Advances in Economics, Business and Management Research, volume 71
4th International Conference on Economics, Management, Law and Education (EMLE 2018)

A Study on Rater Reliability Under Holistic and Analytic Scoring of CEPT Writing by Using Generalizability Theory and Many-facet Rasch Model

Chun Lin
College of Foreign Languages
Hunan University
Changsha, China 410082

Yunnan Xiao
College of Foreign Languages
Hunan University
Changsha, China 410082

***Abstract*—To explore the scoring reliability of College English Placement Test (CEPT) writing, generalizability theory (GT) and many-facet Rasch model (MFRM) were applied to analyze 15 raters' holistic and analytic ratings of 300 writing samples. The results were as follows: (1) Raters' scoring method had significant impact on their rating of CEPT writing; under either method, one rater was enough to ensure that the generalizability coefficient was 0.8 or above. (2) Whichever method was adopted, raters differed significantly from each other in severity, but they had sound intra-rater consistency; raters were most biased towards task, then grammar, mechanics, vocabulary, and were least biased towards structure; if the dimensions at the discourse level were scored severely, then those at the linguistic level were scored leniently, and vice versa; raters tended to be severely biased towards the low, the intermediate level groups, and the lowest proficiency examinee, while be leniently biased towards the high level group, and the highest proficiency examinee. GT and MFRM proved from macro- and micro-levels respectively that the CEPT writing scoring enjoyed high reliability.**

Keywords—placement test; writing scoring reliability; holistic scoring; analytic scoring; generalizability theory; many-facet Rasch model

I. INTRODUCTION

Writing scoring reliability has always been the research focus in the field of language testing. During the writing scoring process, the rater, the examinee (essay), and the scoring method (rating scale) will interact more or less with one another, which is one of the sources of score discrepancy. The scoring method may not only affect the reliability of the scoring results, but also may damage the validity of the test [1], [2], [3]. The holistic scoring method and the analytic scoring method are the most commonly used scoring methods in writing test. Among studies on the holistic and the analytic scoring reliability of L2 writing raters, the situation of the majority of studies beyond China is that their participants are native English speaker (NES) raters [1], [4], [5] or that the number of raters and examinees selected for their experiments is small, while the situation of studies in China is that little research has been conducted into scoring reliability of College English Placement Test (CEPT) writing. A scientific and rational CEPT is the basic guarantee for implementing the stratified college English teaching that aims at teaching students in accordance with their ability. In order to better implement CEPT and the stratified college English teaching which are both based on the students' ability and needs, it is a must to ensure the reliability and validity of placement test results, while the reliability and validity of the whole placement test are largely determined by the scoring of the subjective test, namely, writing test, therefore it is necessary to deeply study the holistic and the analytic scoring variability of raters in CEPT writing test, and also to monitor and correct, then to minimize such variability, finally to achieve greater fairness and accuracy in scoring.

Studies on the reliability of writing raters in the field of language testing mostly adopt quantitative methods based on classical testing theory, such as *t*-test [6], two-way ANOVA with repeated measures [5], three-way ANOVA [7], [8], MANOVA [9], ANOVA and MANOVA [10], but these quantitative methods do not provide the micro details that cause the score discrepancy. Generalizability (G-) theory has great advantages in exploring the source of measurement error and defining the measurement situation and generalization range. There are studies which use G-theory to investigate the reliability of writing raters in the field of language testing, for instance, reference [11] applied G-theory to explore to the rater reliability and item effect in the scoring of twenty high school students' three compositions of different forms; reference [12] used G-theory to examine the task and rater effects in second language speaking and writing; reference [13] adopted G-theory to investigate how raters and tasks influenced the reliability of writing scoring for 211 children in Grades 3 and 4. Many-facet Rasch model has prominent advantages in estimating parameters at the individual level of the test scenario. It can analyze on the same logit (log odds units) scale different facets, such as the rater severity, examinee ability and item difficulty, and can provide fitness analysis and bias analysis. Through bias analysis, researchers can accurately and effectively identify the source of the rater bias that affects scoring reliability [14]. A lot of researches have applied many-facet Rasch model to study the reliability of writing raters [15], [16], [17], and mainly to study the two most direct sources of scoring errors: the interaction between

rater and item [18], [19], [20], [21], [22], [23], and the interaction between rater and examinee [14], [20], [21], [24]. Although there are a few studies using G-theory and many-facet Rasch model to study the reliability of writing scoring, for instance [25], [26], [27], few studies have made combination of these two models to analyze the reliability of writing raters under different scoring methods.

