The Potential of Empirically Derived Rating Scales for Inexperienced Raters: A Comparative Study on Rating Scales

Keisuke KUBOTA
Graduate School, Fukushima University

Abstract
Peer assessment is currently popular as an alternative evaluation tool in the context of EFL classrooms. It is also recognized as a useful method for active and collaborative learning. There is, however, the issue of rater bias (Farrokhi, Esfandiari, & Schaefer, 2012; Matsuno, 2009), which can cause negative consequences in instances of peer assessment. One way to avoid such unfairness is to design a better measurement scale that would work equally well for inexperienced evaluators such as the participants of peer assessments and for experienced raters. This study thus aims to explore the potential of employing an assessment instrument that can help to mitigate rater bias in peer assessment. To do this, the author adopted an empirically derived, binary-choice, boundary-defined (EBB) rating scale. The investigation used the Many-Facet Rasch Measurement (MFRM) to analyze 45 sets of essay writing data scored by 5 Japanese raters (3 experienced and 2 inexperienced). The evaluation tool was employed to investigate rater severity or leniency and rater bias patterns in comparison with two types of rating scales. The results revealed that despite the utilization of the rating scales, the same patterns of rater bias occurred as were found in previous studies (Farrokhi et al., 2012; Matsuno, 2009). Moreover, the present study confirmed that the language ability of inexperienced raters could influence their rating tendency. In other words, an inexperienced rater may behave like an experienced evaluator if the person's language proficiency levels are relatively high. On the basis of the findings of the present study, this paper discusses the implications of the use of empirically developed rating scales in EFL classroom writing assessments.

Keywords: rater bias, rater experience, EFL writing, many-facet Rasch measurement

Peer assessment, which can be defined as “an arrangement in which individuals consider the amount, level, value, worth, quality, or success of the products or outcomes of learning of peers of similar status” (Topping, 1998, p. 250), has gained popularity in English as a Foreign Language (EFL) classroom contexts as an alternative assessment tool for second language writing (e.g., Azarnoosh, 2013; Fukazawa, 2010; Matsuno, 2009; Saito, 2008). In
recognition of the importance of English performance tests, peer evaluations are considered to offer several benefits. First, peer appraisal can provide students the opportunity to learn collaboratively, and to spend less time being assessed by teachers (Fukazawa, 2010). Next, the method can enhance the efficiency of receiving peer feedback, which can probably make writing assessments more practical (Azarnoosh, 2013). Further, Saito (2008) states that peer assessments allow students to focus on their own learning, and to become familiar with performance criteria. These advantages imply that the practice of peer assessment creates a positive impact on a teacher's classroom management and on the learning accomplished by students.

However, just like other assessment methods, there are some problems related to peer evaluations with regard to the subjectivity of assessment. One of them is rater bias, which describes the effect of fellow students assessing the writing performances of their peers more severely/leniently, or more inconsistently than expected because of their personal judgment (Farrokhi, Esfandiari, & Schaefer, 2012; Matsuno, 2009). Such subjectivity may influence the reliability and validity of test scores (Lumley, 2005; Schaefer, 2008) and should be avoided in the interests of fairness. The existence of rater bias in peer evaluations seems to depend on a multiple of factors: the attitudes of the evaluators, their friendships, genders, expertise, and experience in rating, and also their language proficiency. Generally speaking, experience and knowledge about language assessment play crucial parts in the process of proficiency rating (Lumley, 2005). The present study on peer rater bias is restricted to the context of the Japanese classroom and Japanese students are generally inexperienced assessors of language proficiency. Thus, assessors’ experiences need to be considered.

Weigle (2002) states that the scoring of written compositions varies because of factors such as raters and rating scales, which could cause inconsistent ratings. Among the factors affecting evaluators, some investigations have posited rater training as a possible method to improve the consistency of peer appraisals (e.g., Lumley & McNamara, 1995; Weigle, 1998). However, there appears to be scant extant research examining whether the use of a rating scale can influence consistency or severity in appraisers. Moreover, the interaction between the status of the evaluators and the rating scale and the effect of such interaction on rater bias remain unexplored by previous research initiatives.

The objective of this study is to explore the effect of rating scales on inexperienced rater behavior by independently comparing two different types of rating scale and two types of raters. To be specific, using Many-Facet Rasch Measurement (MFRM) as a preliminary investigation leading up to a peer assessment study, the present initiative investigates the extent to which types of rating scales can mitigate the influence of rater bias in the case of inexperienced raters and can be helpful for such raters.

