Chapter 2
The Impact of Rating Scales on the CET-4 Writing: A Mixed Methods Study

Shaoyan Zou

Abstract Holistic rating scale and analytic rating scale are two types of scales which are frequently utilized by raters and EFL teachers in scoring EFL writing performance. Given their significant differences in terms of the underlying assumptions about writing constructs and development, as well as the implications for essay rating processes, whether one type is superior to the other remains a contentious issue. However, to date only a few studies have been conducted to empirically compare the effectiveness of the two types of scales in rating EFL writing performance. Therefore, this study was undertaken to examine the impact of two rating scales, one holistic and the other analytic, on the rating of writing scripts produced in College English Test Band 4 (CET-4) by adopting the Mixed Methods approach. Results of the study indicated that the analytic scale was as effective as the holistic one in terms of scale discrimination, rating reliability and scale level functionality; However, the analytic scale lent itself more easily to the control of rating variations. The raters in general held more positive views toward the analytic scale although suggestions were also proposed to conduct minor revisions to the scale descriptions. In conclusion, findings of the study revealed the potential of analytic rating scales in large-scale EFL writing assessments and also generated implications for rating scale development and validation in other research contexts.

Keywords Rating scales · CET writing · Mixed methods research

2.1 Introduction

In recent decades, spurred by test users' increasing demand for the interpretability of test scores (Chapelle et al., 2008), holistic rating scales, especially those commonly used in large-scale EFL writing assessments have aroused growing concerns (e.g. Knoch, 2009; Lee et al., 2010). The major concerns are that these holistic

S. Zou (✉)
Qingdao Agricultural University, Qingdao, China
e-mail: amandazsy@qau.edu.cn

L. Hamp-Lyons, Y. Jin (eds.), *Assessing the English Language Writing of Chinese Learners of English*, https://doi.org/10.1007/978-3-030-92762-2_2

scales are often designed intuitively, hence may not closely represent the features of the corresponding writing performance for one thing, and for another, the scoring criteria adopted in such holistic scales usually rely on impressionistic terminology which may be open to subjective interpretations (Knoch, 2009).

As such, analytic rating scales have gained renewed momentum particularly in the domain of EFL writing assessment. Compared with holistic rating scales, analytic rating scales are more advantageous in capturing examinees' specific weaknesses and strengths in writing, which is especially useful for second language learners whose writing skills are still under development (Lee et al., 2010). Therefore, the recent years has witnessed increasing research attempts to develop and validate analytic rating scales for EFL writing tests (e.g., Berger, 2015; Knoch, 2009; Lee et al., 2010).

Inspired by such a research trend, an analytic rating scale was developed for the writing assessment of College English Test Band Four (the CET-4 writing), one of the largest-scale EFL tests in China. Specifically, the scale is expected to adequately address the concerns over the holistic rating scale currently used in the CET-4 writing (e.g., Fei & Zhao, 2008; Jian & Lu, 2005). Featuring a multi-method approach, the scale developing process took advantage of document analysis, intuitive judgment and Rasch model analysis in pinpointing useful rating categories, providing adequate descriptors, and calibrating difficulty level of the descriptors. The analytic rating scale developed in this study is comprised of five levels and four rating categories (see Table 2.1 for a demo of the scale).

However, despite the endeavors involved in the scale developing stage, whether this analytic scale will be robust in scoring the CET-4 writing scripts or not still needs further scrutiny. Therefore, this study undertakes to examine the effectiveness of the analytic scale and the existing holistic scale. The aim is two-fold: one is to gather validity evidence for the newly-developed analytic scale, and to identify potential problems that jeopardize the scale validity, and the other is to enrich the meager literature concerning empirical comparison of the two types of rating scales, thereby exploring the prospect of analytic rating scale in large-scale EFL writing assessments.

2.2 Literature Review

2.2.1 A Review of Theoretical Arguments

According to Weigle (2002), holistic scoring involves assigning a single score to a script based on the rater's overall impression of the whole script. Whereas, analytic scoring takes into account a list of elements, each of which is judged separately and then assigned a single score (Hamp-Lyons, 2016).

Over the past decades, holistic scales have been widely used in writing assessment due to a number of advantages (see Shaw, 2004; Shaw & Weir, 2007; Weigle, 2002; White, 1984, 1985): First, they are highly efficient and practical in terms of