The scoring reliability of CEPT raters under different scoring methods is related to the scoring validity, which is in turn related to the equity of education opportunity and the effectiveness of teaching, therefore the problem of scoring reliability of CEPT raters should be solved urgently. In this paper, G-theory is applied to mainly address the following two questions: What are the main sources of variability of score discrepancy under the two scoring methods? What is the scoring reliability, i.e., generalizability (G-) coefficient of raters under the two scoring methods? While many-facet Rasch model is used to mainly address the following two questions: What bias patterns do raters display towards dimensions? What bias patterns do raters display towards examinees?

II. METHOD

A. Raters

The raters were all from the foreign language college of a Chinese university that participates in China's construction plan of world-class universities and first-class disciplines. All raters had rich experience in teaching and scoring English as a Foreign Language (EFL) writing. There were altogether 15 raters, of whom one was a College English teacher, and the rest 14 were postgraduate students majoring in language testing.

The NULL hypothesis of our experiment is that because all the raters were experienced raters, in this case intuitively their holistic and analytic scorings would show no significant difference.

B. Rating Scales

The original holistic scale "Entry Level Writing Characteristics" (hereinafter referred to as ELWC) was designed based on Canadian Language Benchmarks, diving writing ability into 6 levels, namely, 010, 020, 030, 040, 050, 060. There were 5 intervals between these 6 levels. Set the middle point of the intervals as cutting points, then there would be 11 sub-levels in total. The scale contained five dimensions, namely, task, grammar, mechanics, vocabulary, and structure. For the convenience of analysis, we took each sub-level as 1 point, then we would get 11 points in total.

In order to achieve the purpose of this study, without changing the basic construct of the original ELWC scale, we converted it into an analytic scale. The analytic scale also consisted of five dimensions. Each dimension was also divided into 6 levels according to the original scale, from 010 to 060, therefore added up with 5 sub-levels there were 11 sub-levels. In order to facilitate analysis, we set each sub-level as 1 point, and add up the total points of five dimensions and then converted the total points to the final score.

C. Samples

It was a single topic writing task. The writing samples were 300 essays selected by stratified sampling from those 4861 compositions of the computer-aided CEPT writing test of the very university. Sampling was based on the original scores given by their English experts and teachers. Because few students reached the 060 level, which was the proficiency level of the score 11, no writing samples representing the proficiency level of the score 11 were chosen. From 010 to 055, there were altogether 10 sub-levels, with 30 samples selected from each sub-level, then 300 samples would be obtained.

D. Rater Training

Rater training included the following steps. At first, raters spent time in studying the holistic and the analytic scales. Next, raters were asked to score 10 anchor writing samples. In the following, raters' scorings were compared with the true level of individual sample. At last, raters made group discussion on their scorings. It took one and half days to holistically grade all 300 samples. And it took another eight and half days to analytically grade all 300 samples. Between the two scoring periods, the raters were given a five-day break for the sake of reducing fatigue and possible memory effect.

E. Design

This study will examine two major factors that may have a significant impact on writing scoring: the rater and the scoring method. G-study within G-theory framework was divided into two steps. At the first step, for all the scoring data, the random double-facet crossed design $p \times m \times r$ was used, in which the measuring object p was the examinee's writing ability, m was the facet of scoring method with 2 levels, and r was the facet of rater with 15 levels. Both the facet of the rater and that of the scoring method were random facets. At the second step, for the different data of the two scoring methods, the random single-facet crossed design $p \times r$ was adopted, in which the measuring object p was the examinee's writing ability, and the random facet r was the facet of rater with 15 levels.

Many-facet Rasch analysis used a three-facet (rater, examinee, item) measurement model. This measurement model can be represented by the mathematical model

$$\text{Log}\,(P_{nijk/}\,P_{nij(k\text{-}1)}) = B_n - D_i - C_j - F_k \qquad (1)$$

In (1), P_{nijk} is the probability that candidate n will be scored k by rater j on item i, $P_{nij(k\text{-}1)}$ is the probability that candidate n will be scored k-1 by rater j on item i, B_n is the ability of examinee n, D_i is the difficulty of item i, C_j is the severity of rater j, and F_k is step difficulty from score k-1 to score k. All the 15 raters holistically and analytically scored the 300 samples, therefore in holistic scoring there were 15×300=4500 data, and in analytic scoring 15×300×5=22500 data. These two sample data sizes were in line with the standard (a minimum of 1152) proposed by Linacre [28].

F. Software

GENOVA 3.1 [29] and FACETS 3.71.4 [28] were used for data processing.