**Literature Review**

**Previous Rater Bias Studies**

Bias studies search for sub-patterns of rater behavior, and unexpected interactions
among factors that affect the assessment of test takers’ performance, such as a human rater, a test taker, or a rating scale. Some researchers have conducted bias studies adopting different types of raters (experienced rater or teacher rater and inexperienced rater or peer rater). For example, Schaefer (2008) examined bias patterns of 40 teacher raters, who were Assistant Language Teachers working in Japan. Each rater rated 40 essays written by Japanese university students, using the rating scale Schaefer developed based on previous studies (e.g., Jacobs, Zinkgraf, Wormuth, Hartfiel, & Hughey, 1981). He reported that the inexperienced raters tended to score the essays more leniently than the expectation of MFRM if the writers’ logits (which represent the writers’ ability) were relatively low, or more harshly if relatively high. This result is in line with the study by Farrokhi et al. (2012). In addition, Lumley (2005) conducted a study investigating significant differences between four trained raters in terms of rating criteria in Australia and found that Grammar was the most severely scored criterion. McNamara (1996) also showed a similar finding to Lumley’s study.

In some bias studies focusing on peer assessment, some researchers found the same bias tendency as that of experienced raters or teacher raters, while others found a unique one with peer assessment. As with the results reviewed above, peer raters as well as teacher raters have a tendency to rate essays more leniently or harshly (Farrokhi et al., 2012). Moreover, Matsuno (2009) found that among 16 criteria (e.g., Relevance, Spelling, Punctuation), Grammar was severely scored by teacher raters similar to those in Lumley’s (2005) study, was also the most severely scored by peer raters. She also revealed that Spelling was leniently scored by teacher raters and peer raters. These consistent results have contributed to establishing a bias pattern in EFL writing environments (Lumley, 2005; Schaefer, 2008). On the other hand, it is also noteworthy that peer raters have an inclination to score their partners’ writing more leniently than teacher raters (Farrokhi et al., 2012; Lee, 2005; Matsuno, 2009). This rating pattern in peer assessment also has been identified in a speaking study (Fukazawa, 2010). Such rater bias could negatively affect the reliability of test scores and result in inconsistent ratings. As Knoch (2007) mentions: “one reason for variability found in writing performance might lie in the way rating scales are designed” (p. 1). In order to address the issue of rater bias in peer assessment, this paper explores types of rating scale to measure inexperienced raters’ bias and develops another type of rating scale.

Two Approaches to Designing Rating Scales

There are mainly two approaches for developing a rating scale: the measurement-driven method, and the performance-based method (Fulcher, Davidson, & Kemp, 2011). The former is an a priori method, based on which experts theoretically develop rating scale descriptors. This type of rating scale has been criticized for the vagueness of the descriptors, which can produce scores with low reliability and validity, especially when raters are not trained (Turner & Upshur, 2002; Upshur & Turner, 1995). Turner and Upshur (2002) summarize these criticisms: (a) “the criteria are often irrelevant to the test task and its context” (p. 51), and (b) “the criteria are improperly grouped at descriptor levels” (p. 51). A prevalent
example of such a rating scale is Jacobs et al.’s (1981) ESL Composition Profile, which is the “best known and most widely used” (Weigle, 2002, p. 115), and has been cited in many previous studies (e.g., Matsuno, 2009).

The latter method is performance-driven and requires collecting performance data from test takers and distinguishing the main features of their performances by discussion or discourse analysis. This method has been described as “one response to criticisms of theory-based rating scales” (Turner & Upshur, 2002, p. 51), and has been receiving more attention because of its high reliability and practicality (Hirai & Koizumi, 2013; Turner & Upshur, 1996, 2002). One prominent example of a data-driven rating scale is an Empirically derived, Binary-choice, Boundary-definition (EBB) scale, originally proposed by Upshur and Turner (1995). Although it is difficult to generalize about various test tasks owing to their specific aspects, EBB has some advantages in performance assessment: (a) relatively higher reliability and validity (Hirai & Koizumi, 2013), (b) ease of use (Chan, Inoue, & Taylor, 2015; Turner & Upshur, 1996, 2002), (c) relatively light burden on raters’ memories during rating (Fulcher et al., 2011), and (d) giving more beneficial feedback to students (Knoch, 2007, 2009).

