Evaluating rater judgments on ETIC Advanced writing tasks: An application of generalizability theory and Many-Facets Rasch Model

Jiayu Wang & Kaizhou Luo
National Research Centre for Foreign Language Education,
Beijing Foreign Studies University, China

Developed by China Language Assessment (CLA), the English Test for International Communication Advanced (ETIC Advanced) assesses one's ability to perform English language tasks in international workplace contexts. ETIC Advanced is only composed of writing and speaking tasks, featured with authentic constructed response format. However, the elicitation of extended responses from candidates would call for human raters to make judgments, thus raising a critical issue of rating quality. This study aimed to evaluate rater judgements on the writing tasks of ETIC Advanced. Data in the study represented scores from 186 candidates who performed all writing tasks: Letter Writing, Report Writing, and Proposal Writing (*n*=3,348 ratings). Rating was conducted by six certified raters based on a six-point three-category analytical rating scale. Generalizability theory (GT) and Many-Facets Rasch Model (MFRM) were applied to analyse the scores from different perspectives. Results from GT indicated that raters' inconsistency and interaction with other aspects resulted in a relatively low proportion of overall score variance, and that the ratings sufficed for generalization. MFRM analysis revealed that the six raters differed significantly in severity, yet remained consistent in their own judgements. Bias analyses indicated that the raters tended to assign more biased scores to low-proficient candidates and the Content category of rating scale. The study serves to demonstrate the use of both GT and MFRM to evaluate rater judgments on language performance tests. The findings of this study have implications for ETIC rater training.

Key words: ETIC Advanced, rater judgements, generalizability theory, Many-Facets Rasch Model

Email address for correspondence: kevinlkz@bfsu.edu.cn

Introduction

A defining characteristic of language performance tests is that actual use of language to perform real-world tasks is required of candidates, rather than mere demonstration of language knowledge, often by means of choosing from options (McNamara, 1996). Performance tasks are often considered to establish authenticity, as the task format and the resulting performance share more similarities with those in real-world situations (Bachman, 1990; Bachman & Palmer, 1996).

Nonetheless, the results of performance tasks may face various threats from task design, interlocutors, rating scales, raters, etc. (e.g. Barkaoui, 2010; Eckes, 2005; McNamara, 1996). Among these sources, raters often exert a profound influence on the final scores since they evaluate the resulting performance, use a rating scale, and ultimately assign scores to test-takers. Unlike 'rating machines', raters may differ from each other in their ratings (Eckes, 2005; Yan, 2014), demonstrate inconsistent severity throughout the rating process (e.g. Myford & Wolfe, 2003), and systematically interact with other facets (Kondo-Brown, 2002; Schaefer, 2008; Upshur & Turner, 1999). These inappropriate judgements may hinder the interpretation and use of test scores, thus further increasing concerns in fairness.

Regarding the importance of rater facet, researchers have conducted studies to reduce raters' erroneous judgements, in which a major concern is rater training. In some studies, training programs have contributed to consistent and accurate rater judgments (Davis, 2016; Lim 2011; Lunz, Wright, & Linacre, 1990; Weigle, 1998, 1999; Wigglesworth, 1993), yet the effectiveness of rater training is far from conclusive. Research also indicates that raters still exhibited significant severity differences after trainings of various forms (e.g. Davis 2016; Eckes, 2005; Kondo-Brown, 2002); their systematic interaction with other facets remained after training programs (Knoch, 2011; Kondo-Brown, 2002; Goodwin, 2016; Schaefer, 2008; Youn, 2018); and the effects of rater training may also decrease with the elapse of time (Lumley & McNamara 1995; Shaw, 2002). The mixed results suggest that the effect of rater training may be contextualised, and the condition of rater judgements after training may vary according to test types, training procedures, raters, candidates, tasks, and rating scales. It is thus necessary to evaluate rater judgements, despite training programs, to ensure the quality of final scores in language performance tests. Generalizability theory and many-facets Rasch model are commonly used to conduct such evaluations.

Evaluating rater judgments through GT

Derived from Classical Test Theory (CTT) and ANOVA, Generalizability Theory (GT) distinguishes itself on a unique conceptual framework (Brennan, 2001). Test scores in GT are treated as cases from a universe of testing conditions; generally, higher reliability indicates a higher generalizability of scores to other testing contexts (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Thus, generalizability of scores depends not only on the facets within a specific test, but also on the contexts out of the test, pertaining to the interpretations and decisions made based on the test results (Bachman, 1990).

GT analysis consists of generalizability study (G study) and Decision study (D study). G study deconstructs sources of variance (e.g. tasks, raters or categories) into variance components and evaluates their contribution to score variance in a single observation (Bachman & Kunnan, 2005). Drawing on the information from G study, D study evaluates the generalizability of the test results from operational testing design as well as alternative designs (Cardinet, Johnson, & Pini, 2010).

In language testing, the evaluation of rater judgments by GT is often embedded in a broader estimation of the generalizability of test results (e.g. Bouwer, Béguin, Sanders, & van den Bergh, 2015; Han, 2016; Huang, 2008, 2012; Sudweeks, Reeve, & Bradshaw, 2004). In this process, raters' main effects and their interaction effects with other facets can be identified. For example, in a pilot study of a writing performance test, Sudweeks et al. (2004) had nine raters provide two ratings of two tasks performed by 24 students twice. By evaluating the variance components related to raters, they found that raters' severity differences accounted for 1.7 % of overall score variance; raters' inconsistency in two occasions explained 0.3% of the score variance; and rater-candidate interaction contributed to 3.5% of the score variance. The results indicated that the nine raters can make appropriate judgments in the pilot study. From a macro perspective, In'nami and Koizumi (2016) synthesised previous GT analyses in performance tests, and identified that rater interaction effects generally accounted for higher score variance than did rater main effects, and rater-related variance components were generally smaller compared with those from tasks.

In comparison to CTT, GT enables researchers to evaluate different sources of error variance in norm-referenced and criterion-referenced tests, and to estimate generalizability coefficients in alternative testing conditions to achieve an optimal testing design (Bachman & Kunnan, 2005). However, the outcome of GT analysis may be affected by specific samples, and group-level information provides limited guidance for the improvement of an individual element (e.g. a specific rater) in a test (Lynch & McNamara, 1998; Schoonen, 2012).

Evaluating rater judgments through MFRM

Originating from the basic Rasch model (Rasch, 1960), MFRM is a logistic latent trait model about probabilities. The basic model views the probability of a correct response to be affected by test-takers' ability and difficulty of tasks. Through mathematical and statistical modeling, these two factors would be calibrated and evaluated independently. MFRM advances by examining the facets beyond test-takers and tasks, including raters, categories, occasions, and so forth, that affect candidates' test scores (Eckes, 2015; Lynch & McNamara, 1998). Concerning the impact of each facet on measurement, MFRM estimates test-takers' probability to succeed in a task or within a certain threshold of rating scale (Bond & Fox, 2015), in which all facets will be placed on an interval vertical ruler for comparison. This model is therefore suitable for evaluating performance tests.