III. RESULTS AND DISCUSSION

The G-theory research in this section addressed the two issues: (1) the main sources of variability of scoring difference between raters under two scoring methods, and (2) scoring reliability of raters under two scoring methods. While many-facet Rasch analysis solved the two problems: (1) the bias pattern of rater-dimension interaction, and (2) the bias pattern of rater-examinee interaction.

A. G-theory Research

The G-study results of *p x m x* r random effects are shown in "Table I". It is not hard to find that the largest source of variability was paper-by-method (*pm*), the second largest was the residual (*pmr*), the third largest was paper (*p*), followed by four effects which accounted for less than 5% of the total variance, from large to small, in turn, namely rater (*r*), method-by-rater (*mr*), paper-by-rater (*pr*), method (*m*). Among all the variance components, the variance component yielded by paper-by-method explained 62.61% of the total variance, indicating that there was a big difference between the scores obtained by using two scoring methods to evaluate the same essay. The variance component yielded by the residual accounted for 17.30% of the total variance, suggesting that some other random, systematic or unsystematic, and unmeasured facets had not been explained. The variance component yielded by paper explained 15.04% of the total variance, indicating that the writing test had accuracy in measuring the examinee's writing ability. The variance component yielded by rater only accounted for 3.20% of the total variance, which meant that raters differed in terms of severity. The variance component yielded by method-by-rater was small, only explaining 0.95% of the total variance, implying that no matter what scoring method the raters used, there were more or less differences in the severity of the raters. The variance component yielded by paper-by-rater merely interpreted 0.90% of the total variance, indicating that the paper-by-rater effect did not contribute much to the total variance. The variance component yielded by method explained 0.00%, suggesting that scoring methods were consistently used from beginning to end.

TABLE I. VARIANCE COMPONENTS OF DIFFERENT SOURCES OF VARIABILITY FOR A RANDOM EFFECTS *P X M X R* DESIGN

Source of Variability	df	σ^2	S.E.	%
p	299	0.4767	0.1467	15.04%
m	1	0.0000	0.0009	0.00%
r	14	0.1014	0.0419	3.20%
pm	299	1.9843	0.1647	62.61%
pr	4186	0.0284	0.0089	0.90%
mr	14	0.0300	0.0113	0.95%
pmr	4186	0.5484	0.0120	17.30%
Total	8999	3.1692	0.3864	100.00%

The G-study results of *p x r* random effects are shown in "Table II". "Table II" shows that the largest source of variability for both holistic scoring and analytic scoring was paper (*p*), which indicated that both methods were accurate in measuring the examinees' writing ability. Comparatively speaking, analytic scoring method was more accurate, because the variance component of its measurement object namely paper (*p*) was slightly larger. The variance components of rater (*r*) for both holistic and analytic scoring were less than 5%, indicating that no matter which scoring method the raters used, there were some differences in terms of rater severity, while comparatively speaking, the raters were more consistent when using analytic scoring method because under this method their variance component was slightly smaller.

TABLE II. VARIANCE COMPONENTS OF DIFFERENT SOURCES OF VARIABILITY FOR A RANDOM EFFECTS *P X R* DESIGN

Scoring Method	Source of Variability	df	σ^2	S.E.	%
Holistic Scoring	*p*	299	2.4792	0.2054	76.82%
	r	14	0.1466	0.0526	4.54%
	pr	4186	0.6015	0.0131	18.64%
	Total	4499	3.2273	0.2711	100.00%
Analytic Scoring	*p*	299	2.4427	0.2021	78.52%
	r	14	0.1161	0.0417	3.73%
	pr	4186	0.5522	0.0121	17.75%
	Total	4499	3.1110	0.2559	100.00%

D-study within G-theory framework estimates the relationship between the number of raters and the G-coefficient [30]. The comparison between the raters' holistic and analytic scoring reliability is shown in "Table III". "Table III" shows that as the number of raters increased, so did the G-coefficient of the raters; when the number of raters was the same, the G-coefficient of the analytic method was generally higher than that of the holistic method. Taking this study as an example, if 15 raters were used, the G-coefficient of the holistic method was 0.98, and that of the analytic method was 0.99. In view of the economic cost of large-scale examination scoring and the feasibility of operational practice, in the norm-referenced test, one rater is enough to guarantee the G-coefficients of both scoring methods to reach 0.8 or above.

