International Journal of English Linguistics; Vol. 11, No. 1; 2021
ISSN 1923-869X E-ISSN 1923-8703
Published by Canadian Center of Science and Education

Development and Validation of a Diagnostic Rating Scale for EFL Writing in China

Yixi Lu[1], Qiqi Han[1], Zhaoxu Fang[1] & Antian Shen[1]

[1] Department of Linguistics, Zhejiang University, Hangzhou, China

Correspondence: Yixi Lu, Department of Linguistics, Zhejiang University, Hangzhou, Zhejiang, 310 058, China. E-mail: laoyouaa@163.com

Received: September 16, 2020 Accepted: October 18, 2020 Online Published: October 29, 2020

doi:10.5539/ijel.v11n1p32 URL: https://doi.org/10.5539/ijel.v11n1p32

Abstract

Diagnostic assessment of EFL writing ability is useful yet seldom adopted for Chinese EFL students. In line with this urge, this study intends to design and validate a diagnostic rating scale for EFL writing in China. This rating scale is adapted from China's Standards of English Language Ability (CSE in short) for an argumentative writing assignment of College English III students at a key university in Eastern China. To collect data for validation, four raters were asked to score 67 compositions utilizing the rating scale. A multi-facet Rasch analysis was employed to investigate the validity of the rating scale. Three facets—examinee, rater, and criteria—basically accord with the ideal requirements. Comparing our validated rating scale and rating scales for writing assessment designed in other contexts, the importance of setting rating scales in a specific context is demonstrated. Additionally, our context-specific, CSE-based rating scale once again corroborates the versatility of CSE. This study provides a meaningful examination of the appropriate form of a rating scale for diagnostic assessment in China.

Keywords: diagnostic assessment, rating scale, validity, EFL writing ability, China's Standards of English

1. Introduction

As a major form of language assessment, diagnostic assessment has received increasing attention in the language assessment literature. Following this trend, this study intends to design a diagnostic rating scale for English writing ability based on China's Standards of English Language Ability (CSE) and investigate its validity through a multi-facet Rasch model analysis on the results of an argumentative writing task. Setting in the context of a writing test for Chinese undergraduates, EFL students in a key university in Eastern China, this study illustrates the importance of designing rating scales with the specific context, and provides Chinese EFL teachers with a validated diagnostic assessment rating scale for writing ability.

Firstly, this study reviews previous literature in diagnostic assessment for writing ability, paying extra attention to the rating scale, theoretical framework of diagnostic assessment and validation of rating scales. Secondly, a detailed presentation of our research design is provided. Moreover, the results of our research are analyzed and discussed, and the findings are stated. Implications and limitations of this study are included in the last section of this paper.

2. Literature Review

2.1 Diagnostic Assessment for Writing Ability

Written language is one of the basic ways of communication and verbal expression that enables people from different cultures and backgrounds to participate in all aspects of today's global society. With the rapid development of technology and communication in countries around the world, to write in a second or foreign language has become an important skill. Therefore, improving the ability to write well is a need for all students in academic and second foreign language courses, not only to generate new information but also to impart knowledge (Lee & Sawaki, 2009).

To write efficiently, writers are expected to master a series of abilities and strategies including coherence, logic construction, etc. Lack of these skills can impede the writing quality. Thus, timely feedback is necessary for the future improvement. The last few decades have witnessed the interest in the L2 writing feedback. However, accurate diagnostic information about the strengths and weaknesses of L2 learners' writing is hard to obtain

based on the current feedback system (Llosa et al., 2011).

Diagnostic assessment, as an alternative to existing feedback methods, has recently received more attention from educational experts due to its ability to provide diagnostic information about students' strengths and weaknesses. Diagnostic assessment is defined as using tests "for the purpose of discovering a learner's specific strengths or weaknesses" to inform "decisions on future training, learning, or teaching" (ALTE, 1998), which is found to be effective in identifying students' problems and providing possible solutions, for its integration of assessment and instruction (Pellegrino & Chudowsky, 2003).

Unfortunately, diagnostic assessment of EFL writing ability is seldom adopted for Chinese EFL students. As Ross (2008) and Matoush (2012) pointed out, teachers and researchers are worried by the fact that traditional high-stake, large-scale proficiency tests dominate Chinese EFL teaching and testing, and individualized learning needs have been largely neglected. Moreover, very few studies have explored diagnostic assessment for EFL writing ability in the Chinese context and the need for practical studies is apparent. In line with this urge, this research focuses on the rating scale of diagnostic assessment in the context of a writing test for Chinese undergraduate EFL students.

2.2 Rating Scales for Diagnostic Writing Assessments

As a crucial part of instrumentation, the rating scale is the major concern of diagnostic assessment (Lee, 2015). There has been a collective realization that rating scales designed for conventional tests are considered not suitable for diagnostic assessment (Alderson, 2005, 2015; Knoch, 2011; Upshur & Turner, 1995). To cite an example, Weigle (2002) demonstrated that the analytical feedback of conventional tests given by raters is likely to be imprecise due to the "halo effect", that is, the interference from the overall impression on certain single aspects of the writing script. As a result, rating scales designed specifically for diagnostic assessment are necessary to accomplish the special purpose of diagnostic writing assessment.

Weigle (2002) proposed several questions during rating scale development, which can be summarized as "What type of rating scale is desired?" "Who is going to use the rating scale?" "What criteria should be used as the basis for the rating?" "What will the descriptors look like and how many scoring levels will be used?" and "How will scores be reported?". Given the context of diagnostic assessment, the first two questions are automatically answered (It should be an analytical, assessor-oriented scale). The third question is elaborated on in the following section while the fourth and fifth questions are briefly answered in the methodology section of this paper. As further clarification, Knoch (2011) pointed out that the several factors specified in the questions must be weighed as a whole. That is, the format (descriptors and scoring levels) and theoretical orientation (criteria) of the rating scale must be appropriate for the context (e.g., the characteristics of the test-takers and assessors) in use. To illustrate this point, when Bruce and Hamp-Lyons (2015) tried to design an assessment tool for a new cohort of Hong Kong Dioploma of Secondary Education Examination students in Hong Kong City University, they found opposing tensions between local and international rating scales, claiming that aligning scales based on English for Academic Purposes needs and ones of CEFR or IELTS will cause trouble. Seen in this light, although there are several available standards for assessment of writing such as ACTFL and CEFR, the contexts in which they are set are not even close to the context of writing tests for Chinese undergraduate EFL students, as a result, they may not function satisfactorily in the context of our research. Our research, therefore, has both the theoretical significance of providing a reference for future Chinese diagnostic assessment rating scales designing and the applicable value of helping Chinese undergraduate EFL students with their English writing.