An EBB rating scale consists of a hierarchical set of binary decisions (see Appendix B) and is developed by a rank ordering of descriptors that are determined according to the key features of actual language performance data to distinguish the performance into binary levels. As reviewed above, this unique feature of the EBB can improve the ease of scoring essays and contribute to enhancing test usefulness. However, it has not been confirmed whether this function of EBB works or not in peer assessment. If inexperienced raters are able to receive the benefits of EBB, they can then be involved more efficiently in peer assessment activities. Thus, the practicality of an EBB rating scale for peer assessment needs validating.

**Functions of MFRM**

MFRM is an expansion of the basic Rasch model which enables researchers to add more than two facets, and to produce detailed information, such as rater behavior, rater severity or leniency on the same measurement scale, etc. (e.g., Eckes, 2015; Lumley, 2005; Lumley & McNamara, 1995; Matsuno, 2009; Schaefer, 2008). The traditional true-score approaches (e.g., a *t*-test, or an analysis of variance) can certainly examine issues of rater bias; however, they do not sufficiently address issues of rater bias such as severity/leniency, criterion difficulty/easiness (Eckes, 2015). Moreover, there have been a number of studies of peer assessment using MFRM in EFL writing research (e.g., Farrokhi et al., 2012; Matsuno, 2009). Hence, the present study adopts MFRM analysis in order to investigate inexperienced rater behavior from focusing on rater bias.

**The Present Study**

The present study investigates whether two different types of rating scale (a theory-based one and a data-driven one) (1) affect the type of rater behavior which has been
reported in previous studies, (2) mitigate rater bias, and (3) contribute to enhancing ease of using a rating scale for inexperienced raters. The following are the research questions (RQs) established for the present study.

RQ1. Is the rater bias found in previous studies confirmed in this study?

RQ2. Does the EBB rating scale function to mitigate rater bias?

RQ3. How do inexperienced raters feel about the features of the EBB rating scale compared with other rating scales in terms of practicality?

Methods

Test Participants and Raters

Writing products were obtained from 54 Japanese national university students (three seniors majoring in English education and literature, three juniors majoring in English education, 39 sophomores and nine freshmen majoring in economics or politics). A total of six students (two males and four females) voluntarily participated in this experiment, and 48 others (22 males and 26 females) were involved in it as part of compulsory English subjects.

In the assessment procedure, five Japanese raters were introduced. Among them, three male raters (one associate professor, one graduate English education major, and the author) assessed the students’ writing as experienced assessors, with two female raters (a senior majoring in local politics and a freshman majoring in liberal arts) as inexperienced assessors. The two inexperienced assessors were contrastive in their English proficiency. One of the two scored 705 on the TOEIC Listening & Reading (TOEIC L & R) Test in March 2016, while the other scored 360 on the same test in September of that year. Neither had had any opportunity to assess essays written in English.

Test Materials and Test Procedure

The essay topic was “If you could change one important thing about your hometown, what would you change?”, which was developed with reference to essay questions on other tests (e.g., TOEFL Test of Written English, Eiken).

In the data collection procedure, two tasks including the one above were introduced. The students were given 90 minutes (including 10 minutes for the test introduction) to write the two essays. Totally, 54 essays composed of approximately 100 words were collected from personal requests and two compulsory English classes.

Two Rating Scales and the Rating Procedure

Two analytic rating scales were adopted. One of them was the ESL Composition Profile (Jacobs et al., 1981) (hereafter, Composition), whose scoring dimensions were Content, Organization, Vocabulary, Language Use, and Mechanics (see Appendix A). The author translated it into Japanese and revised the range of the original rating scale to four points (1 to 4), because it has no midpoint, which lessened vagueness in interpreting scores.

The other rating scale, the EBB rating, was an original one developed empirically by
the author and an associate professor who took part in the study as a teacher rater (see Appendix B and C). We developed the new rating scale according to Turner and Upshur (1996) and Upshur and Turner (1995). At first, the author selected 12 samples of each scale band from all the essays, and then with the associate professor individually divided the set of 12 performances into six better and six poorer ones from our own impressions. Next, we discussed their rankings and reconciled the differences, then formulated the simplest question that allowed us to classify the samples into upper half (scores of 3 or 4) and lower half (scores of 1 or 2). This process was repeated in each upper and lower set to complete four-point EBB scales. These steps were run through again in each of the five criteria for Composition.