MFRM has been widely applied in the field of language assessment since the 1990s, including the evaluation of rater judgments (e.g. Eckes, 2005; Knoch, 2011; Kondo-Brown, 2002; Lumley & McNamara, 1995; Schaefer, 2008; Weigle, 1998, 1999; Wigglesworth, 1993). A major indicator of rater judgements from MFRM is rater severity, which has been investigated in the majority of rater-mediated MFRM-based assessments (e.g. Lumley & McNamara, 1995; McNamara, 1996; Yan, 2014). In addition, MFRM also presents information about intra-rater consistency through fit statistics in language performance assessments (e.g. Bachman, Lynch, & Mason, 1995; Eckes, 2005; Fan & Bond, 2016; Yan, 2014). Another important indicator in MFRM studies is bias estimates (McNamara, 1996; Kondo-Brown, 2002; Schaefer, 2008; Upshur & Turner, 1999). Rater bias manifests itself when raters assign particularly severe or lenient scores systematically to a certain group of candidates, tasks, categories in criteria, etc. (Jin & Wang, 2017). In this regard, bias might overlap with inter- and intra-rater consistency, but can describe in detail what causes inconsistency.

It is worth noting that some studies combined GT and MFRM to evaluate rater judgements from macro (GT) and micro (MFRM) perspectives (Bachman et al. 1995; Lynch & McNamara, 1998; Sudweeks et al., 2004; Kim & Wilson, 2009). For example, Bachman et al. (1995) applied both GT and MFRM to examine the scores from speaking tasks in a newly-designed placement test. Results from GT indicated that the inconsistency among the raters and rater-candidate interaction were found not to affect score variance, and rater-task interaction contribute to only 6% of the score variance. MFRM further revealed each rater's severity and presented a severity range of 4.2 logits; all raters' fit statistics, and bias estimates were proved satisfactory. The researchers suggested that these two models are not mutually exclusive. GT is able to identify the effects of each facet and their interactions relative to the overall score variance; MFRM can provide specific information such as the quality of each rater's judgements.

To validate a performance test used for immigration purposes in Australia, Lynch and McNamara (1998) applied GT and MFRM to analyse the trial data in a testing occasion derived from 83 candidates, four raters, seven performance tasks, and a six-point analytical rating scale. Although producing similar results about rater severity, the two approaches caused a "somewhat striking difference" (p. 176) with regard to rater bias. MFRM revealed an extensive case of significant rater-candidate interaction (36%) and rater-task interaction (48%), whereas GT showed that these two interactions contribute to only 3.2% and 0.03% of score variance. The researchers explained that the discrepancy was attributed to different levels of details operated by the two approaches. MFRM functioned as a microscope to detect every case of significant interaction, yet these interactions may not exert huge influences on the final scores and thus were flattened out by GT analysis. By simply relying on either method, researchers may draw a misleading conclusion. While MFRM is capable of providing detailed information about individual raters, GT could detect whether raters' performance would significantly affect rating quality. The complementary role highlights the necessity to use both GT and MFRM in the evaluation of rater judgements.

Context of the study

English Test for International Communication (ETIC)

Launched by Chinese Language Assessment (CLA) in 2016, ETIC is a performance-based criterion-referenced test that consists of two suites: main and translation suites. The main suite covers four levels: Basic, Intermediate, Advanced, and Superior, while translation suite includes three categories: Written Translation, Consecutive and Simultaneous Interpretation. ETIC main suite is held twice a year, in May and in November, aiming to assess one's ability to perform English language tasks in the international workplace. Each level of main suite is composed of writing and speaking communication tasks. Most of these tasks are integrated, performance-based, and constructed-response in nature. All the tests are administered online via computer (Luo & Han, 2018).

Writing tasks in ETIC Advanced

ETIC Advanced evaluates candidates' ability to perform English tasks in international workplace contexts. Candidates with the level C1 in CEFR (Council of Europe, 2001) are expected to pass the test (China Language Assessment, 2018). ETIC Advanced writing component consists of three tasks (see Appendix 1 for task formats). The first task of Letter Writing requires the test-takers to write a letter, of about 150 words, in 25 minutes. Functions of letters include issuing an invitation, responding to requests, inquiring about information, etc. It is designed to measure one's ability to describe and explain an issue to a particular person in a specific situation with an assigned role.

In the second task of Report Writing, test-takers are required to read a graph and write a report, of about 150 words, in 25 minutes, to describe the information by summarizing the main features. By completing this task, test-takers can demonstrate their ability to describe, compare, and summarise the key information.

The third task of Proposal Writing requires test-takers to write a proposal with 300-350 words long in 40 minutes. This task asks test-takers to play a specific role and demonstrate their ability to clarify the purpose and argue for the necessity of the proposal.

Ratings in ETIC Advanced

It has been reported that ETIC has taken different measures to reduce the potential problems of rating (Luo & Han, 2018). Prospective ETIC raters have to participate in a one-day training program. They will be provided with detailed information regarding ETIC at that time (e.g. description of participants and format of the test), and the common rater judgment errors in the rating process. They will later be classified into groups according to the test and task categories for a specific apprenticeship, in which the group leader will familiarise them with the tasks, rating scales, and benchmarking samples. Then the raters are required to assign scores to the samples. After that, they will check the results and discuss with the group leader about any uncertainty in their ratings.

During the rating, the raters use a six-point three-category analytic rating scale presented in Chinese (see Appendix 2 for the English translation). The first category of Content deals with topic relevance, fulfillment of task requirements, and sufficiency of supporting details. The second category of Organization examines the development of ideas, coherence and cohesion, and format. The last category of Language concerns word choice, flexibility of sentences, and grammar accuracy. In addition, each task will be double rated on an online scoring system to increase the reliability of the score. Some benchmarking samples will be embedded in the system to examine raters' internal consistency throughout the rating process (Luo & Han, 2018). Notification and suggestions will be provided to raters whose ratings are inconsistent.

After the rating, the online rating system will identify the tasks that received notably discrepant scores from the two raters, and submit them to expert raters for a third rating. This can reduce significant differences between raters in rating pairs.

Although this carefully controlled rating process contributes to the quality of final scores, empirical studies on evaluating raters' judgements remain necessary. On the one hand, the study helps to strengthen the validity of the interpretation and use of ETIC Advanced test scores. On the other hand, results about rater judgements may facilitate future rater training, thus improving rating quality and reducing expert raters' workload and the cost of rating.

Research questions

This study aims to evaluate the quality of rater judgements (inter-rater consistency, intra-rater consistency, and rater bias) in ETIC Advanced writing tasks. Three research questions are proposed:

> RQ1: Do raters of ETIC Advanced writing tasks demonstrate a satisfactory level of inter-rater consistency?
>
> RQ2: Do raters of ETIC Advanced writing tasks demonstrate a satisfactory level of intra-rater consistency?
>
> RQ3: Do raters of ETIC Advanced writing tasks assign scores consistently across different candidates and categories in the rating scale?

Methodology

Data collection

The study analysed 3,348 ratings from 186 test-takers who completed all the ETIC Advanced writing tasks in November 2017. The ratings were randomly sampled from the data pool of ETIC Advanced, and were authorised by CLA after the researchers assigned a confidentiality agreement. Third ratings from expert raters were not used because third ratings will improve the quality of rating and flatten out potential problems of rater judgements.