2.3 Theoretical Frameworks of Writing Development

It has been widely argued that the criteria in a diagnostic rating scale for writing should be based on a theory or a model of language or writing development (Behizadeh & Engelhard, 2011; Dolgova & Siczek, 2019; Knoch, 2011; Read & von Randow, 2013). Furthermore, in a project of examining how assessors inform themselves of using assessment for Canadian university-learning EAP courses, Doe (2015) chose The Canadian Academic English Language (CAEL) as the assessment framework, explaining that the students should be familiar with assessments based on the framework and the raters should be familiar with the framework itself. Therefore, the appropriate theoretical model used in our research should not only be comprehensive enough but also suitable for Chinese undergraduate EFL students and familiar enough for the raters.

With the task-based, topic-based writing test as our research context in mind, we compared several available frameworks for English writing ability. The result was immediately clear. As the recent assessment framework developed by the National Education Examinations Authority, CSE provides a comprehensive description of English proficiency skills based on English language education in China. Besides, English teachers and researchers, as assessors, are adequately familiar with rating scales developed from CSE. Thus, CSE can best

reflect our understanding of the model where we should base our rating scale.

Although no study has explored any assessor-oriented rating scale for diagnostic writing assessment for Chinese students, relevant studies on assessment in China provide us with important information and caution on designing our rating scale based on CSE. To name a few, in a project of designing a self-assessment scale, Pan, Song and Deng (2019) specified that CSE Band 4–7 aligns with the non-English major undergraduate students.

Fulcher (2003) offered a provisional model consisting of four factors (rater, rating scale, task, and candidate) that influence the score of a test taker, and indicated that the probable interaction between rating scales and the others factors requires further research. Although former studies (Knoch, 2009, 2011) have supported the notion that rating scales should be designed with the specific testing context (e.g., task and candidate) in mind, few studies have explored to which extent does a different context (e.g., a test for Chinese undergraduate EFL students in the case of our research) influence the diagnostic rating scale designed accordingly. By qualitative comparison of the similarities and differences between our rating scale and rating scales designed in other contexts (e.g., Hawkey, 2001; Kuiken & Vedder, 2017; Madsen, 1983), the influence of context-specific factors on rating scales and the importance of setting rating scales in a specific context can be demonstrated (or disproved).

2.4 Validation of Diagnostic Rating Scales

Validation of a rating scale is needed since it is considered a fundamental concern of testing (Messick, 1989) and is associated with fairness and social responsibilities of a test (Kane, 2010). Multi-facet Rasch model analysis (MFRM) and generalizing theory (G-theory) are two mainstream validating approaches in recent studies (Beglar, 2009; Hsieh, 2013; Bochner, 2015). Through comparison, a multi-facet Rasch analysis is selected as the specific method for validation, as MFRM provides finer and more individual analyses and performs better in analyzing how test bias occurs than G-theory does (Kim & Wilson, 2009; McNamara & Knoch, 2012; Sudweeks, Reeve, & Bradshaw, 2004). In more concrete terms, MFRM is an extended Rasch model where test setting factors (e.g., task difficulty, test-taker ability, rater severity) are named facets. In writing assessment, ratings are viewed as the outcome of interactions between different facets. By MFRM, analyses of different facets and interactions of facets are conducted on the same logit scale in a statistical software called FACETS. Besides, different elements of a certain facet are detected, and fit statistics of each element underlying facets are provided, thus yielding detailed identification of validity of the assessment measure (Barkaoui, 2013; Linacre, 1993, 1994).

Based on Knoch's (2007, 2009, 2011) validation framework, to validate a rating scale, one needs to explore the validity of five dimensions of the scale, which are discrimination of rating scale, rater separation, rater reliability, variations in rating, and scale step functionality. Knoch (2007) compared the validation of two rating scale—Diagnostic English Language Needs Assessment scale and a self-developed one—by examining the aforementioned five dimensions, and acquired abundant evidence to support the validity of the self-developed scale. Grounded in Knoch's validation framework, we adjust our validation process to seven dimensions in accord with features of FACETS software, which is presented in the Results section.

3. Methodology

Based on the aforementioned literature, this research aims to investigate the validity of a CSE-adapted diagnostic rating scale for English writing ability.

3.1 Research Question

This study intends to answer the following research question: How valid it is of the CSE-adapted diagnostic assessment scale for English writing ability, in the context of Chinese EFL?

3.2 Participants

Sixty-seven students (seventeen females and fifty males) from different majors were selected from the course "College English III" at a key university in Eastern China to participate in the writing test task. After a year of English learning on the course of College English I & II, the English proficiency levels of the majority lied in the range from intermediate English learners to advanced ones, while the most were evaluated by the teachers as upper-intermediate English learners.

3.3 Instruments

3.3.1 The Writing Task

The task adopted in this study is an argumentative writing assignment of College English III students. A statement and a question are given as the prompt:

> Children today do not have much play time as parents believe a busy schedule of study, arts class and

> sports, etc., is good for their development. Write an essay entitled Should Children be Allowed More Time to Play? You should use specific reasons and examples to support your view. Your essay should be no shorter than 150 words.

Participants performed the untimed task without being informed of the purpose of the study beforehand, to make sure the authenticity of their writings.

3.3.2 The Rating Scale

Based on China Standards of English, the rating scale consists of two columns. The left is descriptors and the right marks. It assesses writings from three aspects: language quality, essay structure, and task completion. Language quality contains two dimensions: vocabulary and syntax; essay structure also contains two dimensions: discourse and coherence. Band 4–7 are chosen because they are reported to be in accordance with the ability of college students apart from English majors in China (Pan, 2019).

As Alderson (2005) and Weigle (2002) have suggested, current diagnostic rating scales usually employ vague or impressionistic terms which might confuse raters and hence impede both the validity of the scale and informativity of the test. From this perspective, we try to confine the descriptors to clear, concrete, and objective (Raffaldini, 1988) terminology as much as possible. We set five bands for each aspect of the scale, in part because the reliability was reported to be highest for scales with 5–9 bands (Miller, 1956; Myford, 2002). We also decide that both the combined score and sub-scores should be reported, for it is an accepted view that the scores should offer as much feedback as possible to students (Knoch, 2009; Alderson, 1991).