In the scoring procedure, the five raters individually assessed students’ writing using the two types of analytic rating scale. The anonymity of both the raters and the participants was preserved. In order to avoid some influences on scoring such as the practice effect of using the same rating scales, scoring order was counterbalanced among the raters. Two raters (one teacher and one student) were required to score the products from Composition to EBB in order for each criterion (Content, Organization, Vocabulary, Language Use, and Mechanics). On the other hand, the other two raters, except the author, were required to do this in reverse order. The author evaluated the writing in a nonconsecutive manner, in which the two rating scales were used in turns.

On the assumption that the rating scales could be used in classroom contexts, rater training should be conducted for fair assessment. However, rater training was not done in the present study, because the study purely focused on functions of the two rating scales. Instead, I arranged scoring samples indicating the scale bands. In the samples, nine essays were adopted; as a result, a total of 45 essays were assessed and analyzed.

After the rating procedure, interviews were conducted with the inexperienced raters to investigate how they used the EBB scales, and how they felt about the EBB scale in terms of practicality. The inexperienced raters were asked three questions: (1) how they felt about each rating scale, focusing on good points and bad points; (2) how they felt about the ease of use of the two rating scales by comparing them; and (3) how they felt about the two rating scales from other viewpoints. The results were recorded and investigated.

Data Analysis

MFRM was conducted using FACETS computer program, version 3.71.4 (Linacre, 2014), to examine the participants’ ability, rater severity, severity of the rating scales, and bias patterns. The scoring data was analyzed using the Rating Scale Model of MFRM, since the two rating scales used in this study had the same score levels and criteria and each criterion of the rating scales was expected to function similarly. In the analyses, participants, raters, scale types, and criteria were specified as facets.

This study aims to explore functions of rating scales. Thus, following Bond and Fox (2015), infit and outfit mean-squares were set from 0.6 to 1.4 to assess the model fit to the
Rasch model.

**Results and Discussion**

**FACETS Map and Measurement Reports**

The vertical ruler in Figure 1 shows the relationships among ability, rater severity, scale type, and criteria, and the first column shows the Rasch logit scale. Unlike raw test scores, which do not mean the same distance between scores in the rating scale, the logit scale is on the same measurement scale (Eckes, 2015).

The second column in Figure 1 shows participants. Each asterisk represents one test taker. The logits varied from -1.70 to 4.16, and participants are distributed on the map according to their abilities.

The third column in Figure 1 shows rater severity. The most severe rater is at the upper part, and the most lenient rater is at the lower part. *Ex1* indicates the author, *Ex2* the associate professor, and *Ex3* the graduate student; *Inex1* the relatively novice rater, and *Inex2* the relatively proficient rater. Rater severity differed from -1.36 to 0.64; *Ex1* (0.64 logits) is the most severe rater, followed by *Ex2* (0.60 logits), *Ex3* (0.18 logits), *Inex2* (-0.06 logits); *Inex1* (-1.36 logits) was the most lenient. The fourth column in Figure 1 indicates the severity of each scale type. It shows that Composition is relatively more severe than EBB; the severity varied from -0.29 to 0.29.

The fifth column in Figure 1 illustrates the severity of each of the five criteria of the rubrics. Criterion 1 is *Content*, Criterion 2 *Organization*, Criterion 3 *Vocabulary*, Criterion 4 *Language Use*, and Criterion 5 *Mechanics*. Their severity differs from -0.54 to 0.97; Criterion 4 is the most severe category (0.97 logits), Criterion 3 is relatively severe (0.29 logits), Criteria 5 and 2 are relatively lenient (respectively -0.28 logits, -0.44 logits), and Criterion 1 is the most lenient (-0.54 logits).

The sixth column in Figure 1 shows the four-point rating scale measure, and the distance between each threshold on the scale. For instance, a student at 1.00 logits has a 50% possibility of getting three points in each criterion from a rater at 1.00 logits.