The examinees in November 2017 consisted of EFL/ESL learners, government officials, and employees from various occupations such as English teachers in high schools and universities, and staff in enterprises. The test takers' writing performance were rated by six certified ETIC raters who had experience in rating large-scale high-stake English tests in China, including the College English Test (CET), the Test for English Major (TEM), and the China Accreditation Test for Translators and Interpreters (CATTI).

Data analysis

Integrating GT and MFRM to evaluate rater consistency

In order to answer the research questions, this study applied GT and MFRM to analyse the data. GT is able to evaluate rater judgements from an aggregated level. MFRM serves as a magnifying lens to investigate rater judgements at an individual level. Researchers can investigate which rater tends to judge idiosyncratically in terms of inter-rater consistency (rater severity), intra-rater consistency (rater fit), and rater bias.

Generalizability theory analysis

The study used EduG 6.1e (Swiss Society for Research in Education Working Group, 2010) to conduct GT analysis. This program features a graphical user interface and drop-down menus, similar to SPSS, but uses slightly different terminologies compared with those in GENOVA suites. In EduG settings, the relationship among the facets was identified as PC(R:T) because in the rating design of ETIC Advanced in 2017, the six raters used the same rating scale and every two raters were classified as a group to rate one task. Therefore, candidates were fully crossed with tasks and the three categories of the rating scale, whereas the six raters were nested within three tasks.

Concerning the status of the facets, Person and Raters were set as infinite random as they were treated as samples from the universe. Category was set as a fixed facet, since they were carefully developed and considered to stand for the criteria used in a specific target language use domain. Task was also regarded as a fixed facet because, on the one hand, the three task types, derived from the results of needs analyses, were considered to be representative of the actual writing contexts in international workplace communication, and were not exchangeable with other tasks (see Shavelson & Webb, 1991, pp. 65-66). On the other hand, the study intended to restrict the writing test scores to the three tasks, since the primary focus was on the quality of rater judgements instead of the quality of tasks.

Additionally, a D study was also conducted to evaluate the G-coefficients in operational rating designs and different potential rating designs. Although the primary focus is on the dependability of the results in current rating designs, the G-coefficients from different testing conditions may provide insight in optimizing future rating designs.

To analyse the data, EduG will first present the results through G study regarding the effects of variance components (VC) and their interactions that contribute to score variance in a single observation. Brown (2011) pointed out that the VC for rater main effect reveals their severity differences; VCs about raters' interaction with other facets pertain to their fluctuation of severity when encountering different elements in those facets (e.g. rating different test-takers). Accordingly, these results will partly address the first and third research questions.

While G study in this study identifies the contribution of inter-rater consistency and rater bias to the score variance in a single observation, the D study, drawing on the baseline data from G study, examines the extent to which raters will influence the generalizability or the dependability of test results. It is possible that the rater effects identified by G study may not significantly impact generalizability of test results as long as the G coefficients from D study are satisfactory (above 0.8). In this sense, the results from G study and D study should be jointly interpreted in order to answer the research questions.

Many-Facets Rasch Model analysis

This study used FACETS (Version 3.81.2, Linacre, 2019) to conduct MFRM analysis. Specifically, the facets of candidates, tasks, raters, and categories were included in the analysis. The rater facet was set as partial credit with an assumption that the raters had their own understandings of rating scales. The results of rater severity differences and their use of scale steps will address RQ1, and rater fit statistics are able to address intra-rater consistency (RQ2). To answer RQ3, which is focused on rater bias, we conducted rater-candidate and rater-category interaction analysis. Since raters were nested within tasks, it was unnecessary to analyse rater-task interaction.

Linking design for MFRM analysis

As mentioned above, ETIC Advanced adopts a nested linking design in actual ratings. Every two raters were grouped to exclusively rate one task (Table 1). Nested design is economical and practical, and has the advantage of diminishing the rater-task interaction.

Table 1. Rating assignment of ETIC Advanced writing tasks (nested design)

Rater	Task	Examinee									
		1	2	3	4	5	6	7	8	…	186
1	1	√	√	√	√	√	√	√	√	√	√
2		√	√	√	√	√	√	√	√	√	√
3	2	√	√	√	√	√	√	√	√	√	√
4		√	√	√	√	√	√	√	√	√	√
5	3	√	√	√	√	√	√	√	√	√	√
6		√	√	√	√	√	√	√	√	√	√

However, the nested design in actual ETIC ratings violated the requirement of connectedness in MFRM. In table 1, although the severity of the raters within the same group can be directly evaluated through MFRM analysis, the raters among groups cannot be compared as the tasks they rated had no connection.

Table 2. Rating assignment for remedying disconnected data

Rater	Task	Examinee										
		1	2	3	4	5	6	7	8	…	186	
1	1	√	√	√	√	√	√	√	√	√	√	Round one
2		√	√	√	√	√	√	√	√	√	√	

3	2	√	√	√	√	√	√	√	√	√	√	
4		√	√	√	√	√	√	√	√	√	√	
5	3	√	√	√	√	√	√	√	√	√	√	
6		√	√	√	√	√	√	√	√	√	√	
7	1, 2, 3	√	√	√	√	√	√	√	√	√	√	Round two

To conduct MFRM analysis, the current study added an extra round after the official rating condition (Table 2). This linking design transformed the nested linking design into a mixed design. The extra rater 7 in round two is a certified ETIC rater who assigned scores to all three tasks written by all the test-takers. The disconnected data were linked because all the raters can be compared by referring their scores to those from the second round[2]. Rater 7 in the design only served a linking purpose and will not be analysed in the study.

Results

Results from GT analysis

Generalizability study

Based on current rating design, G study decomposed the facets and their interactions into variance components and identified their contribution to overall score variance (Table 3). Percentage of variability (POV) was chosen for reporting since it was more explicit compared to the variance components.

Table 3. Amount of variability due to each source

Source	Variance components (VC)	Standard error (SE)	Percentage of variability (POV)
Person (P)	0.269	0.031	27.4
Task (T)	0.071	0.077	4.8
Rating (R:T)	0.064	0.041	6.5
Criteria (C)	0.014	0.012	1.0
Person by task (PT)	0.188	0.020	19.2
Person by rating (PR:T)	0.156	0.010	15.9
Person by criteria (PC)	0.019	0.004	1.9
Task by criteria (TC)	0.006	0.008	0.3
RC:T	0.011	0.006	1.1
PTC	0.038	0.008	3.9
PRC:T, e	0.175	0.007	17.9
Total			100%

[2] This linking design was confirmed by Mike Linacre though a personal communication in April 1, 2018.

The POV for Person demonstrated the extent to which test-takers' ability differed from each other (Brown, 2011). Although there are no agreed standards for the POV for Person, researchers generally expect the percentage to be larger than any other sources (e.g. Sudweeks et al. 2004), thus indicating that test-takers' ability differences account for the largest proportion of test score variance. In table 3, the POV for Person (27.4 %) showed that test-takers had different levels of ability and could be differentiated by the writing tasks.

It is worth noting that, although POV for Person was 27.4%, we would not interpret this result as test-takers' ability differences account for only 27.4 % of the overall score variance, due to the G study results being based on a single observation. Specifically, although the current testing design contains 186 test-takers, three tasks, six raters, and three categories, the G study results only derive from one candidate, rated by one rater using one of the categories in one task. In this sense, the variance components from the G study are for reference only, and we have to resort to the results from the D study to get an overall evaluation of the quality (generalizability) of the test results.