TABLE III. THE RELATIONSHIP BETWEEN THE NUMBER OF RATERS AND THE CHANGE OF G-COEFFICIENT UNDER TWO SCORING METHODS

Number of Papers	Number of Raters	G-Coefficients	
		Holistic Scoring	***Analytic Scoring***
300	1	0.80	0.82
300	2	0.89	0.90
300	3	0.93	0.93
300	15	0.98	0.99

B. Many-facet Rasch Analysis

1) Rater consistency: "Table IV" shows that there were significant differences in severity between CEPT writing raters no matter which method was used. When 15 raters used the holistic scoring method, the rater severity separation ratio was 8.42, and the separation reliability was 0.99, which meant the difference between the raters' severity was significant. The chi-square test showed that the fixed (all-same) chi-square value=1111.6, the degree of freedom=14, and p< 0.00,

indicating that there was a significant difference in severity among the raters. When 15 raters used the analytic scoring method, the rater severity separation ratio=7.60, and the separation reliability=0.98, implying that there was significant different in severity between raters. The chi-square test showed that the fixed (all-same) chi-square value=950.8, the degree of freedom=14, and p< 0.00, suggesting that the severity between raters was statistically significant different. The mean of all the 15 raters' severity was 0.30 logits under holistic scoring and 0.26 logits under analytic scoring, indicating that the raters were slightly lenient when using the analytic score. This finding is contrary to that of [20].

"Table IV" shows that, regardless of the method used, all CEPT writing raters had good internal consistency, except for the five raters, R1, R4, R6, R7, and R15. Raters' internal consistency is reflected by infit mean square. The CEPT test in this study belongs to intermediate risk test, and the range of acceptable infit mean square value of the subjective scoring of low and intermediate risk test is from 0.6 to 1.4 [31]. Infit mean square being larger than 1.4 means that the data is not fit with the model, while it being smaller than 0.6 means that the data is excessively fit with the model, i.e., the raters use some scores of the rating scale in a concentrated way. Under holistic scoring, R15 underfit the model, meanwhile, under analytic scoring, R6, R7, and R15 underfit the model. Under analytic scoring, R1 and R4 overfit the model with infit mean square values both less than 0.6, indicating that their analytic scoring showed a central tendency. It was worth noting that holistic scoring did not show central tendency, and under holistic scoring the score probability curve showed that each score had a sharp peak, suggesting that holistic rater could well grasp the difference of these scores. (See "Fig. 1")

TABLE IV. COMPARISON OF THE RATER CONSISTENCY UNDER TWO DIFFERENT SCORING METHODS

Rater	Holistic Scoring					Analytic Scoring				
	Severity	*Infit*		*Outfit*		*Severity*	*Infit*		*Outfit*	
		MnSq	*ZStd*	*MnSq*	*ZStd*		*MnSq*	*ZStd*	*MnSq*	*ZStd*
1	0.89	1.07	0.7	1.08	0.9	0.36	**0.57**	-6	0.64	-4.8
2	0.54	0.97	-0.3	0.94	-0.6	0.58	1.18	1.9	1.15	1.6
3	-0.1	0.79	-2.6	0.75	-3.3	0.1	0.93	-0.8	0.89	-1.3
4	0.76	0.7	-3.8	0.73	-3.5	0.56	**0.59**	-5.6	0.59	-5.5
5	1.1	1.04	0.4	0.96	-0.4	0.73	0.69	-4	0.66	-4.4
6	-0.84	1.18	2.1	1.15	1.7	-0.07	**1.6**	5.9	1.64	6.4
7	-0.3	0.98	-0.2	0.93	-0.8	0.06	**1.49**	4.9	1.43	4.5
8	0.85	1.15	1.7	1.13	1.5	0.11	0.81	-2.3	0.78	-2.7
9	0.66	0.78	-2.7	0.73	-3.5	1.02	0.6	-5.4	0.59	-5.5
10	0.88	0.74	-3.3	0.72	-3.6	0.62	0.75	-3.2	0.74	-3.2
11	1.07	1.13	1.4	1.12	1.3	1.07	0.78	-2.7	0.74	-3.2
12	-0.1	0.79	-2.6	0.75	-3.3	0.34	0.73	-3.4	0.73	-3.5
13	0.49	1.24	2.6	1.18	2	0.46	0.83	-2	0.86	-1.7
14	-0.69	0.66	-4.5	0.68	-4.4	-0.53	1.13	1.4	1.08	0.9
15	-0.68	**1.88**	8.4	1.84	8.3	-1.51	**2.27**	9	2.35	9
Mean	0.3	1.01	-0.2	0.98	-0.5	0.26	1	-0.8	0.99	-0.9
S.D.	0.66	0.3	3.2	0.29	3.3	0.62	0.46	4.4	0.47	4.4
	Separation: 8.42, Reliability: 0.99 Fixed (all same) chi-square: 1111.6, d.f.: 14, significance: p<0.00					Separation: 7.60, Reliability: 0.98 Fixed (all same) chi-square: 950.80, d.f.: 14, significance: p<0.00				