As mentioned above, to assess the model fit, the infit and outfit mean-squares between 0.6 and 1.4 were regarded as reasonable (Bond & Fox, 2015; participants below the overfit range are too consistent; participants above the underfit range are too inconsistent). Following this criterion, ten participants misfitted to the model. The infit mean square value of all participants was 2.03 (Participant 23), 1.76 (P 11), 1.62 (P 14), 1.55 (P 20), 1.48 (P 7 and P 36), 1.41 (P 34), 0.57 (P 3), 0.47 (P 24), and 0.45 (P 18). Among them, 6 underfit participants (P 7, P 11, P 14, P 20, P 23, and P 36) were eliminated from the data set, except for P 34 (slightly higher than the range at 1.40), because “misfit can greatly change the substantive meaning of the resulting measure and thus threaten the validity of the interpretations and uses that draw on these measures” (Eckes, 2015, p. 77). All the raters, scale types, and criteria were within the range: 0.79 to 1.39 for raters, 0.87 to 1.14 for scale type, and 0.83 to 1.13 for criteria.
Figure 1. Variable map of MFRM analysis. Each asterisk (*) indicates one participant. Ex indicates an experienced rater. Inex indicates an inexperienced rater. In Criteria column, 1 indicates Content, 2 Organization, 3 Vocabulary, 4 Language Use, and 5 Mechanics.

Table 2 shows the percentages of mean-square fit statistics after eliminating the data of underfit participants. Underfit test participants were 5.13% of the total, which showed a
certain pattern; specifically, that they could not complete the test task appropriately. In this study, the percentage of underfit participants was not high, and it seems to be appropriate.

### Table 1
**Descriptive Statistics of Four Facets**

<table>
<thead>
<tr>
<th>Facets</th>
<th>M (SD)</th>
<th>Min to Max</th>
<th>Range</th>
<th>Separation</th>
<th>Strata</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants</td>
<td>1.05 (1.31)</td>
<td>-1.31 to 4.16</td>
<td>5.47</td>
<td>5.43</td>
<td>7.57</td>
<td>.97</td>
</tr>
<tr>
<td>Raters</td>
<td>0.00 (0.73)</td>
<td>-1.36 to 0.64</td>
<td>2.00</td>
<td>8.72</td>
<td>11.95</td>
<td>.99</td>
</tr>
<tr>
<td>Scale type</td>
<td>0.00 (0.29)</td>
<td>-0.29 to 0.29</td>
<td>0.58</td>
<td>5.53</td>
<td>7.71</td>
<td>.97</td>
</tr>
<tr>
<td>Criteria</td>
<td>0.00 (0.56)</td>
<td>-0.54 to 0.97</td>
<td>1.51</td>
<td>6.68</td>
<td>9.24</td>
<td>.98</td>
</tr>
</tbody>
</table>

### Table 2
**Percentages of Mean-Square Fit Statistics**

<table>
<thead>
<tr>
<th>Facets</th>
<th>Overfit (value &lt; 0.60)</th>
<th>Fit (0.60 ≤ Value ≤ 1.4)</th>
<th>Underfit (1.4 &lt; Value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants</td>
<td>Infit 5.13</td>
<td>89.74</td>
<td>5.13</td>
</tr>
<tr>
<td></td>
<td>Outfit 5.13</td>
<td>87.18</td>
<td>7.69</td>
</tr>
<tr>
<td>Raters</td>
<td>Infit 0.00</td>
<td>100</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Outfit 0.00</td>
<td>100</td>
<td>0.00</td>
</tr>
<tr>
<td>Scale Type</td>
<td>Infit 0.00</td>
<td>100</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Outfit 0.00</td>
<td>100</td>
<td>0.00</td>
</tr>
<tr>
<td>Criteria</td>
<td>Infit 0.00</td>
<td>100</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Outfit 0.00</td>
<td>100</td>
<td>0.00</td>
</tr>
</tbody>
</table>

### Rater Bias Confirmed in This Study (RQ1)

RQ1 investigates whether the rater bias patterns of the inexperienced raters accord with the results of previous studies. According to Figure 1, the two inexperienced raters (each severity measure = -1.36, -0.06) tend to assess the essays more leniently than the three experienced raters (each severity measure = 0.64, 0.60, 0.18), which is in line with the results of the following previous studies (Farrokhi et al., 2012; Lee, 2005; Matsuno, 2009). In particular, Inex1, who was a relatively more novice rater, rated the essays the most leniently, which may imply that rater experience could influence the rated scores.