Concerning the rater judgements, the POV for rating (R: I) indicated the extent to which the raters differed from each other. The POV of 6.5 % suggested that the raters had somewhat different degrees of severity. This figure can be ignored in norm-referenced tests since it would not affect the rank order of candidates. In criterion-referenced test, as is the case in ETIC Advanced, this is of concern because raters of different severity levels would demonstrate discrepancy in assigning scores and thus potentially decrease the inter-rater reliability. Table 3 also revealed rater-candidate and rater-category interaction. The POV of 15.9 % meant that the raters' inconsistent judgements to different test-takers may contribute substantially to score variance. The raters may therefore receive further training to reduce the interaction. With regard to rater-category interaction, the raters did not demonstrate biased use of the categories in the rating scale (POV = 1.1%).

Table 3 showcased the importance of each facet that contributed to overall score variance. While the discrepancy among raters accounted for a rather small proportion of score variance, the rater-candidate interaction had a larger impact on final scores. The following D study will evaluate whether the test scores suffice for generalization after taking these error variances into account.

Decision study

While G study estimates the results from a single observation, D study uses those results as building blocks to evaluate generalizability of the test scores in operational testing design as well as various potential designs. Two crucial indicators in D study are Coefficient G-relative (used in NRT) and Coefficient G-absolute (used in CRT).

They are similar to the reliability coefficients in CTT, and we expect the value to be above 0.8 (Cardinet et al., 2010). Since ETIC Advanced is a criterion-referenced test, we focused on Coefficient G-absolute, which indicates the extent to which the object of measurement can be placed above or below particular cut scores in criterion-referenced tests.

Table 4 revealed the generalizability of the writing scores of ETIC Advanced in the operational and various potential testing designs. Firstly, the columns with italics indicated that the generalizability of the operational testing design that contains three writing tasks, two raters nested within each task, and three categories, was satisfactory. That is, while G study revealed that the raters differed in severity and demonstrated interaction with candidates, the test scores remained satisfactory to be generalised (Coefficient G-absolute= 0.88).

Table 4. Results of D study (measurement design p/ric)

Tasks	Raters (nested within tasks)	Categories	Coefficient G-relative	Coefficient G-absolute
1	2	3	0.597	0.569
2	2	3	0.885	0.845
_3_	_2_	_3_	**_0.912_**	**_0.88*_**
4	2	3	0.928	0.902
3	**1**	3	0.838	0.786
_3_	**_2_**	_3_	**_0.912_**	**_0.88*_**
3	**3**	3	0.939	0.917
3	**4**	3	0.954	0.936
3	2	**1**	0.75	0.724
3	2	**2**	0.898	0.867
_3_	_2_	**_3_**	**_0.912_**	**_0.88*_**
3	2	**4**	0.919	0.887

*Italics refer to actual rating design used in ETIC Advanced

In alternative testing designs, we first discovered that tasks may exert the largest impact on score generalizability. When increasing the number of tasks from one to four, the Coefficient G-absolute increased from 0.569 to 0.902. Remarkably, a single task rated by a rater pair using a three-category rating scale merely produced the Coefficient G-absolute of 0.569, but the coefficient escalated to 0.845 when adding one more task to the writing section. This suggested that at least two tasks should be used in order for the test outcome to be generalizable.

In terms of raters, the Coefficient G-absolute increased notably when increasing the elements of the Rater facet, indicating that the raters exerted influence on the generalizability of writing scores, in line with the results about raters in the G study. In a condition where each task was single rated, the overall Coefficient G-absolute turned out to be 0.786. When the three tasks were double rated in the current rating design, this coefficient climbed to 0.88. This proved that the double rating process was qualified to provide generalizable scores.

The information about categories exhibited that applying one of the categories of Content, Organization, and Language, may be under-representative of test-takers' ability, because the rating design with a single category elicited the Coefficient G-absolute of 0.724. This coefficient increased when using one more category, rising to 0.867. The coefficient grew slowly, however, when adding more categories in the rating scale (from 0.867 to 0.887).

In summary, the D study revealed that raters' erroneous judgements identified in the G study did not significantly affect the generalizability of test scores (Coefficient G-absolute=0.88), which proved that the raters could demonstrate satisfactory inter-rater consistency, and remained consistent across different candidates and categories. However, room still existed for the improvement of the generalizability of test scores. The suggestion from the D study was to increase the elements of tasks and raters from a macro perspective. The MFRM in the following section will contribute from a micro perspective.

Results from MFRM analysis

Candidates

Results from FACETS revealed that the writing tasks can successfully divide candidates into different ability levels. According to the vertical ruler (Figure 1), the most proficient candidates were about 4.2 logits and the least proficient candidate was about -2.4 logits, having a spread of 6.43 logits. The significant Chi-square rejected the null hypothesis that candidates have the same level of ability (see Table 5). The separation index of 3.96 further indicated that the candidates were divided into about 4 levels according to their ability. The reliability of 0.94 proved that such separation is very likely to happen repeatedly.

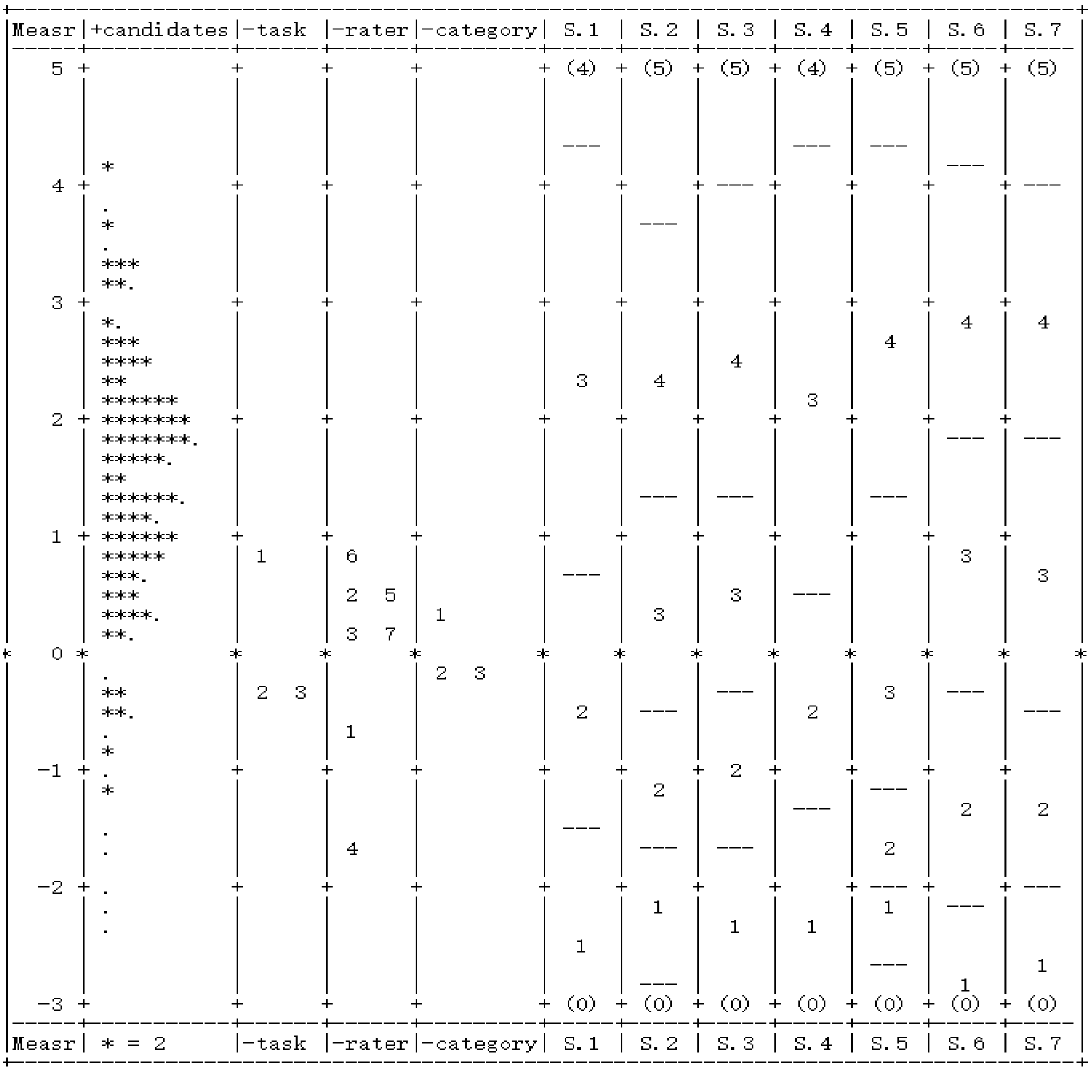
Table 5. Summary of MFRM analysis results