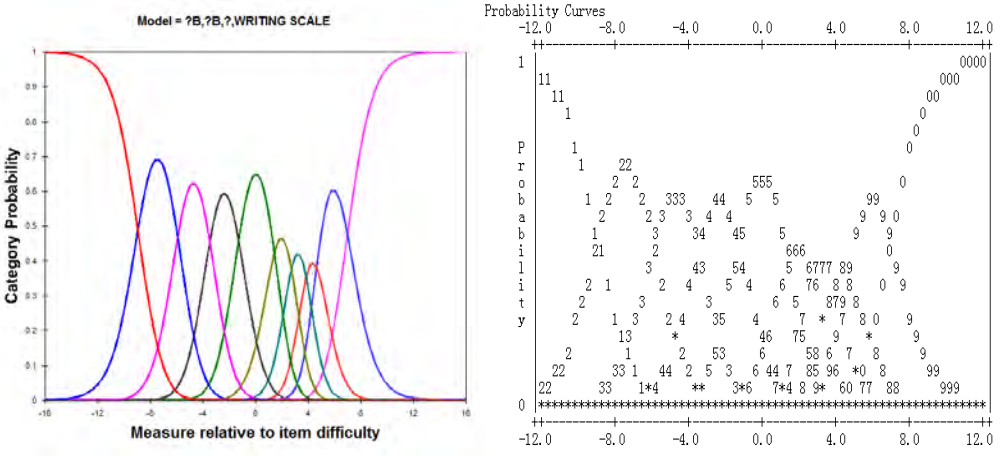

Fig. 1. Score probability curve under holistic scoring.

2) Rater-dimension bias analysis: Rater-dimension bias analysis aims to evaluate how the raters' self-consistency performs on five dimensions of analytic rating scale. As can be known from sub-section 3.2.1, in this study, the self-consistency of R1, R 4, R6, R7 and R15 is poor, so the scorings of these five raters will not be considered in the following bias analysis. "Table V" shows the significant rater-dimension bias frequency statistics, merely listing the significant rater-dimension interaction whose z-score is greater than +2 or smaller than -2 [32]. This table is arranged in the ascending order of rater number. Across the top of the table are the name of each dimension. Immediately below is the difficulty of each dimension represented by logit measure. At the third line is the sort of rater-dimension bias: severe or lenient. Next comes the frequency statistics of the severe or the lenient biases each rater shows towards each dimension, with 1 signifying existence and 0 signifying absence.

"Table V" shows that raters produced 8 biases on task, 6 biases on grammar, mechanics and vocabulary respectively and 5 biases on structure. Generally, in terms of the amount of biases, raters displayed the most biases towards task, while the least towards structure. In terms of total number, 10 raters produced 31 significant biases (17 severe ones and 14 lenient ones) towards the five different dimensions, which accounted for 0.21% of the total rater-dimension interaction (10×300×5=15000). Two raters (R8 and R10) displayed significant biases towards only one dimension, and the other eight raters showed significant biases toward two or more dimensions, for instance, R9 and R11 displayed biases towards all the five dimensions and they tended to show more severe biases than lenient ones in their scoring. This result verifies the rater severity listed in "Table IV", i.e., under analytic scoring, R11 was the most severe grader, and R9 came the second.

"Table V" reveals some systematic rater bias patterns in rater-dimension interaction: when a rater displayed significant biases towards the writing's macro level and micro level at the same time, he who was severe at the discourse level would be lenient at the linguistic level; on the contrary, i.e., he who was severe at the linguistic level would be lenient at the discourse level. Task and structure are related to macro requirements on the discourse level, while mechanics and vocabulary are related to micro requirements on the linguistic level. "Table V" shows that if a rater displayed significant biases towards task and mechanics at the same time, his biases towards task and his biases towards mechanics went reverse, i.e., if a rater was severely biased towards task, then he was leniently biased towards mechanics (R3, R11, and R14), on the contrary, if a rater was lenient towards task, then he was severe towards mechanics (R5 and R9); if a rater displayed significant biases towards structure and vocabulary at the same time, his biases towards structure and his biases towards vocabulary went reverse, i.e., if a rater was severely biased towards structure, then he was leniently biased towards vocabulary (R5 and R9), on the contrary, if a rater was lenient towards structure, then he was severe towards vocabulary (R3 and R11). This finding verifies observed by [14] and [19] the "rater compensation strategies", which says that a rater frequently subliminally decides to make compensations for being over-severe or over-lenient with any particular rating dimension.