Table 2.1 A demo of the analytic rating scale

	Linguistic range and control	Content and idea	Discourse organization	Linguistic appropriacy
Level A	Can use a sufficient range of vocabulary to convey information and ideas effectively. Can use some fairly advanced words or phrases as well as common idioms, proverbs, sayings or expressions to convey meaning more precisely or to enhance the expressing effect. Both the choice of words and the use of collocations are accurate and idiomatic, thus making the expression clear and distinct. Can effectively and accurately use a variety of sentence patterns, syntactic or grammatical structures.	The content is closely related to the topic and the opinions are expressed in a clear and distinct manner. Can explain or illustrate issues and ideas with adequate examples or in proper ways. Can effectively organize and use facts and details to support one's arguments. Can produce clear, well-organized and well-developed text, demonstrating a sufficient understanding of the topic.	Can appropriately and flexibly use a wide variety of cohesive devices to indicate the logical relations between sentences and to make the text clear and well-organized. Can maintain the cohesion and consistency of the content through echoing the points mentioned before. Good paragraph structures, with clearly expressed main points and adequate supporting details. Has an adequate control of the connections between paragraphs as well as the logical relations between parts and the whole.	Can express appropriately and idiomatically in written language based on specific contexts. Has certain awareness of the audience, and the language style and register (formal or informal) are rather appropriate. Can express feelings or attitudes in a manner that is appropriate to the context. Can idiomatically use cultural references and figures of speech to convey information effectively.

time and cost. Second, they are useful for discriminating across a narrow range of assessment bands and suitable for arriving at a rapid overall rating, especially for large-scale assessments, thus enhancing rating reliability. Third, they focus the rater's attention on the strengths of the writing rather than the weaknesses so that writers are rewarded for what is well done instead of what is underachieved. However, holistic scales have been criticized for being devoid of any real theoretical underpinning, thus leading some researchers to challenge their validity (Shaw & Weir, 2007). With multiple categories collapsed into a single score, the same score from different raters may reflect vastly different constructs because raters, more often than not, bear their own rating criteria in mind during the rating process. For example, some raters pay more attention to grammatical accuracy, while others attach great importance to syntactic complexity. Another significant drawback of holistic scales lies in their inability to capture examinees' specific weaknesses and strengths in writing. According to Lee et al. (2010), this drawback can be even more conspicuous for EFL learners whose writing skills are still under development and who are more likely to show uneven profiles across different aspects of writing.

In contrast to holistic scales, the primary advantage of an analytic rating scale is that it can provide useful diagnostic information about a test taker's performance, especially his/her literacy progress (Hamp-Lyons, 1986, 1991; Shaw & Weir, 2007; Weigle, 2002). Moreover, analytic scales are considered more helpful in rater training as inexperienced raters can understand and apply rating criteria more easily (Weigle, 2002; Weir, 1990). Notwithstanding all the advantages, there are also some concerns over analytic rating scales, such as the high cost and the 'halo effect' problem associated with the use of such scales (Weigle, 2002), thus posing questions for the effectiveness of analytic scales in essay rating.

2.2.2 A Review of Empirical Research

Despite the seemingly abundant theoretical discussion, only a few empirical studies have been conducted to examine the effectiveness of the two types of scales. Moreover, findings of the existing research are a little mixed. On one hand, some studies found that the use of holistic scales could result in a higher generalizability coefficient and dependability coefficient, thus contributing to satisfactory inter-rater reliability (Barkaoui, 2007; O'Loughlin, 1994). For instance, in Barkaoui's (2007) study, four experienced English teachers scored EFL learners' writing scripts using holistic scale and analytic scale alternatively. The results showed that holistic rating scale contributed to higher rating reliability, while analytic rating scale caused greater variation between raters, which could be addressed only through increasing the number of writing scripts.

On the other hand, some research discovered that analytic rating scales could lead to higher rating reliability due to their advantages in making finer and more accurate distinctions between test takers' performances (Barkaoui, 2008; Li, 2014; Li, 2015; Song & Caruso, 1996; Sun & Han, 2013; Wiseman, 2008). For instance, Song and Caruso (1996), after scrutinizing holistic and analytic scores assigned by English and ESL teachers, found significant differences in the holistic scores, but not the analytic ones. They concluded that raters' teaching and rating experience might have an impact on their holistic scores, however, these factors didn't seem to affect analytic scoring as analytic scales focus raters' attention on the same aspects of the essays, thus mitigating the effects of rater-related variables. Li (2014) also compared two sets of data elicited through holistic scoring and analytic scoring of the writing scripts produced in Test for English Majors Band-4 (TEM-4) separately. He found that analytic rating scale was more advantageous than holistic scale in terms of scale discrimination, inter- and intra-rater reliability, and interaction bias between raters and examinees.

In addition to the above two lines of conclusions, some researchers argue that there's no significant differences between holistic rating scale and analytic rating scale in terms of their rating reliability (e.g., Bacha, 2001). They can both lead to higher agreement within and between raters, although analytic scales can generate more diagnostic information for EFL learners' writing proficiency.

Considering the ongoing arguments on the two types of scales and the mixed findings yielded from existing research, the present study is undertaken to re-examine the impact of two rating scales, one holistic and the other analytic, on the CET-4 writing. Specifically, the study is intended to address the following questions:

1. To what extent do the two scales differ in terms of discriminating power, rater reliability, rating variation and scale level functionality?
2. To what extent can the rating categories on the analytic scale function as intended?
3. How do raters of the CET-4 writing perceive the effectiveness of the two scales?

