41	-2.23	.7	-2.54	.43	-3.61	-4.17		-3.70	16%
2	84	8%	9%	-1.21	-1.38	1.3	-4.29	.31	-2.65	-3.16	-3.42	-3.36	50%
3	300	28%	38%	-.23	-.14	.9	-2.11	.13	-.76	-1.91	-2.11	-2.03	64%
4	428	40%	78%	1.78	1.84	1.6	.51	.09	1.84	.52	.51	.51	65%
5	204	19%	97%	3.07	2.96	.9	3.21	.09	4.24	3.10	3.21	3.15	57%
6	30	3%	100%	3.90	3.60	.8	5.22	.20	(6.39) (Mean)	5.48	5.22 (Modal)	5.32 (Median)	100%

4.4 Random Effect

Random effect is the situation when raters show apparent inconsistency on certain or a number of rating scales, and the inconsistency indicates great randomness. Random effect may happen in situations when a rater exhibits inconsistency during the overall rating period in his/her rating, or when two raters show overly significant consistence with each other. If random effect exists in the rating process, it will be difficult for raters to correctly evaluate examinees' real ability. Examinees' performance and the measurement accuracy can be indicated by Separation and Reliability, especially when the value of the former or the latter is low. Table 5 demonstrates proficiency of the examinees. As mentioned in the central tendency part, the index of Separation is 1.63, G is 2.51, so proficiency of the examinees can be divided into 7 levels. And with a relatively high Reliability (0.73), it can be inferred that there doesn't exist random effect.

Table 5. Examinee measurement report

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Average	Measure	Model S.E.	Infit MnSq	ZStd	Outfit MnSq	ZStd	Estim. Discrm	Num Examinees
33.5	9.0	3.7	3.74	1.17	.56	.96	-.2	.96	-.2		Mean(Count: 119)
4.8	.0	.5	.53	1.11	.13	.81	1.3	.80	1.3		S.D. (Populn)
4.8	.0	.5	.53	1.11	.13	.82	1.3	.80	1.3		S.D.(Sample)

With extremes, Model, Populn: RMSE .58 Adj (True) S.D. .94 Separation 1.63 Reliability .73

With extremes, Model, Sample: RMSE .58 Adj (True) S.D. .95 Separation 1.64 Reliability .73

Without extremes, Model, Populn: RMSE .56 Adj (True) S.D. .80 Separation 1.43 Reliability .67

Without extremes, Model, Sample: RMSE .56 Adj (True) S.D. .80 Separation 1.44 Reliability .67

With extremes, Model, Fixed (all same) chi-square: 472.3 d.f.: 118 significance (probability): .00

With extremes, Model, Random (normal) chi-square: 88.3 d.f.: 117 significance (probability): .98

4.5 Discussion

An initial objective of the study is to identify how automatic rater and expert teacher raters differ with each other in rating severity and reliability. In the respect of severity, results differ from Zhou and Zeng's (2016) findings which showed that automatic rater and human raters were significantly different concerning severity but exerted no decisive influence on examinees' proficiency distribution. This study reveals that automatic rater and expert teacher raters display similar severity. This finding broadly supports the work of Wang (2015), which confirmed the rating accuracy and construct generalizability of the automatic rater.

Reliability is examined from two perspectives, intra-rater and inter rater reliability. As to the intra-rater reliability, one unanticipated finding is that college teacher raters and high school teacher raters both demonstrate good rating consistency, while the infit values for the automatic rater is slightly beyond the upper limit of 1.3 (infit value 1.37). Myford and Dobria (2006, Workshop Notebook) suggest three possible reasons for the misfit of

larger infit values: first, the rater may not possess a comprehensive understanding of the scale structure and thus is incapable of distinguishing the categories; second, the rater may be biased towards certain kinds of examinees or certain characteristics on examinees and thus can't apply the rating scale consistently; and third, fatigue may have caused the rater fail to apply the scale consistently over time, which is not considered as a factor for automatic rater in this study. A possible explanation for this might be that an excessively stringent limit range is adopted. Moreover, the rating rationale of automatic rating should be investigated for possible reasons. Actually, 1.37 is not significantly different from 1.3, thus automatic rating is deemed acceptable in this case. While for the inter-rater reliability, both the computer automatic rater and expert teacher raters manifest perfect self-consistency. It is possible that these results are only valid for small sample rating, thus human raters are not subject to factors affecting rating validity, such as fatigue. Also, the experience of teachers in CELST rating plus adequate rater training have secured rating quality.