TABLE V. FREQUENCY STATISTICS OF SIGNIFICANT RATER-DIMENSION BIASES

Dimension	Task	Grammar	Mechanics	Vocabulary	Structure	Total (S/L)	Sum
Logit	-0.09	0.02	-0.16	0.22	0.01		
Rater	S/L[a]	S/L	S/L	S/L	S/L		
R 2			0/1		1/0	1/1	2
R 3	1/0		0/1	1/0	0/1	2/2	4
R 5	0/1		1/0	0/1	1/0	2/2	4
R 8	0/1					0/1	1
R 9	0/1	1/0	1/0	0/1	1/0	3/2	5
R 10		1/0				1/0	1
R 11	1/0	1/0	0/1	1/0	0/1	3/2	5
R 12	1/0	0/1		1/0		2/1	3
R 13	0/1	1/0				1/1	2
R 14	1/0	1/0	0/1	0/1		2/2	4
Total (S/L)	4/4	5/1	2/4	3/3	3/2	17/14	31
Sum	8	6	6	6	5	31	

[a] S=Severe; L=Lenient.

3) Rater-examinee bias analysis: Rater-examinee bias analysis purposes to investigate how the raters' self-consistency performs across examinees' ability levels. In this study, there were altogether 833 significant rater-examinee bias interactions, from which 370 ones displayed by the misfitting raters (R1, R4, R6, R7, and R15) should be eliminated, then 463 ones produced by the rest 10 raters were left. These 463 significant bias interactions between 10 raters and examinees with different writing abilities were sorted out and the frequency statistics of significant rater-examinee biases is shown in "Table VI". "Table VI" is arranged in the ascending order of rater number. Across the top of "Table VI", taking the mean value of the examinees' logit measures (-0.03 logits) as the center and the standard deviation (2.12) as an interval, the examinees are divided into three groups: Group 1 (low level group) are the examinees whose logit measures are at least one standard deviation less than the mean value; Group 2 (intermediate level group) are the examinees whose logit measures are within the range of one standard deviation less than the mean value and one standard deviation larger than the mean value; Group 3 (high level group) are the examinees whose logit measures are at least one standard deviation larger than the mean value. To examine how the raters interacted with the lowest and the highest proficiency examinees, the ratings of the two examinees, namely E54 and E4, were listed out singly. In the second line of "Table VI" is the amount of examinees belonging to the corresponding ability group. Below comes the sort of bias (severe or lenient one) the raters exhibited. The body of "Table VI" displays the number of severe or lenient bias a rater produces towards a specific ability group. The last row and the last column are the sum of frequency of each column and each row respectively.

Raters tended to be severely biased towards the low, intermediate level groups, and the lowest proficiency examinee, while be leniently biased towards high level group, and the highest proficiency examinee. "Table VI" shows that there were altogether 463 significant rater-examinee bias interactions (257 severe ones and 206 lenient ones), accounting for 15.43% of all the biases that the 10 raters displayed towards all the 300 examinees (10×300=3000). With the exception of R14, the vast majority of raters presented more severe biases than lenient ones. When the same rater evaluated the essays of the low and the high level groups at the same time, his bias interactions towards these two level groups presented a completely different tendency, i.e., if he was severely biased towards one level group, then he would be leniently biased towards the other one. Six raters' (60% of the raters) biases towards the low and the high level groups go reverse. For example, R2, R10, R11, and R12 showed more severe biases towards the compositions of examinees in the low level group, while showed more lenient biases towards the compositions of examinees in the high level group; the bias pattern of R8 and R14 and that of R2, R10, R11, and R12 went reverse.

The findings of this study are different from those of [14]: he studied the scoring reliability of native English-speaker raters, and found that the raters were lenient towards the compositions of low level examinees. The findings of this study also have some distinctions from those of [20]: her research revealed that raters tended to grade the compositions of both the lowest and the highest proficiency examinees severely. In this study, raters exhibited more severe biases towards the low level group and more lenient biases towards the high level group. There may be two reasons: firstly, low level compositions are characterized by short passage lengths, unfinished task, unintelligible/ ungrammatical sentences, inappropriate use of mechanics, few low-frequency vocabulary, and bad structure, which makes it easier for the raters to grade them severely and give them a low score decisively; what is more, it may be that the raters had lower expectations of the high level writers out of the freshmen, so they may be more lenient in scoring and gave higher marks to the higher level compositions.