2.3 Research Design

To solve the research questions, a convergent parallel design of the mixed-methods approach was adopted. According to Creswell and Clark (2017), the advantage of this type of design lies in that the researcher can collect and analyze both quantitative and qualitative data during the same research phase, and the two sets of data can then be merged to provide an overall interpretation of the research questions.

2.3.1 Research Instruments

2.3.1.1 Instruments of the Rating Experiment

The first instrument adopted in the rating experiment is the existing holistic rating scale for the CET-4 writing which consists of five band levels and three rating categories: content relevance, language quality and discourse coherence. As with the operational rating of the CET-4 writing, the holistic scale is presented in Chinese to facilitate the raters' interpretation of the scale. The second instrument is the draft version of the analytic scale. At this stage, the scale is comprised of four sub-scales: *Linguistic Range and Accuracy*, *Content and Idea*, *Discourse Organization*, and *Linguistic Appropriacy*, with each substantiated by detailed level descriptions. As with the holistic scale, the analytic scale also entails five band levels.

In addition, 30 writing scripts produced in the operational CET-4 in June 2011 and June 2016 respectively were authorized to be utilized by the CET committee. These scripts which had been used as benchmark essays during the operational rating process address two different topics, with half of them entitled *Online Shopping* and the other half involving writing a *Thank-you* letter.

2.3.1.2 Instruments of the Interview

The instrument adopted in the interview was a semi-structured interview guideline which was designed based on the scale usefulness framework proposed by Knoch (2009). The guideline mainly deals with the raters' perceptions of the two scales, including scale clarity (perceived validity), scale completeness (perceived validity), scale operability (practicality), feedback for teaching (impact), correspondence with the CET-4 writing performance (authenticity), as well as its efficiency in rater training (perceived reliability).

2.3.2 Participants

21 raters selected from various CET-4 marking centers in China took part in the rating experiment. Among them, 15 were females and 6 males. 19 of them had more than 5 years of teaching experience and had been involved in the operational marking of CET-4 writing for more than 3 times. 6 of the raters held a PhD degree while the others held a master degree.

For the semi-structured interviews, 10 participants (4 males and 6 females) were invited. All of them were experienced raters of the CET-4 writing and several of them were rating experts appointed by the CET committee.

2.3.3 Data Collection

The rating experiment began in early January, 2017 and lasted for approximately one and a half month. To control the potential order effects caused by the use of the two rating scales, a counterbalanced design was adopted. Specifically, the participants were divided into two groups with 10 in the first group and 11 in the other. The rating procedures were as follows: in the first session, Group1 started to rate the 30 CET-4 writing scripts using analytic scale while Group 2 accomplished the same assignments by using the existing rating scale; Whereas, in the second session, the two groups utilized the two scales in reverse order.

Following the rating experiment, the interviews were conducted one on one through WeChat, a popular communication software in China. The raters were encouraged to comment on the analytic scale and further, to make suggestions for the potential improvement of the scale. All the interviews were live recorded by a digital voice recorder.

2.3.4 Data Analysis

2.3.4.1 Quantitative Analysis

The rating data were submitted to Many-faceted Rasch Model (MFRM) analysis using the computer program of FACETS version 3.80.0 (Linacre, 2013). MFRM can be used to calibrate a number of facets involved in the rating onto a common logit scale, thus allowing for a direct comparison of the two scales. Specifically, indices like examinee separation, rater separation, scale level functionality and rating category functionality were closely looked at. According to Fisher (1992), the examinee separation ratio is conventionally used as an effective indicator of the discrimination of the rating scale because it measures the spread of examinees' proficiency relative to their precision. Meanwhile, rater separation demonstrates how the raters as a group differ in terms of their severity or leniency. The scale level response structure and category response structure are indices indicating whether all the scale levels and scale categories can be used as intended in the rating process. Of particular interest to this study were infit and outfit mean squares (MnSq) which revealed variations in rating and assessed global model fit.

2.3.4.2 Qualitative Analysis

Recordings of the interviews were transcribed and coded. A coding scheme was developed *a priori* by taking into account the aspects involved in the interviews. The interview data were then classified according to the themes. It is worth mentioning that to ensure the accuracy and appropriateness of the coding schemes, a second qualified researcher was invited for a double check.

2.4 Results and Discussion

2.4.1 Effectiveness of the Two Scales

To facilitate comparison of the two scales, the key statistics resulting from the two MFRM analyses are summarized in Table 2.2.

Examinee Discrimination

As can be seen from Table 2.2, the examinee separation ratio for the holistic rating session is 14.88, indicating that the measures of examinee proficiency are statistically different, and the examinee separation indices by the two scales are quite similar, although the one yielded by the analytic scale is a little higher. Meanwhile, the separation reliability estimates of the two scales are also very close to each other, with the one by the analytic scale being slightly higher.