3.4 Procedure

Firstly, participants were assigned the aforementioned writing task. After finishing the untimed argumentative writing, they sent the electronic version of the writings back to the researchers. Before the official rating process started, to make sure raters understand the descriptors unanimously and correctly, a pilot rating was implemented by four raters. When the agreement of ratings on 5 samples reached 80%, the four raters were allowed to rate all the sixty-seven compositions according to the diagnostic assessment rating scale. All four raters, who were both postgraduate students in English linguistics and teaching assistants of the course College English, had received the professional rating training. Once finished, the rating results were sent back to the researchers.

3.5 Data Analysis

The software *FACETS* for Windows No.3.80.0 (Linacre, 2017) was selected as the instrument of multi-facet Rasch analysis. Two steps were involved in the data analysis process. At first, a multi-facets model was established. Then, a multi-facet Rasch analysis was taken to evaluate the scoring outcome to investigate the validity of a self-developed diagnostic rating scale through FACETS.

Based on the research aim, a multi-facets model was established:

$$\log\left(P_{irjk}/P_{irj(k-1)}\right) = \theta_i - R_r - \delta_j - F_k \quad (1)$$

P_{irjk} is the probability that on the certain item j, the rater r gives k points to student i. $P_{irj(k-1)}$ is the probability that on the certain item j, the rater r gives $(k-1)$ points to student i. θ_i is the student's writing skill. R_r is the harshness of the rater. δ_j is the mean item difficulty estimated for each item. F_k is the difficulty of receiving a k based on $(k-1)$ points.

4. Results

The results will be presented from seven aspects: overall statistics, Knoch's (2007, 2009) five components for validation of a rating scale (discrimination of the rating scale, rater separation, rater reliability, variation in ratings, scale step functionality), and criteria facet.

4.1 Overall Statistics

Figure 1 presents the overall status of all three facets which are all measured by the same vertical axis in the unit of logit. With the "examinee" facet being tagged as positive, it is suggested that the higher the measure is, the more proficient in writing the examinee is. And from the figure, a normal distribution can be roughly observed in the column of "examinee", which means the examinees are rated appropriately and differentially using the rating scale. While the "rater" facet was tagged as negative, it indicates that the first rater (R1) and the fourth rater (R4) is the most severe in giving high scores and the third rater (R3) is the most lenient. Seeing that all the raters crowd together, which implies that there is no deviation for raters' ratings, the inter-rater agreement is satisfied. The "criteria" facet is tagged as negative as well, from which we may state that, among five aspects of the

criteria, in the item of "coherence", raters tend to give lower marks, and in "argument", raters are likely to give a high mark, but it can be noted that five criteria cluster together, implying that there is no significant difference among criteria.

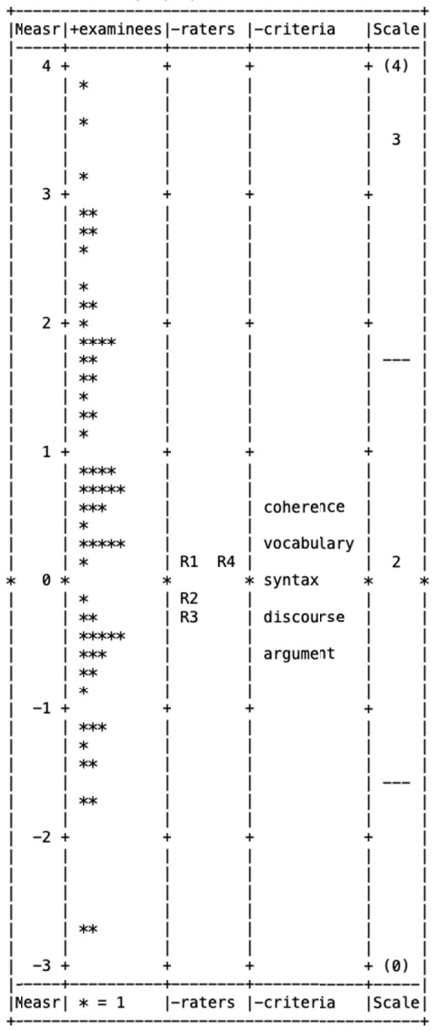
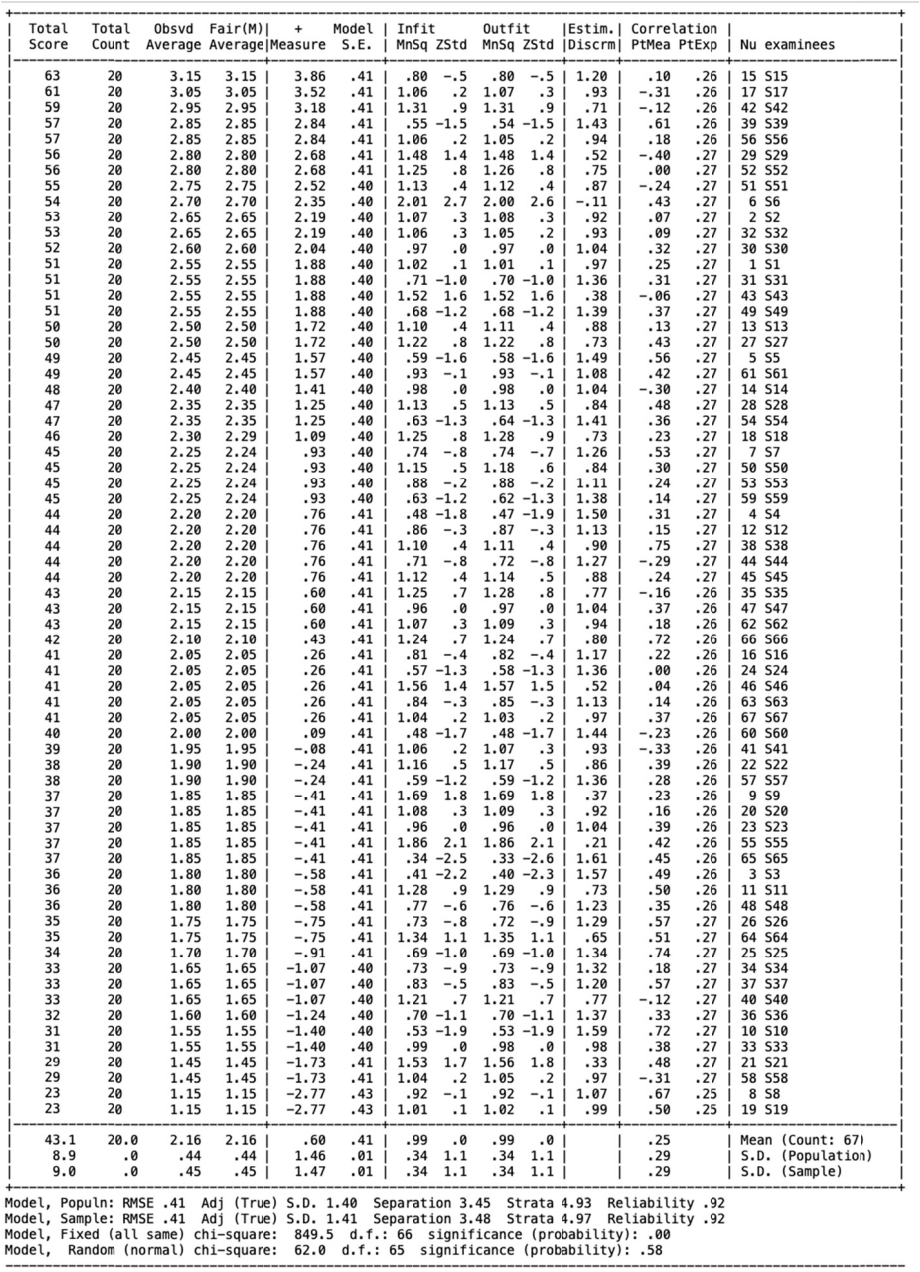