Statistics	Candidates	Tasks	Raters	Categories
Range (logits)	6.43	1.13	2.73	0.53
M SE	0.30	0.04	0.06	0.04
df	185	2	6	2
χ^2	3738.6*	549.5*	813*	134.6*
Separation index	3.96	14.07	12.41	6.57
Separation reliability	.94	.99	.99	.98

*$p < 0.01$

Tasks

The three writing tasks were shown to have different levels of difficulty (χ^2 =549.5, df=2, p<0.01). According to the vertical ruler, Letter Writing was the most difficult task (0.75 logits) while Report Writing and Proposal Writing shared a similar level of difficulty (-0.38 logits), indicating that the tasks were relatively easy (see Figure 1). As a criterion-referenced test, ETIC Advanced does not require a strict targeting between candidates and tasks. Therefore, the relatively easy tasks may indicate that a higher number of candidates are able to pass the test. In terms of fit statistics, the three tasks were within a satisfactory range (Letter Writing, infit MnSq=0.95 and outfit MnSq=0.97; Report Writing, Infit MnSq=1.03 and outfit MnSq=1.00; Proposal Writing, infit MnSq=1.03 and outfit MnSq=1.02), and this served as supporting evidence of the quality of tasks.



```
+-----------------------------------------------------------------------------------+
|Measr|+candidates|-task |-rater|-category| S.1 | S.2 | S.3 | S.4 | S.5 | S.6 | S.7 |
|-----+-----------+------+------+---------+-----+-----+-----+-----+-----+-----+-----|
|   5 +           +      +      +         + (4) + (5) + (5) + (4) + (5) + (5) + (5) |
|     |           |      |      |         |     |     |     |     |     |     |     |
|     |           |      |      |         |     |     |     |     |     |     |     |
|     |           |      |      |         |     |     |     |     |     |     |     |
|     |           |      |      |         | --- |     |     | --- | --- |     |     |
|     | *         |      |      |         |     |     |     |     |     | --- |     |
|   4 +           +      +      +         +     +     + --- +     +     +     + --- |
|     | .         |      |      |         |     |     |     |     |     |     |     |
|     | *         |      |      |         |     | --- |     |     |     |     |     |
|     | .         |      |      |         |     |     |     |     |     |     |     |
|     | ***       |      |      |         |     |     |     |     |     |     |     |
|     | **.       |      |      |         |     |     |     |     |     |     |     |
|   3 +           +      +      +         +     +     +     +     +     +     +     |
|     | *.        |      |      |         |     |     |     |     |     |  4  |  4  |
|     | ***       |      |      |         |     |     |     |     |  4  |     |     |
|     | ****      |      |      |         |     |     |  4  |     |     |     |     |
|     | **        |      |      |         |  3  |  4  |     |     |     |     |     |
|     | ******    |      |      |         |     |     |     |  3  |     |     |     |
|   2 + *******   +      +      +         +     +     +     +     +     +     +     |
|     | *******.  |      |      |         |     |     |     |     |     | --- | --- |
|     | *****.    |      |      |         |     |     |     |     |     |     |     |
|     | **        |      |      |         |     |     |     |     |     |     |     |
|     | ******.   |      |      |         |     | --- | --- |     | --- |     |     |
|     | ****.     |      |      |         |     |     |     |     |     |     |     |
|   1 + ******    +      +      +         +     +     +     +     +     +     +     |
|     | *****     | 1    | 6    |         |     |     |     |     |     |  3  |     |
|     | ***.      |      |      |         | --- |     |     |     |     |     |  3  |
|     | ***       |      | 2  5 |         |     |     |  3  | --- |     |     |     |
|     | ****.     |      |      | 1       |     |  3  |     |     |     |     |     |
|     | **.       |      | 3  7 |         |     |     |     |     |     |     |     |
*   0 *           *      *      *         *     *     *     *     *     *     *     *
|     | .         |      |      | 2  3    |     |     | --- |     |     |     |     |
|     | **        | 2  3 |      |         |     |     |     |     |  3  | --- |     |
|     | **.       |      |      |         |  2  | --- |     |  2  |     |     | --- |
|     | .         |      | 1    |         |     |     |     |     |     |     |     |
|     | *         |      |      |         |     |     |     |     |     |     |     |
|  -1 + .         +      +      +         +     +     +  2  +     +     +     +     |
|     | *         |      |      |         |     |  2  |     |     | --- |     |     |
|     |           |      |      |         |     |     |     | --- |     |  2  |  2  |
|     | .         |      |      |         | --- |     |     |     |     |     |     |
|     | .         |      | 4    |         |     | --- | --- |     |  2  |     |     |
|     |           |      |      |         |     |     |     |     |     |     |     |
|  -2 + .         +      +      +         +     +     +     +     + --- +     + --- |
|     | .         |      |      |         |     |  1  |     |     |  1  | --- |     |
|     | .         |      |      |         |     |     |  1  |  1  |     |     |     |
|     |           |      |      |         |  1  |     |     |     |     |     |     |
|     |           |      |      |         |     |     |     |     | --- |     |  1  |
|     |           |      |      |         |     | --- |     |     |     |  1  |     |
|  -3 +           +      +      +         + (0) + (0) + (0) + (0) + (0) + (0) + (0) |
|-----+-----------+------+------+---------+-----+-----+-----+-----+-----+-----+-----|
|Measr| * = 2     |-task |-rater|-category| S.1 | S.2 | S.3 | S.4 | S.5 | S.6 | S.7 |
+-----------------------------------------------------------------------------------+
```

Figure 1 Vertical rulers of all facets

Categories

Similar to the task facet, significant differences in difficulty of the three categories were also found (χ^2=134.6, df=2, p<0.01), causing a narrow range of 0.53 logits. Fit statistics showed that the three categories functioned satisfactorily, as they fit the Rasch model well (Content, infit MnSq=1.10 and outfit MnSq=1.11; Organization, infit MnSq=1.06 and outfit MnSq=1.05; Language, infit MnSq=0.86 and outfit MnSq=0.83). Nevertheless, the vertical ruler revealed that the scale steps were used differently by the raters, which may be attributed to both raters and categories (e.g. descriptors in each category).