Furthermore, among the category of Criteria in Figure 1, criterion 4 (Language Use) was at the top, meaning Language Use (difficulty measure = 0.97), which can be considered the same criterion as Grammar, was rated the most severely among the five criteria. This result shows a similar tendency with the studies of Lumley (2005), McNamara (1996) and Matsuno (2009). McNamara pointed out that trained raters in his study unconsciously put emphasis on grammar, while Matsuno considered this result an influence of the Japanese English education style that focuses on grammar. The result in the present study seems to be
influenced by the same factor as in Matsuno because of the similar education context and needs to be investigated further.

This study also finds other rater behavior patterns as some previous studies. For example, Mechanics (difficulty measure = -0.28), which can be considered as a part of Spelling, was scored relatively leniently like Matsuno’s (2009) study. Confirming the same rating behavior as previous studies can contribute to establishing systematic sub-patterns of peer assessment even if a different type of rating scale is adopted for writing assessment, and this has implications for rater training.

Table 3
Summary of Significant Rater Bias

<table>
<thead>
<tr>
<th>Raters</th>
<th>Participants</th>
<th>Scale type</th>
<th>Criteria</th>
<th>Obs - Exp average</th>
<th>Bias size</th>
<th>t-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inex1</td>
<td>P1</td>
<td>Composition</td>
<td>Content</td>
<td>-0.88</td>
<td>-3.27</td>
<td>-2.01</td>
</tr>
<tr>
<td>Inex1</td>
<td>P1</td>
<td>Composition</td>
<td>Mechanics</td>
<td>-0.91</td>
<td>-3.38</td>
<td>-2.23</td>
</tr>
<tr>
<td>Inex1</td>
<td>P15</td>
<td>EBB</td>
<td>Content</td>
<td>-1.86</td>
<td>-4.34</td>
<td>-2.22</td>
</tr>
<tr>
<td>Inex1</td>
<td>P16</td>
<td>EBB</td>
<td>Mechanics</td>
<td>-2.08</td>
<td>-3.68</td>
<td>-2.07</td>
</tr>
<tr>
<td>Inex1</td>
<td>P17</td>
<td>Composition</td>
<td>Content</td>
<td>-1.43</td>
<td>-3.93</td>
<td>-2.24</td>
</tr>
<tr>
<td>Inex1</td>
<td>P2</td>
<td>Composition</td>
<td>Content</td>
<td>-1.59</td>
<td>-4.46</td>
<td>-2.57</td>
</tr>
<tr>
<td>Inex1</td>
<td>P3</td>
<td>Composition</td>
<td>Mechanics</td>
<td>-0.89</td>
<td>-3.16</td>
<td>-2.08</td>
</tr>
<tr>
<td>Inex1</td>
<td>P34</td>
<td>EBB</td>
<td>Content</td>
<td>-2.36</td>
<td>-5.53</td>
<td>-2.91</td>
</tr>
<tr>
<td>Inex1</td>
<td>P38</td>
<td>Composition</td>
<td>Vocabulary</td>
<td>-0.88</td>
<td>-3.20</td>
<td>-2.01</td>
</tr>
<tr>
<td>Inex1</td>
<td>P4</td>
<td>Composition</td>
<td>Content</td>
<td>-1.46</td>
<td>-4.03</td>
<td>-2.29</td>
</tr>
<tr>
<td>Inex1</td>
<td>P4</td>
<td>EBB</td>
<td>Organization</td>
<td>-1.67</td>
<td>-4.37</td>
<td>-2.45</td>
</tr>
<tr>
<td>Inex1</td>
<td>P42</td>
<td>EBB</td>
<td>Content</td>
<td>-0.94</td>
<td>-3.93</td>
<td>-2.42</td>
</tr>
<tr>
<td>Inex2</td>
<td>P1</td>
<td>Composition</td>
<td>Organization</td>
<td>-1.70</td>
<td>-4.47</td>
<td>-2.51</td>
</tr>
<tr>
<td>Inex2</td>
<td>P1</td>
<td>EBB</td>
<td>Vocabulary</td>
<td>-1.51</td>
<td>-4.55</td>
<td>-2.11</td>
</tr>
<tr>
<td>Inex2</td>
<td>P13</td>
<td>EBB</td>
<td>Organization</td>
<td>-1.78</td>
<td>-5.53</td>
<td>-2.72</td>
</tr>
<tr>
<td>Inex2</td>
<td>P19</td>
<td>Composition</td>
<td>Mechanics</td>
<td>-1.61</td>
<td>-3.77</td>
<td>-2.59</td>
</tr>
<tr>
<td>Inex2</td>
<td>P38</td>
<td>EBB</td>
<td>Organization</td>
<td>-0.93</td>
<td>-3.60</td>
<td>-2.40</td>
</tr>
<tr>
<td>Inex2</td>
<td>P39</td>
<td>EBB</td>
<td>Language Use</td>
<td>-2.31</td>
<td>-3.63</td>
<td>-2.08</td>
</tr>
<tr>
<td>Inex2</td>
<td>P40</td>
<td>EBB</td>
<td>Mechanics</td>
<td>-2.36</td>
<td>-4.56</td>
<td>-2.45</td>
</tr>
<tr>
<td>Inex2</td>
<td>P44</td>
<td>EBB</td>
<td>Mechanics</td>
<td>-1.36</td>
<td>-2.99</td>
<td>-2.06</td>
</tr>
</tbody>
</table>