Table 2.2 Key statistics for the two rating scales

Qualities	Indices	The existing rating scale	The analytic scale
Examinee discrimination	Examinee separation ratio	14.88	14.99
	Separation reliability	.98	1.00
Rater separation	Rater separation ratio	5.18	4.11
	Separation reliability	.87	.94
Variation in ratings: 0.7~1.3	% Unexpected responses	0.8%	0.6%
	% Rater misfit	9.5%	4.8%
	% Rater overfit	28.6%	9.5%
Scale properties	Level functionality	Well-spread	Well-spread

According to Knoch (2009), a higher examinee separation ratio indicates a more discriminating rating scale since more levels on the rating scale are attended to in the rating process. Following this logic, we can conclude that the two scales are both effective in discriminating the examinee proficiency levels, and the analytic scale is slightly more powerful than the holistic scale.

Rater Separation and Reliability

As shown in Table 2.2, the rater separation ratio resulted from the analytic rating process is 4.11 which is moderately lower than the one by the holistic rating session (5.18). According to Knoch (2009), "a well-functioning rating scale would result in small differences between raters in terms of their leniency and harshness as a group" (p. 205). Eckes (2011) also echoed this point by stating that the rater separation ratio close to 1 would be ideal because all raters would form a single, homogeneous class. Following these criteria, it is fair to say that the raters as a group were more similar in terms of their severity in using the analytic scale than in using the holistic scale.

Variation in Ratings

In assessing global model fit, the first index to be examined is the standard residual. As Table 2.2 shows, only 0.8% of standard residuals caused by the holistic rating session are higher than 3, indicating a satisfactory model fit. Likewise, in the analytic rating session, the standard residuals greater than 3 merely take up 0.6%, suggesting a good model fit as well. However, as Eckes (2011) noted, "such a result does not preclude that specific parts of the measurement system exhibit deviations from model expectations" (p. 58). Therefore, a closer look at the rater statistics is necessary so as to uncover potential variation in ratings.

Considering that the primary intention of the MFRM was to stringently compare the effectiveness of the two scales in the CET-4 writing and to inform further scale revision where necessary, a more severe control range of 0.7~1.3 (Bond & Fox, 2015; McNamara, 1996; Wright & Linacre, 1994) is adopted when examining the rater fit statistics. Following this criterion, raters with fit values exceeding 1.3 are considered misfitting because they assign ratings too inconsistently and show more variation than the model would predict; Whereas, raters with fit values less than 0.7 are overfitting as their ratings exhibit less variation than expected. Knoch (2009) holds that a well-functioning rating scale would result in fewer raters who rate either inconsistently or unduly consistently. According to Knoch, an overfit can be attributed to two possible reasons: one reason is that the raters were rating too consistently, the other involves the raters' overuse of the inner levels (i.e. Levels 2, 3, 4) of a rating scale in order to play safe. As is shown by Table 2.2, in the holistic rating session, there are 9.5% raters with fit values greater than 1.3, suggesting that their ratings display too much variation than expected. Meanwhile, 28.6% raters are overfitting with fit values less than 0.7, indicating that their ratings exhibit less variation than intended.

By contrast, when using the analytic scale, the raters' fit statistics are much more satisfactory. Only one rater shows more variation than predicted, and 9.5% raters are lack of supposed variation. Taking these findings together, it can be concluded that the analytic scale can lend itself more easily than the holistic scale to the control of rating variation.

Scale Level Functionality

The first commonly used indicator of the functionality of the scale levels is the average measure of each scale level, which represents the average of the examinee measure modeled to generate the observations in a given level (Eckes, 2011). The underlying assumption is that average measures should progress monotonically with the increase of scale levels. As Table 2.3 shows, the observations of the five levels on the holistic scale range from 83 to 169, all of which greatly exceed the minimum requirements of 10. Meanwhile, the average measures of each level increase monotonically from −6.51 logits to 5.41 logits, indicating that the five levels on the holistic scale are used reasonably by the raters.

Another indicator of scale level functionality is the outfit MnSq value for each level which compares the average examinee measure with the expected examinee measure that the Rasch model would predict for each scale level. According to Bond and Fox (2015), this outfit MnSq value should not exceed 2. As is shown in Table 2.3, the outfit MnSq values of the holistic scale are all around the expected value of 1. Finally, the level thresholds progress clearly from −5.27 logits to 4.66 logits. All of the threshold calibrations stay well within the expected range of 1.4~5.0 logits (Bond & Fox, 2015). Based on these findings, it is fair to say that the levels on the existing rating scale are properly ordered and generally functioned as intended.

Table 2.3 Level statistics for the existing rating scale

Level	Observed count (%)	Average measure	Outfit MnSq	Threshold calibration
1	85 (14)	−6.51	1.0	NONE
2	140 (23)	−3.16	.9	−5.27
3	169 (28)	.56	.9	−1.39
4	132 (22)	3.01	.9	1.99
5	83 (14)	5.41	1.1	4.66

The same set of statistics for the analytic scale are listed in Table 2.4. As the table shows, the observations for each of the five levels on the analytic scale are all well above 10, suggesting that the five levels on this scale are also reasonably attended to by the raters. Besides, the average measures increase from −4.66 logits to 4.10 logits as the scale levels move up. The outfit MnSq values are either equal or very close to the expected value of 1. Finally, the level thresholds advance monotonically with the increase of the levels. Specifically, there is a clear progression from −3.93 logits to 4.27 logits. Given these findings, we can conclude that the five levels on the analytic scale can generally be used as intended.