Rater measurement report

Table 6 presents detailed information about rater judgements. Significant Chi-square statistics rejected the null hypothesis that raters have the same level of severity. The large separation estimates and separation reliability also confirmed the variation among the raters. To be specific, rater 1 and 4 were particularly lenient rater, and thus, considerably widened the spread of rater severity. With the exception of raters 1 and 4, other raters created a narrow spread of severity (0.98 logits). The fit statistics indicated the extent to which the raters demonstrated unexpected judgements in the rating process, and thus, served as an indicator of intra-rater consistency. It was indicated that raters' MnSq of infit and outfit were both within the actable range of 0.7 to 1.3 (Bond & Fox, 2015), suggesting that the raters were generally consistent in their rating judgements.

Table 6. Rater measurement report

Rater	Measure	Model S.E.	Infit		Outfit	
	(logits)		MnSq	ZStd	MnSq	ZStd
6	0.88	0.06	0.86	-2.4	0.86	-2.4
2	0.56	0.06	0.97	-0.5	0.95	-0.8
5	0.47	0.07	1.09	1.2	1.03	0.5
7	0.08	0.04	1.05	1.3	1.03	0.7
3	-0.10	0.06	1.06	0.9	1.03	0.5
1	-0.25	0.07	0.94	-0.9	0.94	-0.8
4	-1.85	0.08	1.05	0.7	1.07	0.9

Note: Separation reliability: 0.99; Separation: 12.41; Fixed (all same) Chi-square: 813, df: 6, p= .00

Rater bias estimates

Since the nested rating design had eradicated raters' bias to the tasks, this section mainly presented the condition of rater-candidate bias and rater-category bias. Table 7 displays the pattern of raters' bias when rating essays from candidates of different abilities. Among the 1,116 rated essays (186 candidates × 3 essays × 2 raters), 129 received biased judgements, occupying a relatively small proportion (11.7 %). This proved that the rater-candidate interaction was satisfactory in general. With respect to bias pattern, the raters' tendency to assign biased ratings varied inversely with the ability spectrum. Specifically, the raters assigned the least biased scores when rating the extremely competent candidates (3 logits or above). The percentage of biased ratings increased gradually as the raters faced less competent candidates and was the highest (16.7 %) when they rated extremely unproficient candidates. In terms of individual raters, all of them demonstrated around 20 biased ratings. Rater 1 had the most biased judgements in the rating process (*n*=25). Although rater 4 was the most lenient rater, he or she showed the least biased judgements when rating different candidates (*n*=16).

Table 7. Rater-candidate bias report

Ability estimate (logits)	**Number of rated essays**	**Number of significant biased cases**						**Total**	**%**
		Rater 1	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 5**	**Rater 6**		
3.01 or higher	102	0	0	1	1	0	0	2	2
2.01 to 3.00	246	6	4	4	3	2	8	28	11.4
1.01 to 2.00	396	12	7	8	4	9	6	46	11.6
0.00 to 1.00	246	4	5	5	6	7	7	34	13.8
-1.00 to -0.01	78	2	3	3	1	1	1	11	14.1
-2.00 to -1.01	36	1	0	3	1	0	1	6	16.7
-2.01 or lower	12	0	1	0	0	1	0	2	16.7
Total	1116	25	20	24	16	20	23	129	11.6

Figure 2 provides the information on rater-categories bias. The X-axis listed the three categories of Content, Organization and Language, and the Y-axis presented "t" value which pertains to the hypothesis "there is no bias apart from measurement error" (Linacre, 2017, p. 218). When the absolute value of "t" is greater than 2.0, we can state that the case of bias is statistically significant instead of happening by chance. In the case of rater-category interaction, the "t" value above 2.0 indicates that raters consistently interpret the category more severely than normal; the "t" value below -2.0 means that raters are more lenient in using that category.

The figure reveals that four of the six raters demonstrated at least one case of bias in using the categories. Raters 3, 5, and 6 had one case of biased judgements. Rater 4 showed bias toward both Content and Language in the ratings, suggesting that he or she may have problems in understanding and applying the rating scale. Raters 1 and 2 were immune to rater-category bias.

Three of the six raters demonstrated biased judgements in using Content, which implied that the training program was not effective in clarifying its meaning. Most raters interpreted and applied Organization appropriately, with the exception of the third rater who used the category more severely. In terms of Language, the raters tended to use it either harshly (raters 1, 6) or leniently (raters 2, 3, 4, 5), with rater 4 demonstrating significantly lenient judgements. In this sense, the category of Language may deserve further investigation.

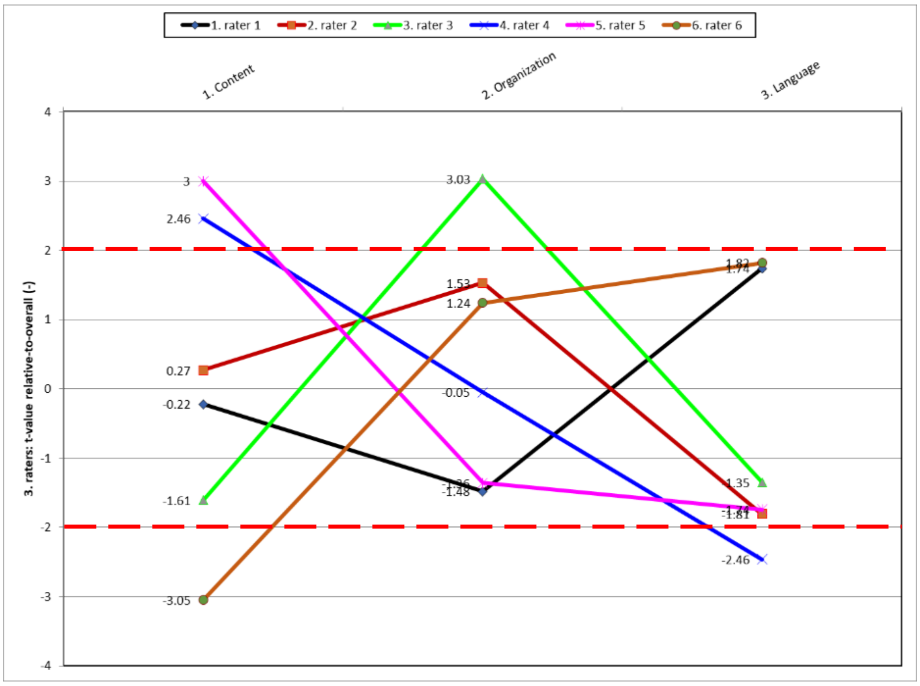


Figure 2. Rater-category bias

Discussion

Question 1: Do raters of ETIC Advanced writing tasks demonstrate a satisfactory level of inter-rater consistency?