TABLE VI. FREQUENCY STATISTICS OF SIGNIFICANT RATER-EXAMINEE BIASES

Examinee (Logit) / Rater	E54 -7.66	Group 1 ⩽−2,16	Group 2 -2.15~2.09	Group 3 ⩾2.10	E4 3.06	Total (S/L)	Sum
No. of examinee	1	40	226	32	1		
Severe/Lenient	S/L	S/L	S/L	S/L	S/L		
R 2		13/0	20/25	4/8		37/33	70
R 3		2/1	20/25	5/0		27/26	53
R 5		5/0	11/10	3/1		19/11	30
R 8		0/4	17/13	8/0		25/17	42
R 9			9/14	6/0		15/14	29
R 10		5/0	21/10	1/7		27/17	44
R 11		3/0	20/13	0/3	0/1	23/17	40
R 12		3/1	19/7	1/6		23/14	37
R 13	1/0	1/0	27/21	3/3		32/24	56
R 14		0/1	28/22	1/10		29/33	62
Total (S/L)	1/0	32/7	192/160	32/38	0/1	257/206	463
Sum	1	39	352	70	1	463	

IV. CONCLUSION

Based on the G-theory study and many-facet Rasch analysis of 15 raters' holistic and analytic ratings of 300 CEPT compositions, this study obtained abundant data on the scoring reliability of CEPT writing. The main findings are as follows:

Firstly, raters' scoring method had significant impact on their rating of CEPT writing, but some of the residual was still unexplained, and the unexplained residual indicated that the raters' scoring reliability might well be affected by other factors, such as raters' educational background, raters' tolerance of language errors, and so on. In the norm-referenced test, under either method, one rater was enough to ensure that the generalizability coefficient was 0.8 or above.

Secondly, whichever method was adopted, raters differed significantly from each other in severity, but they had sound intra-rater consistency. Raters were most biased towards task, then grammar, mechanics, vocabulary, and were least biased towards structure. When a rater displayed significant bias towards the writing's macro level and micro level at the same time, he who was severe at the discourse level would be lenient at the linguistic level, and vice versa. Raters tended to be severely biased towards the low, intermediate level groups, and the lowest proficiency examinee, while be leniently biased towards the high level group, and the highest proficiency examinee.

Based on the above empirical results, the following two suggestions for rater training are put forward. Firstly, it is hoped that raters' intra-reliability can be enhanced through training methods such comparison method, think-aloud protocol, and interview. Taking this study as an example, under analytic scoring, R1 and R4 should be trained through these methods to use the whole range of scores to grade the examinees with reference to the examinees' writings. Secondly, it is expected that the number of rater-dimension/ rater-examinee biases will be decreased through individualized rater training, during which raters are divided into several groups based on their scoring traits and each group are delivered different training contents. In the case of

this study, according to the raters' bias patterns at the macro discourse level and the micro linguistic level, R3, R11, and R14 can be classified into one group, while R5 and R9 another group. According to the raters' bias patterns with the low level group and the high level group, R2, R10, R11, and R12 can be classified into one group, while R8 and R14 another group.

This paper explored only one of the rater-related factors that might affect the scoring reliability, that is, raters' scoring method, but did not dig out raters' scoring patterns with reference to other rater-related factors, such as age, gender, or personality and so on, and was thus not possible to provide further qualitative data to explain the differences in scoring reliability. Future studies may further explore other factors that cause scoring discrepancy.

REFERENCES

[1] N. Bacha, "Writing evaluation: what can analytic versus holistic essay scoring tell us? ", System, vol. 29, pp. 371-383, March 2001.

[2] K. Barkaoui, "Rating scale impact on EFL essay marking: a mixed-method study", Assessing Writing, vol. 12, pp. 86-107, October 2007.

[3] U. Knoch, "Diagnostic assessment of writing: a comparison of two rating scales", Language Testing, vol. 26, pp. 275-304, April 2009.

[4] N. T. Carr, "A comparison of the effects of analytic and holistic rating scale types in the context of composition tests", Issues in Applied Linguistics, vol. 11, pp. 207-241, December 2000.

[5] B. Song and I. Caruso, "Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? ", Journal of Second Language Writing, vol. 5, pp. 163-182, May 1996.

[6] R. Sheorey, "Error perceptions of native-speaking and nonnative speaking teachers of ESL", ELT Journal, vol. 40, pp. 306-312, April 1986.