To sum up, the results as presented above offer answers to the first research question: in general, the two scales are both effective in terms of scale discrimination, rating reliability and scale level functionality; However, the analytic scale lends itself more easily to the control of rating variation as the raters' behaviors are more consistent both at group level and at individual level.

2.4.2 Category Functionality of the Analytic Scale

According to Eckes (2011), the analysis of the category facet is able to generate insight into the relative difficulty of each rating category and to test the assumption that these categories can work together to define a single latent dimension. The statistics relating to the category utility of the analytic scale are presented in Table 2.5.

Table 2.4 Level statistics for the analytic scale

Level	Observed count (%)	Average measure	Outfit MnSq	Threshold calibration
1	361 (14%)	−4.66	1.0	NONE
2	517 (21%)	−2.46	1.1	−3.93
3	714 (28%)	−.02	1.0	−1.53
4	673 (27%)	2.34	1.0	1.20
5	254 (10%)	4.10	1.1	4.27

Table 2.5 Category measurement report

			Infit		Outfit		Correlation
Category	Measure	S.E.	MnSq	Std	MnSq	Std	
Linguistic appropriacy	.74	.07	.85	−2.6	.92	−1.1	.89
Linguistic range and accuracy	.20	.07	.84	−3.0	.86	−2.4	.90
Discourse organization	−.30	.07	1.12	2.0	1.13	2.1	.87
Content and idea	−.65	.07	1.09	1.5	1.13	2.2	.86
Mean	.00	.07	.97	−.5	1.02	.3	.88
S.D.	.61	.00	.09	1.7	.08	1.3	.01

As is shown by the table, the four categories differ significantly from each other in terms of their difficulty measures, with *Linguistic appropriacy* being the most difficult one (0.74 logits), while *Content and idea* being the easiest (−0.65 logits). This result implies that it is relatively more difficult for examinees to gain a higher score on the category of *Linguistic appropriacy* than on *Content and idea.* Meanwhile, the mean square fit indices of the four categories all stay within the narrow range of 0.7~1.3, which indicates satisfactory data-model fit. Besides, the correlation estimates of the four rating categories range from 0.86 to 0.90, all of which are well above 0.3, suggesting that the four rating categories can work together to measure one single latent construct.

In all, these results confirm the robustness of the four rating categories. That is to say, they could function ideally both as an individual and as a whole in discriminating the CET-4 writing performance. The results could offer adequate answers to the second research question.

The findings elicited thus far through quantitative analysis enabled a straightforward comparison of the effectiveness of the two scales in the CET-4 writing. However, to gain a more comprehensive and more in-depth understanding of the two scales, a closer look at the interview data should be essential.

2.4.3 Raters' Perceptions of the Two Scales

The raters' perceptions of the two scales were transcribed and then classified based on the themes entailed in the interview guide.

Overall Effectiveness

Almost all the raters expressed positive views on the overall effectiveness of the analytic scale. Rater 10 believed that compared with the existing rating scale, the analytic scale made the scoring of the CET-4 writing performance more reliable and more evidence-based. Also, the raters considered the analytic scale 'more detailed and more specific' (R1), thus 'it gives more useful guidance in rating' (R3) and 'it makes the rating well-grounded and more objective' (R2). To quote R2:

> In using the existing rating scale, I often hesitated in awarding scores for fear that some important information might be ignored. However, with the analytic scale, I feel more confident because everything is clearly set out in the scale.

There was only one rater (R7) who was slightly concerned over the effectiveness of the analytic scale, holding that:

> Although the analytic scale looks reasonably good, it may not be practical to rely solely on the scale.
>
> In the operational rating of CET-4, benchmark essays played a crucial role which sometimes even outweighed the role of the rating scale. However, it should be admitted that in the operational rating process, interpretations of the benchmark essays often varied from person to person, which to some extent would compromise the objectivity of holistic scoring.

Scale Clarity

As for the clarity of the two scales, the raters unanimously held more positive views toward the analytic scale. They thought that the level descriptions on the analytic scale were more distinct than those on the existing scale. Specifically, as R4 said:

> There are some key words at different levels of the analytic scale which can help to discriminate the CET-4 writing performance at different levels. For example, in the sub-scale of *Linguistic range and accuracy*, Level A is characterized by expressions like 'rich vocabulary', 'accurate expressions', 'diversified sentence structures' and so on, while Level B is filled with key words like 'basic vocabulary', 'circumlocutions' , 'lexical gaps', etc.

There are five raters who spoke highly of the sub-scale of *Linguistic range and accuracy* on the analytic scale, deeming it to be the most explicit one among the four sub-scales. As Rater 7 put it:

> When describing language errors, expressions like 'grave language errors' and 'occasional language errors' are frequently used by the existing scale of the CET-4 writing. The interpretations of such expressions, however, rely significantly on modifiers like 'grave' and 'occasional'. By contrast, descriptions on the sub-scale of *Linguistic range and accuracy* are more informative and more helpful, say, 'some grave grammatical errors occur in writing which distort the meaning conveyed', and also 'errors in collocation and usage occur occasionally, which after all do not hinder comprehension'.