Figure 1. All facets vertical ruler

4.2 Discrimination of the Rating Scale

Discrimination of the rating scale can be reflected in two indicators: separation and fixed chi-square (see Figure 2). Firstly, the separation ratio of candidates is 3.45, higher than 2, which is considered as the threshold of the existence of individual differences (Myford & Wolfe, 2004). In other words, we can infer from this result that the rating scale is capable to differentiate examinees of different abilities. It is notable that the reliability here specifically refers to that of separation ratio, instead of rater reliability, and the reliability of .92 further affirms the significant differences among candidates. The second indicator is fixed chi-square, which implies candidates' differences. Here, with the value of fixed chi-square of 849.5 and significance of .00, it suggests that there is not a group-level central tendency and the degree of discrimination is acceptable.



Total Score	Total Count	Obsvd Average	Fair(M) Average	+ Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	PtExp	Nu examinees
63	20	3.15	3.15	3.86	.41	.80	-.5	.80	-.5	1.20	.10	.26	15 S15
61	20	3.05	3.05	3.52	.41	1.06	.2	1.07	.3	.93	-.31	.26	17 S17
59	20	2.95	2.95	3.18	.41	1.31	.9	1.31	.9	.71	-.12	.26	42 S42
57	20	2.85	2.85	2.84	.41	.55	-1.5	.54	-1.5	1.43	.61	.26	39 S39
57	20	2.85	2.85	2.84	.41	1.06	.2	1.05	.2	.94	.18	.26	56 S56
56	20	2.80	2.80	2.68	.41	1.48	1.4	1.48	1.4	.52	-.40	.27	29 S29
56	20	2.80	2.80	2.68	.41	1.25	.8	1.26	.8	.75	.00	.27	52 S52
55	20	2.75	2.75	2.52	.40	1.13	.4	1.12	.4	.87	-.24	.27	51 S51
54	20	2.70	2.70	2.35	.40	2.01	2.7	2.00	2.6	-.11	.43	.27	6 S6
53	20	2.65	2.65	2.19	.40	1.07	.3	1.08	.3	.92	.07	.27	2 S2
53	20	2.65	2.65	2.19	.40	1.06	.3	1.05	.2	.93	.09	.27	32 S32
52	20	2.60	2.60	2.04	.40	.97	.0	.97	.0	1.04	.32	.27	30 S30
51	20	2.55	2.55	1.88	.40	1.02	.1	1.01	.1	.97	.25	.27	1 S1
51	20	2.55	2.55	1.88	.40	.71	-1.0	.70	-1.0	1.36	.31	.27	31 S31
51	20	2.55	2.55	1.88	.40	1.52	1.6	1.52	1.6	.38	-.06	.27	43 S43
51	20	2.55	2.55	1.88	.40	.68	-1.2	.68	-1.2	1.39	.37	.27	49 S49
50	20	2.50	2.50	1.72	.40	1.10	.4	1.11	.4	.88	.13	.27	13 S13
50	20	2.50	2.50	1.72	.40	1.22	.8	1.22	.8	.73	.43	.27	27 S27
49	20	2.45	2.45	1.57	.40	.59	-1.6	.58	-1.6	1.49	.56	.27	5 S5
49	20	2.45	2.45	1.57	.40	.93	-.1	.93	-.1	1.08	.42	.27	61 S61
48	20	2.40	2.40	1.41	.40	.98	.0	.98	.0	1.04	-.30	.27	14 S14
47	20	2.35	2.35	1.25	.40	1.13	.5	1.13	.5	.84	.48	.27	28 S28
47	20	2.35	2.35	1.25	.40	.63	-1.3	.64	-1.3	1.41	.36	.27	54 S54
46	20	2.30	2.29	1.09	.40	1.25	.8	1.28	.9	.73	.23	.27	18 S18
45	20	2.25	2.24	.93	.40	.74	-.8	.74	-.7	1.26	.53	.27	7 S7
45	20	2.25	2.24	.93	.40	1.15	.5	1.18	.6	.84	.30	.27	50 S50
45	20	2.25	2.24	.93	.40	.88	-.2	.88	-.2	1.11	.24	.27	53 S53
45	20	2.25	2.24	.93	.40	.63	-1.2	.62	-1.3	1.38	.14	.27	59 S59
44	20	2.20	2.20	.76	.41	.48	-1.8	.47	-1.9	1.50	.31	.27	4 S4
44	20	2.20	2.20	.76	.41	.86	-.3	.87	-.3	1.13	.15	.27	12 S12
44	20	2.20	2.20	.76	.41	1.10	.4	1.11	.4	.90	.75	.27	38 S38
44	20	2.20	2.20	.76	.41	.71	-.8	.72	-.8	1.27	-.29	.27	44 S44
44	20	2.20	2.20	.76	.41	1.12	.4	1.14	.5	.88	.24	.27	45 S45
43	20	2.15	2.15	.60	.41	1.25	.7	1.28	.8	.77	-.16	.26	35 S35
43	20	2.15	2.15	.60	.41	.96	.0	.97	.0	1.04	.37	.26	47 S47
43	20	2.15	2.15	.60	.41	1.07	.3	1.09	.3	.94	.18	.26	62 S62
42	20	2.10	2.10	.43	.41	1.24	.7	1.24	.7	.80	.72	.26	66 S66
41	20	2.05	2.05	.26	.41	.81	-.4	.82	-.4	1.17	.22	.26	16 S16
41	20	2.05	2.05	.26	.41	.57	-1.3	.58	-1.3	1.36	.00	.26	24 S24
41	20	2.05	2.05	.26	.41	1.56	1.4	1.57	1.5	.52	.04	.26	46 S46
41	20	2.05	2.05	.26	.41	.84	-.3	.85	-.3	1.13	.14	.26	63 S63
41	20	2.05	2.05	.26	.41	1.04	.2	1.03	.2	.97	.37	.26	67 S67
40	20	2.00	2.00	.09	.41	.48	-1.7	.48	-1.7	1.44	-.23	.26	60 S60
39	20	1.95	1.95	-.08	.41	1.06	.2	1.07	.3	.93	-.33	.26	41 S41
38	20	1.90	1.90	-.24	.41	1.16	.5	1.17	.5	.86	.39	.26	22 S22
38	20	1.90	1.90	-.24	.41	.59	-1.2	.59	-1.2	1.36	.28	.26	57 S57
37	20	1.85	1.85	-.41	.41	1.69	1.8	1.69	1.8	.37	.23	.26	9 S9
37	20	1.85	1.85	-.41	.41	1.08	.3	1.09	.3	.92	.16	.26	20 S20
37	20	1.85	1.85	-.41	.41	.96	.0	.96	.0	1.04	.39	.26	23 S23
37	20	1.85	1.85	-.41	.41	1.86	2.1	1.86	2.1	.21	.42	.26	55 S55
37	20	1.85	1.85	-.41	.41	.34	-2.5	.33	-2.6	1.61	.45	.26	65 S65
36	20	1.80	1.80	-.58	.41	.41	-2.2	.40	-2.3	1.57	.49	.26	3 S3
36	20	1.80	1.80	-.58	.41	1.28	.9	1.29	.9	.73	.50	.26	11 S11
36	20	1.80	1.80	-.58	.41	.77	-.6	.76	-.6	1.23	.35	.26	48 S48
35	20	1.75	1.75	-.75	.41	.73	-.8	.72	-.9	1.29	.57	.27	26 S26
35	20	1.75	1.75	-.75	.41	1.34	1.1	1.35	1.1	.65	.51	.27	64 S64
34	20	1.70	1.70	-.91	.41	.69	-1.0	.69	-1.0	1.34	.74	.27	25 S25
33	20	1.65	1.65	-1.07	.40	.73	-.9	.73	-.9	1.32	.18	.27	34 S34
33	20	1.65	1.65	-1.07	.40	.83	-.5	.83	-.5	1.20	.57	.27	37 S37
33	20	1.65	1.65	-1.07	.40	1.21	.7	1.21	.7	.77	-.12	.27	40 S40
32	20	1.60	1.60	-1.24	.40	.70	-1.1	.70	-1.1	1.37	.33	.27	36 S36
31	20	1.55	1.55	-1.40	.40	.53	-1.9	.53	-1.9	1.59	.72	.27	10 S10
31	20	1.55	1.55	-1.40	.40	.99	.0	.98	.0	.98	.38	.27	33 S33
29	20	1.45	1.45	-1.73	.41	1.53	1.7	1.56	1.8	.33	.48	.27	21 S21
29	20	1.45	1.45	-1.73	.41	1.04	.2	1.05	.2	.97	-.31	.27	58 S58
23	20	1.15	1.15	-2.77	.43	.92	-.1	.92	-.1	1.07	.67	.25	8 S8
23	20	1.15	1.15	-2.77	.43	1.01	.1	1.02	.1	.99	.50	.25	19 S19
43.1	20.0	2.16	2.16	.60	.41	.99	.0	.99	.0		.25		Mean (Count: 67)
8.9	.0	.44	.44	1.46	.01	.34	1.1	.34	1.1		.29		S.D. (Population)
9.0	.0	.45	.45	1.47	.01	.34	1.1	.34	1.1		.29		S.D. (Sample)