The second question in this research is to investigate whether the two types of raters manifest central tendency and random effect. Through examining the statistics generated by Rasch, and analyzing measurement report and the probability curves, no central tendency effect is found for both automatic and human raters. Both of the two types of raters can appropriately differentiate examinees' proficiency and well apply the rating scale. Moreover, the automatic rater's rating on 4 candidates and the college teacher raters' rating on 1 candidate were discovered to have outfit values larger than infit statistics. For the college teacher raters, this unexpected results affirm previous research conclusions on human rating in performance test that rating errors are inevitable for human raters regardless of the training they are given and the experience they have (McNamara, 1996; Upshur & Turner, 1999; Eckes, 2005). The unexpected ratings pointed out by the Rasch model provides information for the test administrators to seek for possible causes that leads to this rating result, whether they lie in the incapability of applying the rating scales or biases on certain tasks or certain kinds of test takers (Myford, 2006).

All in all, automatic rating has high reliability and consistency and displays no central tendency and random effect concerning the rating of integrated spoken test like CELST. This study produced results corroborating the findings of a great deal of the previous work which favor automatic rating.

5. Conclusion

5.1 Major Findings and Implications

The current research of rating differences between automatic and teacher raters on CELST is conducted to evaluate the rating validity of automatic rater synthetically through comparison. This research firstly examines the severity and reliability of different groups of raters. Then, possible central tendency and random effect are also investigated to make the comparison more comprehensive. Though the college teacher raters and high school teacher raters perform more consistently than the automatic rater in applying the rating scales, the automatic rating is still considered acceptable. As to the severity, their ratings are acceptable in the model and their severity differences are not statistically different. Moreover, central tendency and random effect did not appear. Four ratings from the teacher raters and one from the automatic rater are unexpected compared with those in the model but are acceptable.

There is an increasing trend of the application of automatic scoring in various types of items. Thus the investigation of automatic rating's validity, reliability and rating efficiency and accuracy became an urgent issue. Prior studies have noted the importance of automatic rating validity, but there's a lack of synthetical evaluation of automatic scoring. Therefore, by evaluating the differences between automatic and human rating, this research sheds some light on the evaluation and inspection on subjective rating qualities in the CELST, and will provide some suggestions to the rating process and rater training of the test, as well as the improvement and future development of automatic rating.

5.2 Limitations and Recommendation for Future Research

Firstly, the sample size (altogether 6 raters) of this research is so small that the generalizability of the discoveries is limited. Therefore, in future research, a larger number of raters with different backgrounds can be employed to explore whether there exists larger differences from various perspectives between automatic and human raters, for more reliable statistics and better generalizability. Secondly, this research employs only quantitative rating data, which is surface information rather than deeper-level information. Future research can focus on investigating the setting and rating rationale of automatic rating programs. In the end, the rating of subjective performance should be a combination of human and automatic scoring to insure rating quality as well as efficiency.

References

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing, 12*(2), 238-257. https://doi.org/10.1177/026553229501200206

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental measurement in the human sciences*. (2nd ed.). Lawrence Erlbaum.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20*(1), 89-110. https://doi.org/10.1191/0265532203lt245oa

Brown, A. (2007). An investigation of the rating process in the IELTS oral interview. In M. Milanovic, & C. J. Weir (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 98-141). Cambridge: Cambridge University Press.

Caban, H. L. (2003). Rater group bias in the speaking assessment of four L1 Japanese ESL students. *Second Language Studies, 21*(2), 1-43.