Note: P indicates participants; Obs indicates observed scores in the original data; Exp indicates expected scores from MFRM.

The Function of the EBB Rating Scale to Mitigate Rater Bias in Inexperienced Raters (RQ2)

RQ2 uses bias analysis to address the influence of differences between rating scales on rater bias. More specifically, this RQ examines whether the EBB rating scale works well
mitigating rater bias because of its features. Following Eckes (2015, p. 314), \( t \) statistics greater than or equal to |±2.00| were considered a significant bias. Table 3 shows a summary of significant bias among four facets, with 2.22% (20/900) of the significant biases being confirmed. All the significant bias was severe, that is, observed scores (Obs) were unexpectedly lower than expected scores (Exp). This result might be because the raters were inexperienced, implying the lack of rating experience could lead to lenient ratings. Although only two raters are not enough to build systematic patterns of inexperienced rater behavior, the result seems to be worth investigating further.

Table 4 and Table 5 show the totals for significant bias for the two inexperienced raters. EBB revealed 11 cases of significant bias, and Composition revealed nine for both raters, which means there was no systematic pattern to mitigate rater bias. This result suggests that the EBB developed for this study may have less ability to mitigate rater bias. However, this study’s rating scale was a data-driven type, which seems to need continuous development based on diagnosis by MFRM. Thus, mitigating rater bias should be examined further, after additional development of the EBB rating scale.

Table 4

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Content</th>
<th>Organization</th>
<th>Vocabulary</th>
<th>Language use</th>
<th>Mechanics</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBB</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Composition</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

Table 5

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Content</th>
<th>Organization</th>
<th>Vocabulary</th>
<th>Language use</th>
<th>Mechanics</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBB</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Composition</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

Inexperienced Raters’ Perception of the EBB Rating Scale Compared With the Composition Scale in Terms of Practicality (RQ3)

RQ3 examines how inexperienced raters who have no experience in rating English essays feel about the two rating scales used in this study (Composition and EBB). For this, we conducted interviews following the rating procedure. The first question was how they felt about each rating scale, focusing on good and bad points. They mentioned ease of use as the chief merit of the EBB rating scale. For example, Inex2’s comment was:

*The yes/no rating style of the EBB was new and easy for scoring essays, and the...*
description of the EBB was good, because it was more concrete than the other rating scale.

This comment indicates that inexperienced raters can appreciate the benefit of the EBB’s hierarchical set of binary decisions, which lightens the memory burden on the raters (Fulcher et al., 2011).

The second question was how they felt about the two rating scales’ comparative ease of use. In this question, Inex1 made an interesting comment about Composition:

Composition had many descriptions about quality of the essays at each level, and I spent much time rating them. Both rating scales were hard for scoring the essays, but Composition was harder.

This comment reveals how the EBB can be relatively time-saving because of the features reviewed in the previous studies, such as ease of use (Chan et al., 2015; Turner & Upshur, 1996, 2002).

The third question was how they felt about the two rating scales from other viewpoints. Here, Inex2 commented on a controversial point of the EBB rating scale:

The first decision seems to be important, because it chooses 1 or 2 points and 3 or 4 points. I think the maximum point should be decided firstly. At least for me, it is favorable.

This comment points out one aspect of this EBB which should be studied carefully. In this EBB rating scale, which we developed following Turner and Upshur’s (1996) procedure, the first decision point divides the upper level (3 or 4 points) or the lower level (1 or 2 points). However, the order of EBB’s hierarchical descriptors can vary depending on the developers’ aim. In fact, Hirai and Koizumi’s (2013) EBB rating scales firstly required raters to decide whether the performance assessed matched the maximum score or minimum score; that is, the first decision is made at a point of either 5 or lower or at a point of 1 or higher. As a result, their EBBs require more decisions to reach the score. Further study is needed to examine how this difference can influence rating results.