Model, Populn: RMSE .41 Adj (True) S.D. 1.40 Separation 3.45 Strata 4.93 Reliability .92
Model, Sample: RMSE .41 Adj (True) S.D. 1.41 Separation 3.48 Strata 4.97 Reliability .92
Model, Fixed (all same) chi-square: 849.5 d.f.: 66 significance (probability): .00
Model, Random (normal) chi-square: 62.0 d.f.: 65 significance (probability): .58

Figure 2. Examinee measurement report

4.3 Rater Separation

Rater separation is an important factor in the applicability of rating scales. A valid rating scale will result in small differences between raters in terms of harshness and leniency. The higher the rater separation ratio, the greater the difference in the harshness of the ratings. Lower rater separation ratios are reported in this study. The rater separation ratio is represented in the results table as a mixed chi-square as well as reliability. It is calculated by dividing the precision of these measures by the distribution of the measure of rater harshness (Myford & Wolfe, 2004).

For intra-rater reliability, "Measure" indicates the Rasch measure of the raters' leniency. Because the facet of rater leniency is negatively oriented, a higher score shows lower leniency, that is, a higher harshness of a rater. From Figure 1, R4 is the harshest while R3 is the most lenient. Another indicator "Infit MnSq" is the information-weighted, inlier-pattern-sensitive, mean-square fit statistic, meaning how much do the actual ratings fit the Rasch model, with its expectation being 1.0, and value between (0.6–1.5) is acceptable. From Figure 3, R1-R4 respectively report 1.19, .96, .90, and .93. Thus, these raters are qualified for the ratings.

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N raters
703	335	2.10	2.09	.19	.10	1.19	2.3	1.19	2.3	.81	.72	.64	4 R4
705	335	2.10	2.10	.17	.10	.96	-.5	.96	-.4	1.05	.62	.64	1 R1
733	335	2.19	2.13	-.10	.10	.90	-1.3	.90	-1.3	1.11	.67	.64	2 R2
750	335	2.24	2.23	-.27	.10	.93	-.9	.93	-.8	1.07	.58	.64	3 R3
722.8	335.0	2.16	2.15	.00	.10	.99	-.1	.99	-.1		.64		Mean (Count: 4)
19.7	.0	.06	.05	.19	.00	.11	1.4	.11	1.4		.05		S.D. (Population)
22.8	.0	.07	.07	.22	.00	.13	1.7	.13	1.6		.06		S.D. (Sample)

Model, Populn: RMSE .10 Adj (True) S.D. .17 Separation 1.67 Strata 2.56 Reliability .74
Model, Sample: RMSE .10 Adj (True) S.D. .20 Separation 2.01 Strata 3.02 Reliability .80
Model, Fixed (all same) chi-square: 15.2 d.f.: 3 significance (probability): .00
Model, Random (normal) chi-square: 2.5 d.f.: 2 significance (probability): .28

Figure 3. Rater measurement table

4.4 Rater Reliability

The next parameter to be examined in the Rasch analysis is rater reliability (Davies & Elder, 2005). Ideally, a scale that has higher rater reliability, it will be considered dependable. FACETS can provide an indicator of rater reliability, the rater point-biserial correlation index, which is the indicator of how similar the raters are to each other in ranking candidates.