Coniam & David. (2009). Experimenting with a computer essay-scoring program based on ESL student writing scripts. *ReCALL, 21*(02), 259. https://doi.org/10.1017/S0958344009000147

Dai, Z. (2011). A study of the reliability of Computerized Oral Proficiency Test. *Computer-assisted Foreign Language Teaching, 138*, 45-50. https://doi.org/10.3969/j.issn.1001-5795.2011.02.008

Dai, Z., & You, Q. (2010). Analysis of rater bias on Computerized College English Oral Proficiency Interview. *Foreign Language World, 5*, 87-95. https://doi.org/CNKI:SUN:WYJY.0.2010-05-016

Duan, R. (2011). A Many-facet Rasch Model Analysis of rater effects in CET-SET. Dissertation for Master Degree. Hebei University of Science and Technology.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197-221. https://doi.org/10.1207/s15434311laq0203_2

Ge, S. (2010). A comparative study of automated essay scoring techniques for college students' English writing. *Journal of Guangdong University of Foreign Studies, 21*(3), 87-90. https://doi.org/10.3969/j.issn.1672-0962.2010.03.019

Goodwin, S. (2016). A many-facet rasch analysis comparing essay rater behavior on an academic English reading/writing test used for two purposes. *Assessing Writing, 30*. https://doi.org/10.1016/j.asw.2016.07.004

He, L., & Zhang, J. (2008). A Many-Facet Rasch Analysis of the reliability in CET-SET English oral test. 31 (4): 388-39. https://doi.org/CNKI:SUN:XDWY.0.2008-04-008

Huang, S., Li, H., Wang, S., Liang, J., & Xu, B. (2009). Automatic assessment of speech fluency in computer aided speech grading systems. *Tsinghua Univ (Sci & Tech), 49*(S1), 1349-1355.

Ishi, C. T., Ishiguro, H., & Hagita, N. (2008). Automatic extraction of paralinguistic information using prosodic features related to Fo, duration and voice quality. *Speech Communication, 50*(6), 531-543. https://doi.org/10.1016/j.specom.2008.03.009

Li, X., & Liu, J. (2013). Ensemble learning based essay automated scoring algorithm for Chinese English learners. *Journal of Chinese Information Processing, 27*(5), 100-107. https://doi.org/10.3969/j.issn.1003-0077.2013.05.014

Linacre, J. M. (1994). *Many-facet Rasch Measurement*. MESA Press: Chicago.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*(2), 103-122.

Linacre, J. M. (2005). *A User's Guide to FACETS: Rasch-Model Computer Program* (Computer program manual). MESA Press: Chicago.

Liu, J. (2010). A Many-Facet Rash Analysis of rater effects. *Modern Foreign Languages (Quarterly), 33*(2), 185-193.

McNamara, T. F. (1996). *Measuring Second Language Performance*. London: New York, Longman.

Mcnamara, T. F., & Lumley, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational 1 settings. *Language Testing, 14*(2), 140-156.

https://doi.org/10.1177/026553229701400202

Meiron, B. E., & Schick, L. S. (2000). Ratings, raters, and test performance: An exploratory study. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 153-176). Cambridge: Cambridge University.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189-227

Myford, C. M., & Dobria, L. (2006). *Facets Workshop*. Chicago.

Myford, C. M., & Wolfe, E. W. (2000). Monitoring sources of variability within the Test of Spoken English assessment system, (TOEFL Research Report No. 65). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2000.tb01829.x

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386-422.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413-428. https://doi.org/10.1037/0033-2909.88.2.413

Tian, Q. (2006). Application of Many-faceted Rasch Modeling in performance rating. *Journal of Psychological Exploration*, 70-73. https://doi.org/10.3969/j.issn.1003-5184.2007.01.014

Upshur, J. A., & Turner, C. E. (1999). Systematic effects in the rating of second-language speaking ability: test method and learner discourse. *Language testing, 16*(1), 82-111. https://doi.org/10.1177/026553229901600105

Wang, H. (2015). *Validating the automated scoring of the computer-based English listening and speaking test*. (Unpublished doctoral dissertation). Guangdong University of Foreign Studies, China.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287. https://doi.org/10.1191/026553298670883954

Wolfe, E. W., & Chiu, W. T. (1997). Detecting rater effects with a Multi-Faceted rating scale model. Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, March 25-27).

Wolfe, E. W., Moulder, B. C., & Myford, C. M. (2001). Detecting differential rater functioning over time (drift) using a rasch multi-faceted rating scale model. *Journal of Applied Measurement, 2*(3), 256-80.