Scale Completeness

According to the raters, the rating criteria and the information entailed in the analytic scale were more complete than those in the existing scale, and the analytic scale was sufficient for describing CET-4 writing performance. No criterion or information was reported unattended. For example, Rater 4 said that:

> The new scale has already accounted for everything in my mind, so I am not able to put forward any other rating categories.

Similarly, Rater 5 also mentioned that:

> The new scale is all-encompassing, including almost all the features characterizing the CET-4 writing scripts.

Scale Operability

Compared with the aspects discussed above, the raters' attitudes toward the operability of the analytic scale were a little mixed. Of them, five raters made quite positive comments on the operability of the scale. For example, Rater 4 made the following comments:

> The process of using the analytic scale was not as complicated as I had assumed it to be. In fact, rating analytically can be highly efficient if the raters were well trained. As is known to all, the spoken test of CET adopts an analytic scoring approach. At the beginning, the raters may rate a bit slowly, however, once they become accustomed to the rating scale, they can rate both quickly and efficiently.

Similarly, Rater 5 also reported that he felt a bit unaccustomed when using the analytic scale at first, but after rating a few scripts, he became more confident.

On the other hand, three raters expressed their concerns over the rating efficiency of the analytic scale. Rater 3 thought that scoring the writing scripts with the analytic scale would be extremely time-consuming. As Rater 7 put it:

> There are abundant descriptions on the analytic scale; it is therefore very exhausting to use it in rating the CET-4 writing scripts. It takes more time to rate a single script because each rating criterion has to be taken into account.

In addition to the above comments, there were two raters who seriously doubted the operability of the analytic scale. Their major concern was also related to the time consumed by analytic scoring, thus they considered it to be essentially impractical for such a large-scale test. As Rater 10 said:

> If every single piece of writing were to be assigned four analytic scores, the work load would become extremely heavier and more raters would be recruited, which would lead to a higher cost.

Feedback for teaching

All the raters expressed affirmative viewpoints on the feedback that the analytic scale would be able to provide, highlighting that:

> Such a scale would be very useful in that it could make the teaching of English writing more focused (Rater 6).

As Rater 2 said:

> Compared with the existing rating scale, the analytic scale is more advantageous in providing feedback on the test takers' writing performance. The feedback, I believe, will in turn promote the effectiveness of the teaching and learning of English writing.

Also, this point was further echoed in the following quote:

> It is no denying that the existing rating scale is very effective in terms of discriminating the CET-4 writing performance. However, even writing performance assigned at the same proficiency level may differ from each other in that some of them might excel at language use, whilst the others might do better in organization or content. It is in this respect that the analytic scale would excel the current holistic scale (Rater 3).

Usefulness for Rater Training

As for the usefulness of the scales for rater training, eight raters held a more positive attitude toward the analytic scale. For example, Rater 1 said:

> I think the analytic scale will be more effective in rater training since it is fairly detailed and the level division is also reasonable. At least, I feel, I myself would benefit from rater training featuring such an informative scale.

Even Rater 7 who had previously expressed some doubts over the operability of the analytic scale acknowledged that:

> In the operational scoring of the CET-4 writing, due to the vagueness of the existing rating scale, the raters' interpretations of the scoring points sometimes deviate from what has been informed during the rater training. Well, I think with the analytic scale, this problem might be solved.

Despite these positive remarks, however, two raters cast some doubts on the usefulness of the analytic scale in rater training. As Rater 8 said:

> With so many descriptors, how can the efficiency of rater training be guaranteed? I'm afraid that the rater training will be turned into a laborious process which costs more but gains less.

2.4.4 Discussion

2.4.4.1 Construct Validity of the Scales

Bachman and Palmer (1996) define construct validity as "the meaningfulness and appropriateness of the interpretations that we make on the basis of test scores" (p. 21). Knoch (2009) argues that to establish construct validity for a rating scale, we should not only understand the purpose and context of an assessment, but determine whether the rating scale can effectively help raters in arriving at scores.

According to Jin (2008), the intended purpose of CET-4 is two-fold: one is to provide an objective evaluation of a student' s overall English proficiency, and the other is to exert positive impact on the EFL teaching at the tertiary level in China. In view of the first aspect, the rating scale of the CET-4 writing is supposed to help evaluating the examinees' writing proficiency objectively. In other words, the rating scale should be powerful in discriminating the examinees' writing proficiency. In this study, the construct validity of the analytic scale could be evidenced on one hand by the results of MFRM analysis, and on the other by the raters' opinions elicited through the interviews. Specifically, the analytic scale resulted in higher examinee separation ratio and lower rater separation ratio, which indicated a more satisfactory scale discrimination. Besides, the raters' comments on the overall effectiveness of the two scales also bore out the superiority of the analytic scale in terms of its construct validity. According to the raters, the rating process facilitated by the analytic scale was more objective as the detailed level descriptions made the rating more valid. In other words, the raters didn't have to rely on their own intuitive judgments. As such, it is fair to say that the analytic scale is a little more advantageous than the holistic scale in terms of construct validity.