[7] J. D. Brown, "Do English and ESL faculties rate writing samples differently? ", TESOL Quarterly, vol. 25, pp. 587-603, Winter 1991.

[8] T. Kobayashi, "Native and nonnative reactions to ESL compositions", TESOL Quarterly, vol. 26, pp. 81-112, Spring 1992.

[9] L. Shi, "Native-and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing", Language Testing, vol. 18, pp. 303-325, July 2001.

[10] C. A. Jolivet, "Comparing native and nonnative speakers' error correction in foreign language writing", Texas Papers in Foreign Language Education, vol. 3, pp. 1-14, Fall 1997.

[11] Y. Liu and H. Zhang, "Application of generalizability theory in composition scoring", Acta Psychologica Sinica, vol. 30, pp. 211-218, April 1998.

[12] Y. In'nami and R. Koizumi, "Task and rater effects in L2 speaking and writing: a synthesis of generalizability studies", Language Testing, vol. 33, pp. 341-366, June 2015.

[13] Y.-S. G. Kim, C. Schatschneider, J. Wanzek, B. Gatlin, and S. Al Otaiba, "Writing evaluation: rater and task effects on the reliability of writing scores for children in Grades 3 and 4", Reading and Writing, vol. 30, pp. 1287-1310, February 2017.

[14] E. Schaefer, "Rater bias patterns in an EFL writing assessment", Language Testing, vol. 25, pp. 465-493, October 2008.

[15] D. R. Isbell, "Assessing C2 writing ability on the Certificate of English Language Proficiency: rater and examinee age effects", Assessing Writing, vol. 34, pp. 37-49, September 2017.

[16] S. Liu and J. Zhang, "Multistage rating augmentation - an effective way to improve subjective performance rating", Psychological Exploration, vol. 35, pp. 266-271, June 2015.

[17] J. Wang, G. Engelhard, Jr., K. Raczynski, T. Song, and E. W. Wolfe, "Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach", Assessing Writing, vol. 33, pp. 36-47, April 2017.

[18] T. Eckes, "Rater types in writing performance assessments: a classification approach to rater variability", Language Testing, vol. 25, pp. 155-185, April 2008.

[19] T. Eckes, "Operational rater types in writing assessment: linking rater cognition to rater behavior", Language Assessment Quarterly, vol. 9, pp. 270-292, August 2012.

[20] W. Huang, "Exploration of rater bias patterns of Chinese university EFL teachers with multi-faceted Rasch measurement", Foreign Language Education, pp. 162-169, 2010.

[21] K. Kondo-Brown, "A FACETS analysis of rater bias in measuring Japanese second language writing performance", Language Testing, vol. 19, pp. 3-31, January 2002.

[22] J. Liu and M. Yang, "Quality control for rating in a performance test", Technology Enhanced Foreign Language Education, pp. 26-32, January 2010.

[23] T. Lumley, "Assessment criteria in a large-scale writing test: what do they really mean to the raters? ", Language Testing, vol. 19, pp. 246-276, July 2002.

[24] Y. Wang, Z. Zhu, and H. Yang, "Many-facet Rasch analysis of the reliability of online essay marking", Foreign Language World, pp. 69-76, February 2006.

[25] D. Guan, "Study of writing scoring quality in national postgraduate entrance exam with generalizability theory and many-facet Rasch model", Psychological Exploration, vol. 34, pp. 437-440, October 2014.

[26] H. Li, "Evaluation of the reliability of CET-6 essay scoring using generalizability theory and many-facet Rasch model", Foreign Languages and Their Teaching, pp. 51-56, October 2011.

[27] R. R. Sudweeks, S. Reeve, and W. S. Bradshaw, "A comparison of generalizability and many-facet Rasch measurement in an analysis of college sophomore writing", Assessing Writing, vol. 9, pp. 239-261, January 2005.

[28] J. Linacre, Many-facet Rasch Measurement. Chicago: MESA Press, 1994.

[29] J. E. Crick and R. L. Brennan, GENOVA: A General Purpose Analysis of Variance System. Iowa City, IA: American College Testing Program, 1983.

[30] G. Li and M. Zhang, "Analysis of cross-distribution for estimating variance components in generalizability theory", Psychological Development and Education, vol. 28, pp. 665-672, November 2012.

[31] B. Wright and J. Linacre, "Reasonable mean-square fit values", Rasch Measurement Transactions, vol. 8, pp. 370, 1994.

[32] T. F. McNamara, Measuring Second Language Performance. New York: Longman, 1996.