Xi, X. (2010). Automated scoring and feedback systems—Where are we and where are we heading? *Language Testing, 27*, 291-300. https://doi.org/10.1177/0265532210364643

Xi, X., Higgins, D., Zechner, K. & Williamson, D. M. (2008). Automated scoring of spontaneous speech using SpeechRater v1.0 (ETS Research Report No. RR-08-62), Princeton, NJ: ETS. https://doi.org/10.1002/j.2333-8504.2008.tb02148.x

Xi, X., & Mollaun, P. (2009). How do raters from India perform in scoring the TOEFL iBT speaking section and what kind of training helps? (TOEFL iBT Research Report No. RR-09-31). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02188.x

Yang, H. (1999). Principles in designing college English oral test. *Foreign Language World, 03*, 48-57. https://doi.org/10.1088/1126-6708/1999/03/011

Yang, R. (2010). *A many-facet Rasch analysis of rater effects on an Oral English Proficiency Test*. (Doctoral dissertation). Purdue University.

Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., & Bhola, D. S. (2002). A review of strategies for validating computer automated scoring. *Applied Measurement in Education, 15*, 391-412. https://doi.org/10.1207/S15324818AME1504_04

Zhang, J. (2008). Quality control of subjective scoring using Many-Facet Rasch Model—A case of speaking test in PETS band 3. *Examinations Research, 4*(04), 67-80.

Zhang, M. (2013). *Contrasting automated and human scoring of essays*. Ets R & D Connections. https://doi.org/10.1002/j.2333-8504.2013.tb02325.x

Zhou, Y., & Zeng, Y. (2016). A Many-Facet Rasch Model analysis of Computer-based English Listening-Speaking Test. *Foreign Language Testing and Teaching, 21*(01), 24-33.

Zhou, M., Jia, Y., Zhou, C., & Xu, N. (2019). English automated essay scoring methods based on discourse structure. *Computer Science, 46*(03), 240-247.

Zou, S., & Chen, W. (2010). TEM4 scoring validity and computer-assisted scoring. *Computer-Assisted Foreign Language Education in China, 131*, 56-60.

Appendix A

Task One—Read Aloud in the CELST in 2013 Guangdong NMET

In 1939, on the eve on the Second World War, Albert Einstein wrote a letter to the American President. The letter was about an application of Einstein's famous equation, E equals MC squared, and his fear that the Nazis could use it to build an atomic bomb. E equals MC squared is the symbol of Einstein's genius. It's an equation that sums up one of the most powerful truths about the universe. It combines two ideas, which until Einstein came along, no one had ever dreamed could be connected in such a powerful way. The idea of mass and the idea of energy.

Appendix B

Task Two—Role Play in the CELST in 2013 Guangdong NMET

Part B Role Play

情景介绍(30") 角色：你是团队负责人 Tom。

任务：(1) 和前辈 Mary 谈论团队合作的问题；

(2)根据谈话内容回答同学的提问。

生词：teammates 团队成员

Tapescript:

W: Hi, Tom. You look worried.

M: I'm having trouble with my work.

W: What's the trouble?

M: Uh… I'm in charge of a team, working with five other people on a project. But we are not very productive.

W: Is the project too hard for everyone to work on?

M: No. But no one is working together on it. What do you think the problem is?

W: I guess your group is not good at team work.

M: But I don't know what to do about it. Everyone is just doing his or her own separate work on it.

W: That's too bad. If your group does not come together, then they will not know what is done and what needs to be done.

三问部分：

1. 我怎样才能使他们共同工作呢？

Question 1: How can I make them work together? / How do I make them work as a team? /

What can I do to make them work together?

Answer 1: You should call everyone together for a meeting and you should make them know that working as a team is important.

2. 你认为什么时侯召集开会最好呢？

Question 2: What do you think is best to call them for a meeting? / What do you think is the best time to call everyone for meeting? / When do you think is best to call a meeting?

Answer 2: As soon as possible. You should play the role of a monitor in the meeting, getting to know what they have done so far and finding out what still needs to be done. In this way, you will be sure that the project is done correctly.

3. 作为领导，我应该怎样对待团队成员呢？

Question 3: How should I treat my teammates as a leader? /

How should I treat my team members as a monitor?

Answer 3: You can't just tell them what to do or how to do things. A good leader listens to his teammates as well as directs them. By listening to your teammates, you get to know how much they have achieved and you know.