2.4.4.2 Reliability of the Scales

According to Bachman and Palmer (1996), reliability refers to consistency of measurement, and it is a necessary condition for construct validity. In this study, scale reliability was demonstrated through two major aspects: rater separation and rating variations. For the first aspect, two types of statistics were closely looked at –

rater separation ratio and rater separation reliability. The results showed that the analytic scale could help to yield moderately higher rater separation ratio as well as greater rater separation reliability, hence the scale reliability was more satisfactory.

The other aspect concerning scale reliability is raters' variation in rating which reveals the extent to which the raters' behaviors were consistent with the model expectation. Ideally, neither too much nor too little variation is desired because the two cases suggest that the raters either rate too inconsistently or over consistently (Bond & Fox, 2015). The results of MFRM indicated that compared with the holistic scale, the analytic scale lent itself more easily to control variations in rating. Hence, it can be concluded that the analytic scale is more reliable than the holistic scale in the CET-4 writing.

2.4.4.3 Impact of the Scales

According to Bachman and Palmer (1996), the impact of test use generally operates at two levels: a micro level, which concerns the impact on individuals, and a macro level, which refers to the impact on the educational system or society. As CET-4 is widely perceived to be a high-stakes test, the impact of the rating scale used in CET-4 writing should be considered both at the micro level and at the macro level. At the micro level, the scales are expected to provide useful and meaningful feedback to the test takers. In this regard, the results of the interviews could provide some insights. For instance, the raters mentioned that compared with the existing rating scale, more detailed information regarding the CET-4 writing performance was entailed in the analytic scale, such as the range of lexical use or the grammatical accuracy. If delivered to the test takers along with the score report, the information can supposedly help learners diagnose their strengths and weaknesses in EFL writing. Because the raters involved in the interviews were all experienced college English teachers, their opinions could to a large degree reflect the test takers' real needs.

In addition to the test takers, the raters as a group were also directly affected by the rating scale. During the interviews, the raters all mentioned that the analytic scale was more explicit in terms of level descriptions, enabling them to focus more intensively on the CET-4 writing scripts so as to ensure that the scripts could be assigned accurately at different levels of the analytic scale.

The impact of the rating scales at a macro level mainly concerns college English teaching. During the interviews, the advantage of the analytic scale in this respect had been echoed by the raters as they commented that the detailed information offered by the analytic scale would make the teaching of College English writing more focused than before.

In all, the above evidence indicated that the analytic scale could exert more positive impact both at micro level and at macro level.

2.4.4.4 Practicality of the Scales

According to Bachman and Palmer (1996), practicality denotes “the relationship between the resources that will be required in the design, development, and use of the test and the resources that will be available for these activities” (p. 36). In other words, the practicality of a test is constrained to some degree by the available resources. In a similar vein, the practicality of a rating scale is also limited by such kind of resources.

When being inquired about the implementation of the rating scales in CET-4 writing, the raters showed quite mixed attitudes: first, half of the interviewees exhibited great confidence in the practicality of the scale, commenting positively on the use of it in the CET-4 writing; Second, three of the interviewees expressed some concerns over the intensive labor involved in analytic scoring; Third, two of the interviewees seriously doubted the operability of the analytic scale in scoring the CET-4 writing performance. Admittedly, such mixed attitudes might mitigate the practicality of the analytic scale in the CET-4 writing to some extent. However, the interviewees’ mixed attitudes could be attributed to two possible factors: on one hand, the scale descriptors were established merely on an experimental basis at this stage, hence they still need further refinement; On the other, the interviewees were highly experienced in using the existing holistic scale, and it was somewhat difficult for them to tailor themselves to analytic scoring in such a short period. As such, retraining activities would be called for to help raters become more accustomed to the analytic scale, were it to be adopted in the operation scoring of the CET-4 writing. In addition, considering that large-scale EFL tests have increasingly resorted to automatic essay scoring in recent years, the potential of such an analytic rating scale in the CET-4 writing deserves further exploration because in automatic scoring process, the large rating cost traditionally associated with the use of analytic scales can be reduced significantly (Lee et al., 2010).

2.5 Conclusion and Limitations

Taking advantage of a mixed-methods approach, this study compared the effectiveness of two common types of rating scales in the context of scoring CET-4 writing performance. Results of the study indicated that the empirically developed analytic rating scale was more robust than the holistic scale currently adopted by CET-4 writing although its practicality needed more research.