Although the vertical ruler and rater measurement report in FACETS discovered significant rater differences in severity, GT identified that inter-rater consistency in the test remained satisfactory. On the one hand, the rater main effect in the G study contributed to 6.5 % of score variance in a single observation, far less than did the POV of person effect (27.4 %). On the other hand, the D study demonstrated that the scores from operational testing design sufficed for generalization (Coefficient G-absolute=0.88), which implied that raters' performance were satisfactory in general.

Through the evaluation of inter-rater consistency, this study, akin to previous ones (e.g. Bachman et al., 1995; Lynch & McNamara, 1998), showcased the advantage of applying both GT and MFRM in test analysis. While revealing severity differences, FACETS may not provide a simple indicator that reveals the extent to which such differences will affect the quality of test scores. By contrast, the G study and D study are able to demonstrate raters' contribution in a single observation and various potential test designs. To further extend the rater effects discovered by GT, MFRM presented every rater's specific severity level and identified rater 4 as the most lenient rater (-1.85 logits) who increased the overall severity range to 2.73 logits. A possible reason is that this rater may possess idiosyncratic understandings of candidates' abilities and the rating scale. To improve the overall inter-rater consistency, he or she may require targeted training in the following rating assignments.

The results of inter-rater consistency in the study also served to support that, despite a thorough rater training program, inconsistency among the raters remained, which is consistent with results in other rater-related research (e.g. Eckes, 2005; Kondo-Brown, 2002; Lynch & McNamara, 1998; Upshur & Turner, 1999; Weigle, 1998; Yan, 2014). The results remind us that the rater training program may not be completely successful in eradicating rater inconsistency, albeit it can be effective to a certain degree. Therefore, we have to stress the need for surveillance of rater consistency in spite of a training program, and conduct empirical studies to evaluate rating quality.

Question 2: Do raters of ETIC Advanced writing tasks demonstrate a satisfactory level of intra-rater consistency?

While GT is unable to evaluate intra-rater consistency in only one testing occasion, MFRM can estimate whether the raters exhibit similar severity levels throughout the assignment. The statistics of infit and outfit MnSq showed that all six raters achieved satisfactory fit statistics with a range between 0.86 and 1.09. This meant that the six raters fit the Rasch model and maintained satisfactory intra-rater consistency.

Some researchers suggested that the primary purpose of rater training is to improve raters' internal consistency rather than inter-rater consistency, because an undue emphasis on the latter may hinder raters' normal rating performance (Lunz et al., 1990). McNamara (1996) believed that raters' random error pertaining to internal consistency could be a more serious problem than systematic effects, and accordingly proposed that an appropriate purpose of rater training is "to make raters internally consistent so as to make statistical modelling of their characteristics possible" (p. 127). This view can be supported by the results of the current study. Although rater 4 was particularly lenient in his overall ratings, the lenient scores could be adjusted in FACETS by referring to the fair average scores that the candidates deserved. Since all six raters demonstrated high level of intra-rater consistency, in this regard, the rater training program in ETIC Advanced is rather effective.

Question 3: Do raters of ETIC Advanced writing tasks assign scores consistently across different candidates and categories in the rating scale?

In terms of rater-candidate interaction, the G study discovered that rater-candidate interaction occupied rather large variance components (POV=15.9%) in a single observation. Concerning the results from the D study, fortunately, the operational testing design in the writing part of ETIC Advanced writing elicited a high Coefficient G-absolute of 0.88, and this secured a satisfactory rater-candidate interaction in general.

Nonetheless, the relatively large effect of rater-candidate interaction is still worth further investigation. From an individual level, MFRM disclosed that each rater assigned around 20 significantly biased scores on average, with rater 1 having the most (*n*=25), and rater 4 assigning the least (*n*=16). In addition, the raters tended to assign more biased scores to the candidates with extremely low proficiency, which is in line with Yan's (2014) findings, and partly supports Kondo-Brown's (2002) results where the raters assigned more biased scores to candidates with extremely high or low abilities. Yet, the rater-candidate bias pattern in this study was opposite to Schaefer's (2008) results, as the raters in Schaefer's study assigned more biased scores to high proficiency test-takers. The contradictory results may be attributed to the uniqueness of different contexts. To be specific, there were numerous differences among the studies above regarding test purposes, test-takers, task design, rating criteria, rater training, etc. Moreover, the raters also differed in language background, experience, and nationality. Therefore, they may have different perceptions about rating scales and candidates' performances. Such uncertainty from raters may decrease the consistency both among and within raters, thus calling for further research.

With respect to rater-category bias, GT revealed that rater-category interaction accounted for 1.9 % of the score variance in a single observation, which indicated that the raters elicited a satisfactory result of rater-category interaction in their rating assignment. The results from the D study proved this as well. From a micro perspective, MFRM identified that raters 1 and 2 remained consistent in using different categories, whereas the others more or less demonstrated bias in using the categories. Yet, no obvious bias patterns were found in the present study. Among the four raters who showed bias, rater 4 may be problematic, as he or she showed bias when using both Content and Language. This indicated that this rater may require further training on the use of categories. Notably, the category of Content is more liable to interact with raters, contributing to 60% of the biased cases. This suggested that the rater training section may need to further clarify its meaning. In order to minimise rater-category bias, further studies about raters' mental processes in using rating scales can be conducted, such as probing their decision-making process through verbal reports or eye-tracking technologies. Information of this kind will deepen the understandings about rater judgements and contribute to rating quality.

Conclusion

This study applied both GT and MFRM to evaluate raters' judgements in ETIC

Advanced writing tasks. It found that the raters had significant differences in severity, but such differences accounted for a small proportion of the score variance. Additionally, the raters were internally consistent in the rating process. In terms of rater bias, they tended to judge inappropriately when rating extremely low-proficiency candidates, and their understandings about the category of Content were less aligned. Fortunately, such interactions did not exert a significant impact on the quality of final scores.

This study evaluated rater judgements by using both GT and MFRM. The advantages of the combined application have been demonstrated. Diagnostic information revealed from the study may contribute to improving rating quality in ways such as enhancing the training program and replacing certain problematic raters.

Limitations of the study also exist due to practicality. Since only an authorised portion of scores from writing tasks were accessible for the study, the generalizability of the research outcome to overall test results may require further justification. Besides, explanations of the raters' erroneous judgements are tentative to some extent. In order to improve the quality of rating, further qualitative studies focusing on the factors that affect raters' performance and their decision-making process in rating assignments are demanded.

Acknowledgements

This study was supported by the MOE Project of Key Research Institute of Humanities and Social Sciences and BFSU Project of World Class Academic Discipline Construction.

References

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

Bachman, L. F., & Kunnan, A. J. (2005). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

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

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.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*(1), 54-74.

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in* *the human sciences* (3rd ed). New York, NY: Routledge.

Bouwer, R., Béguin, A., Sanders, T., & van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing, 32*(1), 83-100.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Brown, J. D. (2011). What do the L2 generalizability studies tell us? *International Journal of Assessment and Evaluation in Education, 1,* 1-37.