The rater point-biserial correlation index shows the consistency of a given rater's score with the scores of other raters. Myford and Wolfe (2004) state that the rater point-biserial correlation index of less than 0.30 is considered low, while a correlation index of 0.70 or higher is desirable. In this study, we can tell from Figure 3 that the mean rater point-biserial correlation index ("Correlation PtMea") is 0.64, which roughly meets the criteria.

The Separation (expectation being < 2), Reliability (expectation being 1.0), and Chi-square (expectation being 0) values under the figure are statistical indicators of how different the raters are from each other. In this case, the three values from Figure 3 are 2.01, 0.80, and 0.00 respectively, indicating that the inter-rater difference is not significant and the ratings are consistent.

4.5 Variation in Ratings

The fourth indicator of the scale validity is variation in ratings, which can be observed in the rater infit mean square value. An ideal value of rater infit is 1. According to Myford and Wolfe (2000), the infit mean square value which is higher than 1.3 means that ratings reflect too much variation, and raters rate inconsistently and far away from the predictions by the model, which can be labeled as "misfit". On the contrary, the infit mean square value which is lower than .70 indicates that ratings are closer to each other than what they are supposed to. In other words, the rater is rating over consistently and he or she tends to overuse the intermediate values, such as "3" on a five-point Likert scale. According to Figure 3, the infit mean square value of rater 4 is 1.19, followed by rater 1(.96), rater 3(.93), and rater 2(.90), all of which are between .70 and 1.3. The mean of them is .99, which is close to 1, the ideal value of rater infit. The result shows that variation in ratings is appropriate, which provides the evidence for the scale validity.

4.6 Scale Step Functionality

The last indicator for a well-functioned rating scale is scale step functionality, which mainly examines the functionality of different band levels. FACETS output provides category statistics and probability curves for the examination of scale step functionality.

For the rating scale to be valid, the following requirements should be met: the value of category total need to be larger than 10 for each band level, the values of average measures need to increase monotonically (Knoch, 2007), and so do the values of thresholds measure, besides, the gap among each band level should lie in 1.0–5.0 logits (Linacre, 1999). Here it is noted from Figure 4, all requirements are fulfilled but for band level 0, whose category total is lower than 10. This may be explained by insufficient utilization of band level 0 or ambiguous descriptors in band level 0 in the rating scale. However, overall, the rating scale functions appropriately in each band.

DATA				QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat
Score	Category Total	Counts Used	Cum. %	Avge Meas	Exp. Meas	OUTFIT MnSq	Thresholds Measure	S.E.	Measure at Category	-0.5	PROBABLE from	THURSTONE Thresholds	PEAK Prob
0	6	6	0% 0%	-2.92	-2.18	.7			(-6.25)		low	low	100%
1	212	212	16% 16%	-1.06	-.99	.9	-5.17	.42	-3.33	-5.22	-5.17	-5.19	75%
2	716	716	53% 70%	.41	.37	1.0	-1.54	.09	.09	-1.55	-1.54	-1.55	72%
3	377	377	28% 98%	1.78	1.80	1.0	1.73	.07	3.35	1.73	1.73	1.72	72%
4	29	29	2% 100%	2.72	2.94	1.1	4.97	.26	(6.08)	5.06	4.97	5.00	100%
							(Mean)				(Modal)	(Median)	

Figure 4. Category statistics

Another indicator of scale step functionality is probability curves (see Figure 5). The horizontal axis stands for the candidates' proficiency and the vertical axis is the probability of getting a given score. As is presented in the figure, band levels 1, 2, and 3 all have a peak, which is considered appropriate for a reliable step functionality.

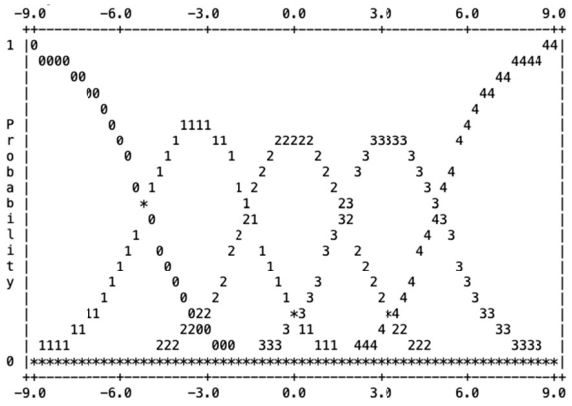


Figure 5. Probability curves

4.7 Criteria Facet

What's more, Figure 6 also reveals the data of the criteria measurement report. Students are easily getting a good score from the aspect of "argument" while it's difficult for them to get a good score incoherence. The infit mean square values of the five criteria are 1.08, 1.03, 1.13, .96, and .76, all of which are between .70 to 1.3., from which we may infer that the five criteria are reasonable and sufficient to measure students' writings.

Total Score	Total Count	Obsvd Average	Fair(M) Average	- Measure	Model S.E.	Infit MnSq	ZStd	Outfit MnSq	ZStd	Estim. Discrm	Correlation PtMea	PtExp	N criteria
532	268	1.99	1.98	.57	.11	1.08	.9	1.10	1.1	.91	.58	.63	4 coherence
550	268	2.05	2.05	.35	.11	1.03	.3	1.02	.2	.98	.74	.63	1 vocabulary
574	268	2.14	2.13	.05	.11	1.13	1.4	1.13	1.4	.87	.61	.63	2 syntax
606	268	2.26	2.25	-.34	.11	.96	-.3	.97	-.2	1.03	.58	.63	3 discourse
629	268	2.35	2.34	-.62	.11	.76	-3.0	.75	-3.1	1.25	.65	.63	5 argument
578.2	268.0	2.16	2.15	.00	.11	.99	-.1	.99	-.1		.63		Mean (Count: 5)
35.5	.0	.13	.13	.44	.00	.13	1.6	.13	1.6		.06		S.D. (Population)
39.7	.0	.15	.14	.49	.00	.15	1.8	.15	1.8		.07		S.D. (Sample)