Methodologically, this study not only offered some valuable insights into the validation of rating scales in the context of large-scale EFL writing assessments, but provided some new evidence for the comparison of the two commonly used types of scales. More practically, by validating an empirically developed analytic scale for CET-4 writing, the study had some implications for the clarification of the CET-4 writing construct and the interpretation of the CET-4 writing scores. As mentioned in the beginning, the interpretability of tests scores has become an increasing demand on behalf of test users. Jin and Yang (2018) also stress that "language assessment researchers in China should attach more importance to the interpretation of test scores, treating it as a crucial part in test validation" (p. 36).

However, this study is not without limitations. The most obvious one lies in that the evidence for the use of the two scales is only one-sided, that is, from CET-4 raters. CET-4 constructors' views are yet to be investigated. Therefore, it would be a gross overstatement to say that the analytic rating scale has been fully validated because "validity is an evolving property and validation is a continuing process" (Messick, 1989: 13). In other words, more evidence relating to the scale validity will be called for, particularly evidence demonstrating the applicability of the analytic scale to the operational scoring of the CET-4 writing as well as the potentiality of the scale in reporting the CET-4 writing scores.

Acknowledgements The author would like to thank the editors and anonymous reviewers for their feedback. The author would also like to thank all the raters who participated in the study. The author wishes to gratefully acknowledge the CET committee for giving permission of the CET-4 writing scripts.

Funding This book chapter is generously funded by Shandong provincial social science planning foundation, China [grant number 19CWZJ22] and a research foundation supported by Qingdao Agricultural University [grant number 663/1118704].

References

Bacha, N. (2001). Writing evaluation: what can analytic versus holistic essay scoring tell us? *System, 29*(3), 371–383.

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

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12*(2), 86–107.

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes*. Dissertation. University of Toronto.

Berger, A. (2015). *Validating analytic rating scales: A multi-method approach to scaling descriptors for assessing academic speaking*. Peter Lang.

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

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. Routledge.

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research* (3rd ed.). Sage.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement*. Peter Lang.

Fei, Q., & Zhao, Y. Q. (2008). Problems in CET-4 writing rubric and scoring method. *Foreign Language Learning Theory and Practice, 4*(45–52), 93.

Fisher, W. P. (1992). Reliability statistics. *Rasch Measurement: Transactions of the Rasch Measurement SIG, 6*(3), 238.

Hamp-Lyons, L. (1986). No new lamps for old yet, please. *TESOL Quarterly, 20*(4), 790–796.

Hamp-Lyons, L. (1991). Scoring procedures. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241–276). Ablex.

Hamp-Lyons, L. (2016). Farewell to holistic scoring. Part Two: Why build a house with only one brick? *Assessing Writing, 29*, A1–A5.

Jian, Q. M., & Lu, J. P. (2005). Inadequacy of the proposition for the writing section of CET-4. *Foreign Languages and Their Teaching, 1*, 32–33.

Jin, Y. (2008). Powerful tests, powerless test designers? Challenges facing the College English Test. *CELEA Journal, 5*, 3–11.

Jin, Y., & Yang, H. Z. (2018). Developing language tests with Chinese characteristics: Implications from three decades of the College English Test. *Foreign Language World, 2*, 29–39.

Knoch, U. (2009). *Diagnostic writing assessment: The development and validation of a rating scale*. Peter Lang.

Lee, Y. W., Gentile, C., & Kantor, R. (2010). Toward automated multi-trait scoring of essays: Investigating links among holistic, analytic, and text feature scores. *Applied Linguistics, 31*(3), 391–417.

Li, Q. H. (2014). *Constructing and validating a rating scale for TEM-4 writing*. Science Press.

Li, H. (2015). The effects of the use of holistic and analytic scales on the reliability of EFL essay scoring. *Foreign Languages and Their Teaching, 2*, 45–51.

Linacre, J. M. (2013). *Facets Rasch measurement computer program* (version 3.80.0). Winsteps.com.

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

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education and Macmillan.

O'Loughlin, K. (1994). The assessment of writing by English and ESL teachers. *Australian Review of Applied Linguistics, 17*(1), 23–44.

Shaw, S. D. (2004). Automated writing assessment: A review of four conceptual models. *Cambridge Research Notes, 17*, 13–18.

Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge University Press.

Song, B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? *Journal of Second Language Writing, 5*(2), 163–182.

Sun, H. Y., & Han, B. C. (2013). A Comparative study of analytic and holistic rating scales for English writing. *Journal of PLA University of Foreign Languages, 6*, 48–54.

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

Weir, C. J. (1990). *Communicative language testing*. Prentice Hall Regents.

White, E. M. (1984). Holisticism. *College Composition and Communication, 35*, 400–409.

White, E. M. (1985). *Teaching and assessing writing*. Jossey-Bass Inc.

Wiseman, C. (2008). *Investigating selected facets in measuring second language writing ability using holistic and analytic scoring method*. Doctoral dissertation. Columbia University.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Shaoyan Zou received her PhD degree in Shanghai Jiaotong University. She is currently Associate Professor of Qingdao Agricultural University. Her research interests mainly include language assessment and evaluation, and foreign language teaching. She has been involved in the China Standards of English project since 2014, hence particularly interested in research on the development and validation of language proficiency scales.