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using EduG*. Taylor & Francis.

China Language Assessment. (2018). *The official guide to ETIC*. Beijing: Foreign Language Teaching and Research Press.

Council of Europe. (2001). *The Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972), *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117-135.

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.

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessment* (2nd ed.). Frankfurt: Peter Lang.

Fan, J. & Bond, T. (2016). Using MFRM and SEM in the validation of analytic rating scales of an English speaking assessment. In Q. Zhang (Ed), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings* (pp. 29-50). Singapore: Springer.

Goodwin, S. (2016). A Many-Facet Rasch analysis comparing essay rater behaviour on an academic English reading/writing test used for two purposes. *Assessing Writing, 30,* 21-31.

Han, C. (2016). Investigating score dependability in English/Chinese interpreter certification performance testing: A generalizability theory approach. *Language Assessment Quarterly, 13*(3), 186-201.

Huang, J. (2008). How accurate are ESL students' holistic writing scores on large-scale assessments? A generalizability theory approach. *Assessing Writing, 13*(3), 201-218.

Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. *Assessing Writing, 17*(3), 123-139.

In'nami, Y., & Koizumi, R. (2016). Task and rater effects in L2 speaking and writing: A synthesis of Generalizability studies. *Language Testing, 33*(3), 341-366.

Jin, K. Y., & Wang, W. C. (2017). Assessment of differential rater functioning in latent classes with new mixture facets models. *Multivariate behavioral research, 52*(3), 391-402.

Kim, S. C., & Wilson, M. (2009). A comparative analysis of the ratings in performance assessment using. *Journal of applied measurement*, *10*(4), 403-423.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior—a longitudinal study. *Language Testing*, *28*(2), 179-200.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*(1), 3-31.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing, 28*(4), 543–560.

Linacre, J. M. (2017). *A user's guide to FACETS: Rasch-model computer programs*. Chicago: Winsteps.com. Retrieved from http://www.winsteps.com/facets.htm.

Linacre, J. M. (2019). Facets Rasch measurement computer program (Version 3.81.2) [Computer software]. Chicago: Winsteps.com.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54-71.

Lunz, M.E., Wright, B.D. & Linacre, J.M. 1990: Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3*(4), 331-345.

Luo, K. & Han, B. (2018), Construct definition, task design and its scoring methods: An introduction to the development of ETIC (国才考试的构念界定、任务设计与评分方法), *Foreign Language Education in China, 11*(1), 40-46.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, *15*(2), 158-180.

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

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

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*(4), 465-493.

Schoonen, R. (2012). The generalisability of scores from language tests. In G. Fulcher

& F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 363-377). Abingdon, UK: Routledge.

Shavelson, R., & Webb, N. (1991). *Generalizability theory: A primer*. Newbury Park, CA: SAGE Publications.

Shaw, S. D. (2002). The effect of standardisation training on rater judgement and inter-rater reliability for the revised CPE writing paper 2. *Research Notes, 8,* 13-17.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2004). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*(3), 239-261.

Swiss Society for Research in Education Working Group. (2010). *EDUG user guide.* Neuchatel, Switzerland: IRDP.

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.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*(2), 145-178.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10*(3), 305-319.

Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed- methods approach. *Language Testing, 31*(4), 501-527.

Youn, S. J. (2018). Rater variability across examinees and rating criteria in paired speaking assessment. *Papers in Language Testing and Assessment, 7*(1), 32-60.

Appendix 1. Samples of ETIC Advanced writing tasks

Writing task 1

You are Robert Chapman, Marketing Director at Best Toys Ltd. Your company has recently developed a new product, Windsor Teddy Bear, and you need to introduce the new product to your long-time retailer Lucy France, Manager of Super Fun Toy Store. Write a **letter**

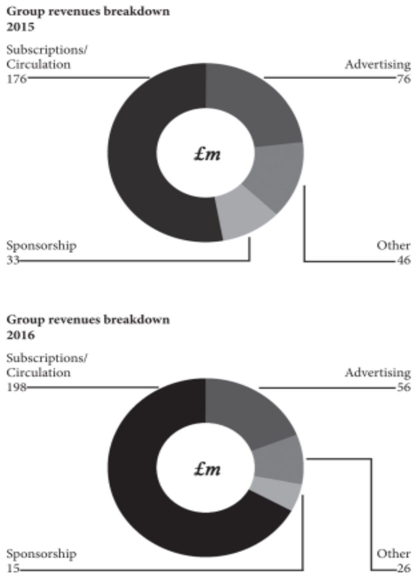
- to describe two main features of your new product;
- to inform Lucy France about your company's discount policy;
- to suggest visiting her in person.

Write about **150** words within **25** minutes.
You do NOT need to write any addresses.

Writing task 2

The graphs below show the breakdown of ABB Group revenues in 2015 and 2016. Using the information from the graphs, write a **report** describing and comparing the changes in the various sectors of ABB Group's revenues.
Write about **150** words within **25** minutes.

ABB Group Revenues Breakdown, 2015 & 2016 (£million)



Writing task 3

Your company is about to release a new sports app: SportsCenter. This mobile app brings users the latest sports news and offers personalized information updates. You are asked to write a proposal to market the app.
Write a **proposal** to your marketing manager, including the following information:

- an outline of the features of the app;
- suggestions for promoting the app;
- an explanation of the benefits of your proposed promotional strategies;
- discussion of the challenges to your proposed strategies.

Write **300-350** words within **40** minutes.

Appendix 2. ETIC rating scale for written tasks (translated version)

Bands	Content	Organization	Language
5	• All content is closely relevant to the topic. • Fully addresses all the requirements of the task. • Provides abundant and accurate supporting information.	• Text is well organized and coherent. • Uses a wide range of cohesive devices naturally. • Accurate format.	• Demonstrates an accurate word choice. • Syntactic variety and flexibility. • Good control of grammar.
4	• All content is relevant to the topic. • Addresses all the requirements of the task. • Provides sufficient supporting information.	• Text is organized and coherent. • Uses a range of cohesive devices effectively. • Appropriate format.	• Demonstrates an effective word choice. • A range of syntactic structures. • Sufficient control of grammar (occasional errors do not obscure meaning).
3	• Most content is relevant to the topic (minor irrelevances). • Basically fulfills the requirements of the task. • Provides some supporting information.	• Text is generally organized and coherent. • Uses linking words and cohesive devices. • Basically correct format.	• Demonstrates an acceptable word choice. • A limited range of syntactic structures. • Limited control of grammar (most errors do not interfere with meaning).
2	• Shows obvious irrelevances to the topic. • Fails to fulfill the requirements of the task (omission of one key point in the requirements).	• Text is less connected and coherent. • Uses a limited number of linking words cohesive devices (inaccurate/repetitive/ under-/over-use). • Inappropriate format.	• Demonstrates an inaccurate choice of words. • Simple sentence structure. • Simple grammatical forms (errors impede meaning at times).
1	• Most part of the response is irrelevant to the topic. • Fails to fulfill the requirements of the task (omission of two key points in the requirements).	• Illogical organization.	• Contains many errors which impede and distort meaning.
0	• Presents totally irrelevant contents or no response.		