Questions to raise:

1. 我怎样才能使他们共同工作呢？

How do I get them to work together (as a team)? (语言 1 分，信息 1 分)

How can I make them work together? /

How do I make them work as a team? /

What can I do to make them work together?

2. 你认为什么时侯召集开会最好呢？

When/What do you think is the best time to call for a meeting?(语言 1 分，信息 1 分)

What do you think is best to call them for a meeting? /

What do you think is the best time to call everyone for meeting? /

When do you think is best to call a meeting?

3. 作为领导，我应该怎样对待团队成员呢？

As a leader, how should I treat my teammates? (语言 1 分，信息 1 分)

How should I treat my teammates as a leader? /

How should I treat my team members as a monitor?

Questions to answer:

1. How many people do you need to work with?

Five. (语言 1 分，信息 1 分)

2. What does Mary think the problem is?

(Mary thinks that) the group is not good at teamwork. (语言 1 分，信息 1 分)

3. What should your teammates know in the meeting?

To know that working as a team is important. (语言 1 分，信息 1 分)

4. What role should you play in the meeting?

A monitor. (语言 1 分，信息 1 分)

5. Why do you need to listen to your teammates?

(By listening to the teammates,) I can know how much they have achieved.

(语言 1 分，信息 1 分)

Appendix C

Task Three—Retelling in the CELST in 2013 Guangdong NMET

Part C: Retelling

你将听到一段独白，独白播两遍。听完独白后，用自己的话复述独白的内容。

梗概：Tom 的车后有团“黑云”，原来是蜂群跟随车旁的蜂王而形成的。

关键词：black cloud 黑云; queen bee 蜂王; police 警察; beekeeper 养蜂人; wheel 车轮

A Black Cloud

After leaving a village, Tom was driving to London. A strange noise made him stop. He got out and examined the car carefully, but he found nothing wrong. So he started driving again. The noise became louder and louder. Tom looked back quickly and saw a great black cloud following the car. When he stopped by a farm, a farmer told him that a queen bee must be hidden in his car, because there were thousands of bees following him. Tom drove away as quickly as he could. He thought it would be the best way to escape.

After one hour’s hard driving, he arrived in London. He parked his car outside a hotel and went in to have a drink. Several minutes later, he finished his drink and went out. He was surprised to find that his car was covered with bees. Then he called the police and explained what had happened. The police called a beekeeper. In a short time, the beekeeper arrived. He found the queen bee near the wheel at the back of the car. Finally the beekeeper put the queen bee and the other bees in a large box and took them home.

Appendix D

Rating Scale of Role Play in the CELST in 2013 Guangdong NMET

	score	standard
language	1	Correct grammatical structure and diction
	0.5	Basically correct grammatical structure and vocabulary, 0.5 point will be deducted for the following mistakes: Mistake on verb tense and voice Transitivity Subject-predicate consistency
	0	Severe error in grammatical and diction which leads to misunderstanding or incomprehension of the examinee’s production, such as: Error in interrogative Incomplete structure and incoherent sense
information	1	Information is delivered according to the requirements on the whole
	0.5	Part of the information is delivered
	0	Cannot deliver information according to the requirements

Appendix E

Rating Scale of Retelling in the CELST in 2013 Guangdong NMET

	Content		Comprehensive rating	
Rating method	Rating by point of information (POI), 1.5/POI		Comprehensive rating according to content, language, fluency and pronunciation	
Level	score	standard	score	Standard
A	1.5	Exactly the same	7-9	·faithful to the original text ·appropriate language conforming to the standard ·fluent expression ·pronunciation and intonation do not affect comprehension
B	0.5-1.0	Roughly correct	4-6	·basically retell the original text ·roughly appropriate language conforming to the standard (a bit of language mistake which doesn't affect comprehension) ·roughly fluent (short pause during retelling) ·slight mistakes on pronunciation and intonation but do not affect comprehension
C	0	Totally mistake	0-3	·cannot retell the main content of the original text ·poor language decency with many language errors ·non-fluent ·pronunciation and intonation affect comprehension
Note on information point	Information is focused and point will be guaranteed once information is delivered; Language error will be deducted in "comprehensive rating"			