Model, Populn: RMSE .11 Adj (True) S.D. .42 Separation 3.79 Strata 5.39 Reliability .93
Model, Sample: RMSE .11 Adj (True) S.D. .47 Separation 4.27 Strata 6.02 Reliability .95
Model, Fixed (all same) chi-square: 76.8 d.f.: 4 significance (probability): .00
Model, Random (normal) chi-square: 3.8 d.f.: 3 significance (probability): .28

Figure 6. Criteria measurement report

5. Discussion

Comparing our validated rating scale and rating scales for writing assessment designed in other contexts, the crucial role of context-specific factors in designing rating scales is highlighted. Task quality and candidates are two domains that can best illustrate our point. In an attempt to design a common scale to describe L2 writing performance, Hawkey (2011) left out descriptors concerning persuasiveness or organization of arguments since these aspects are not of vital importance in all text genre/writing styles. In our context, however, students have to write an argumentative text on the specific topic of childhood playtime, which makes persuasiveness and organization of arguments crucial aspects for written performance. Hence relevant descriptors should be included

in our rating scale. When Kuiken and Vedder (2017) designed a rating scale for Dutch learners' EAP argumentative writing assessment, no descriptors were dedicated to variety/difficulty/complexity of vocabulary and syntax, while varied word choice and syntax are preferred in our rating scale. This is because students in the Netherlands were taught to keep their texts easily comprehensible and smoothly readable (Kuiken & Vedder, 2017), while Chinese EFL undergraduate students are encouraged to make more use of varied and somewhat complex words and syntactic structures as is reflected in CSE. The contrasting pedagogical methods may be explained by the contrasting language distance between English and the native languages. It is conceivable that rating scales in an improper task style or country will fail to distinguish between writing samples at different levels and fail to provide informative feedback for test takers. The importance of setting rating scales in a specific context is demonstrated.

Besides, we may suggest from this study that our context-specific, CSE-based rating scale once again corroborates the versatility of CSE. Different from another distinguished scale, CEFR, which distinctly clarifies its usage, CSE may be limited in that (Wang, 2018). However, with its abundant subscales, explorations of its usages are extensively unfolded, such as on writing (Pan, 2019), listening (Min, 2018), speaking (Jie, 2019), or interpreting (Wang, 2017). Pan (2019) developed a writing assessment scale for college EFL students, using language ability and language use strategy from CSE. Though Pan successfully validates the usage of CSE on writing assessment, his study uses the scale for students' self-assessment. This study, to some extent, provides a new perspective to consider the utility of CSE, and works as a supplement to studies of CSE' application to writing assessment, for what our research concentrates on is the extent to which EFL teachers could use the scale developed from CSE to appropriately and validly rate students' writings.

Based on the results, three variables—examinee, rater, and criteria—basically accord with the ideal requirements. Examinees' scores are presented as normal distribution and examinees of different abilities can be well-differentiated, both intra-rater reliability and inter reliability are satisfied, and five criteria function well in measuring the writing ability. However, as for the scale step functionality, compared with the normal function of band 1–4, band 0 seems to malfunction, which may result from the ambiguous existence of band 0. This is because in the scale, band 1–4 equals to CSE band 4–7, and those who fail to reach the CSE band 4 (including CSE band 1–3) are all scored 0. For one band to cover various proficiency levels, it is conceivable that band 0 malfunctions.

6. Conclusion

Based on the quantitative analysis above, this study successfully designed a diagnostic rating scale for English writing ability based on CSE, which fulfills the core framework of diagnostic assessment and proves to be differentiating for test takers. With the assistance of the multi-facet Rasch model, the study investigated three variables, including examinee, rater, and criteria, which were all satisfied with the model. The model is statistically efficient in discriminating examinees' writing abilities, capable of maintaining the consistency of intra-rater/inter-rater separation and reliability, and reliable with five criteria on assessing the writing skill of the examinees. Thus, it is feasible to design a diagnostic rating scale for assessing EFL students' writing ability.

This study investigated the validity of a rating scale specifically designed to diagnose EFL writing ability. It contributes, in part, to the relatively new field of research in diagnostic assessment. By developing a diagnostic scale, this study has provided a meaningful examination of the appropriate form of a diagnostic rating scale. In conclusion, the efforts of the present study in the experimental and validation process are important for the development of testing tools, especially rating scales, in diagnostic assessments with some insights. Furthermore, rather than simply examining the form or definition of diagnostic assessments, this research focuses on the relative under-researched application of multi-facet models in EFL testing. These attempts also help to clarify factors and methodologies worth noting in future research and even in the development of diagnostic assessments of EFL writing.

However, several improvements should be noted for future research on this topic. First, this study only investigated four raters on the writing performance assessment, and further research involving more raters with a diverse academic background is needed to demonstrate whether similar conclusions would be drawn from the same context. Second, because only one type of essay topic was explored, more research is needed to elucidate whether raters would exhibit bias in other topics. Finally, qualitative methods, such as interviews, may need to be used to examine how raters' academic training affects their perceptions of the scoring categories' perceptions, and how those perceptions can be a factor in scoring bias.

References

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*.

A&C Black. https://doi.org/10.1111/j.1540-4781.2007.00514_23.x

Alderson, J. C., Brunfaut, T., & Harding, L. (2015). Towards a theory of diagnosis in second and foreign language assessment: Insights from professional practice across diverse fields. *Applied Linguistics*, *36*(2), 236–260. https://doi.org/10.1093/applin/amt046

Alderson, J. C., North, B., & Council, B. (Eds.). (1991). *Language testing in the 1990s: The communicative legacy*. Macmillan: London.

ALTE. (1998). *Multilingual glossary of language testing terms* (Vol. 6). Cambridge University Press: Cambridge.

Barkaoui, K. (2013). Multifaceted Rasch analysis for test evaluation. *The Companion to Language Assessment*, *3*, 1301–1322. https://doi.org/10.1002/9781118411360.wbcla070

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, *27*(1), 101–118. https://doi.org/10.1177/0265532209340194

Behizadeh, N., & Engelhard Jr, G. (2011). Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing Writing*, *16*(3), 189–211. https://doi.org/10.1016/j.asw.2011.03.001

Bochner, J. H., Samar, V. J., Hauser, P. C., Garrison, W. M., Searls, J. M., & Sanders, C. A. (2015). Validity of the American sign language discrimination test. *Language Testing*, *15*(2), 158–180. https://doi.org/10.1177/0265532215590849

Bruce, E., & Hamp-Lyons, L. (2015). Opposing tensions of local and international standards for EAP writing programmes: Who are we assessing for? *Journal of English for Academic Purposes*, *18*, 64–77. https://doi.org/10.1016/j.jeap.2015.03.003

Doe, C. (2015). One teacher's take on using a 'test' for diagnostic purposes in the classroom. *Journal of English for Academic Purposes*, *18*, 40–50. https://doi.org/10.1016/j.jeap.2015.03.005

Dolgova, N., & Siczek, M. (2019). Assessment from the ground up: Developing and validating a usage-based diagnostic assessment procedure in a graduate EAP context. *Journal of English for Academic Purposes*, *41*, 100771. https://doi.org/10.1016/j.jeap.2019.100771

Fulcher, G. (2003). *Testing second language speaking*. London: Pearson Longman. https://doi.org/10.4324/9781315837376

Hawkey, R. (2001). Towards a common scale to describe L2 writing performance. *Cambridge Research Notes*, *5*, 9–13.

Hawkey, R., & Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing*, *9*(2), 122–159. https://doi.org/10.1016/j.asw.2004.06.001

Heinemann, A. W., Linacre, J. M., Wright, B. D., Hamilton, B. B., & Granger, C. (1993). Relationships between impairment and physical disability as measured by the functional independence measure. *Archives of Physical Medicine and Rehabilitation*, *74*(6), 566–573. https://doi.org/10.1016/0003-9993(93)90153-2

Hsieh, M. (2013). An application of multifaceted Rasch measurement in the Yes/No Angoff standard setting procedure. *Language Testing*, *30*(4), 491–512. https://doi.org/10.1177/0265532213476259

Jie, W. (2019). Relating English speaking tests to China's Standards of English: CET-SET 4 as a case study. *Foreign Language World*, *1*, 71–80.

Kane, M. (2010). Validity and fairness. *Language Testing*, *27*(2), 177–182. https://doi.org/10.1177/0265532209349467

Kim, S. C., & Wilson, M. (2009). A comparative analysis of the ratings in performance assessment using generalizability theory and the many-facet Rasch model. *Journal of Applied Measurement*, *10*(4), 408–423.

Knoch, U. (2007). *Diagnostic writing assessment: The development and validation of a rating scale*. Unpublished doctoral dissertation. University of Auckland, Auckland, New Zealand.

Knoch, U. (2009). *Diagnostic assessment of writing: The development and validation of a rating scale*. Frankfurt: Peter Lang. https://doi.org/10.3726/978-3-653-00929-3

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing*, *16*(2), 81–96. https://doi.org/10.1016/j.asw.2011.02.003

Kuiken, F., & Vedder, I. (2017). Functional adequacy in L2 writing: Towards a new rating scale. *Language Testing*, *34*(3), 321–336. https://doi.org/10.1177/0265532216663991

Lee, Y. W. (2015). Diagnosing diagnostic language assessment. *Language Testing*, *32*(3), 299–316. https://doi.org/10.1177/0265532214565387

Lee, Y. W., & Sawaki, Y. (2009). Cognitive Diagnosis Approaches to Language Assessment: An Overview. *Language Assessment Quarterly*, *6*(3), 172–189. https://doi.org/10.1080/15434300902985108

Linacre, J. (1994). The structure and stability of the functional independence measure. *Archives of Physical Medicine and Rehabilitation*, *75*. https://doi.org/10.1016/0003-9993(94)90384-0

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

Linacre. J. M. (2017). *Facets Rasch Measurement Computer Program*. Chicago: Winsteps.com

Llosa, L., Beck, S. W., & Zhao, C. G. (2011). An investigation of academic writing in secondary schools to inform the development of diagnostic classroom assessments. *Assessing Writing*, *16*(4), 256–273. https://doi.org/10.1016/j.asw.2011.07.001

Madsen, H. S. (1983). *Techniques in testing*. Oxford, UK: Oxford University Press.

Matoush, M. M., & Fu, D. (2012). Tests of English language as significant thresholds for college-bound Chinese and the washback of test-preparation. *Changing English*, *19*(1), 111–121. https://doi.org/10.1080/1358684x.2012.649176

McNamara, T. (1996). *Measuring second language performance*. Harlow, Essex: Pearson Education. https://doi.org/10.2307/330236

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing*, *29*(4), 555–576. https://doi.org/10.1177/0265532211430367

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11. https://doi.org/10.3102/0013189X018002005

Milanovic, M., Saville, N., Pollitt, A., & Cook, A. (1995). Developing rating scales for CASE: Theoretical concerns and analyses. In A. Cumming & R. Berwick (Eds.), *Validation in language testing*. Clevedon: Multilingual Matters.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97. https://doi.org/10.1037/h0043158

Min, S., He, L., & Luo, L. (2018). Validation of listening descriptors of China's Standards of English: An analysis of self-assessment data using polytomous IRT models. *Foreign Languages in China*, *2*, 72–81. https://doi.org/10.13564/j.cnki.issn.1672-9382.2018.02.010

Myford, C. M. (2002). Investigating design features of descriptive graphic rating scales. *Applied Measurement in Education*, *15*(2), 187–215. https://doi.org/10.1207/S15324818AME1502_04

Pan, M., Song, J., & Deng, H. (2019). Developing and validating the self-assessment scales in an online diagnostic test of English writing. *Foreign Language Education in China*, *2*(4), 33–41.

Pellegrino, J. W., & Chudowsky, N. (2003). FOCUS ARTICLE: The Foundations of Assessment. *Measurement: Interdisciplinary Research & Perspective*, *1*(2), 103–148. https://doi.org/10.1207/S15366359MEA0102_01

Raffaldini, T. (1988). The use of situation tests as measures of communicative ability. *Studies in Second Language Acquisition*, *10*(2), 197–216. https://doi.org/10.1017/S0272263100007312

Read, J., & von Randow, J. (2013). A university post-entry English language assessment: Charting the changes. *International Journal of English Studies*, *13*(2), 89–110. https://doi.org/10.6018/ijes.13.2.185931

Ross, S. J. (2008). Language testing in Asia: Evolution, innovation, and policy challenges. *Language Testing*, *25*(1), 5–13. https://doi.org/10.1177/0265532207083741

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. https://doi.org/10.1016/j.asw.2004.11.001

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal*, *49*(1), 3–12. https://doi.org/10.1093/elt/49.1.3

Wang, S. (2018). The application of China's Standards of English Language Ability in College English teaching. *Foreign Language Education*, *39*(4), 1–4. https://doi.org/10.16362/j.cnki.cn61-1023/h.2018.04.001

Wang, W. (2017). Construction and application of the interpreting scale in China's Standards of English: A case of formative assessment in interpreting teaching. *Foreign Language World*, *6*, 2–10.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511732997