On the whole, inexperienced raters can benefit from the EBB’s features, and recognize its rating system as a useful tool for assessing essays, even if raters have no previous experience in an English writing assessment activity.

Conclusion

This study investigated (1) rater behavior reported in previous studies, (2) mitigation of rater bias, and (3) ease of use with a rating scale for inexperienced raters. Firstly, although the study introduced a relatively new type of rating scale, the EBB, into rater bias study, it found
similar patterns of rater bias as in the previous studies (Farrokhi et al., 2012; Matsuno, 2009), which can contribute to enhancing the efficacy of rater training. Moreover, the language ability of inexperienced raters could influence their severity or leniency, which may imply a new niche for validating peer assessment of writing. Secondly, the EBB rating scale did not function well in mitigating rater bias in inexperienced raters employed in this study, although the EBB seems to have some potential to explore such an aspect as ease of the scale. In order to fully utilize the advantages of an EBB rating scale, more development is needed, based on raters’ feedback, as well as diagnosis by MFRM. Thirdly, inexperienced raters who were not trained for language assessment reported that the EBB was easier to use because of its hierarchical set of descriptors, and they could efficiently use the EBB rating scale, which may indicate that untrained raters can benefit from the features of an EBB rating scale. This might facilitate introducing peer assessment in EFL classrooms with an EBB.

There were five limitations in the present study. The first is the small number of raters. To identify rater patterns in peer assessment more clearly and establish their general tendencies, many more inexperienced raters need to be introduced into the rating procedure. The second limitation is the size of writing performance samples. More various writing performances are desirable, because a data-driven type of rating scale is a rating scale that utilizes prominent features of written performances. The more various the performances are, the more discriminating the rating scale could be. Third, additional development of the EBB rating scale, and further analysis of it, are necessary. In the rating scales used in this study, there are a few questions to be dealt with, including the rater strata, or score distribution (which is not discussed in this paper), which should be tackled in a future study. Fourth, the language proficiency of the inexperienced raters was not uniform. Their language ability as measured by the TOEIC L & R, which was the English proficiency test used in this study, was restricted. Thus, the results may have limitations in terms of the learners’ language proficiency. Fifth, the lab study had no actual classrooms, which means this study did not consider the existence of other bias factors such as friendship or fear of grades. Further analysis will be required in existing classrooms.

Considering the limitations above, future studies should investigate other advantages of an EBB rating scale that were not discussed here. One of them would be giving fruitful feedback to learners, as suggested by Knoch (2007, 2009). Knoch further states that the features of a data-driven rating scale seem to be “more valid and useful for diagnostic writing assessment purposes” (Knoch, 2009, p. 300). Thus, for positive reactions to tests, an EBB rating scale may assume a key role. Furthermore, developing an EBB rating scale may contribute to enhancing rater reliability by training, because the discussion of test takers’ performances by rating scale developers is essential for the development procedure. Discussing test takers’ characteristics and designing an appropriate rating scale will enable raters to interpret both the test takers and their performances, and to share their perceptions.

In sum, although this study cannot confirm a method to mitigate rater bias in an EBB rating scale, the EBB has considerable potential for peer assessment, including ease of use,
meaningful feedback, and being a tool for rater training. After developing an EBB based on the results of MFRM and comments by raters, we expect the EBB’s efficacy to be validated in future studies.

Acknowledgement

I express my deep appreciation to Dr. Shuichi TAKAKI for his helpful advice and warm support. I also appreciate two anonymous reviewers for their thorough reviews and many beneficial comments on this article.

References


Appendix A: ESL Composition Profile Used in This Study (Japanese)

<table>
<thead>
<tr>
<th>評価項目</th>
<th>スコア</th>
<th>評価</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content (内容)</td>
<td>4</td>
<td>Very good to Excellent</td>
</tr>
<tr>
<td>3</td>
<td>Average to Good</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Poor to Fair</td>
<td></td>
</tr>
<tr>
<td>Organization (論理・構成)</td>
<td>4</td>
<td>Very good to Excellent</td>
</tr>
<tr>
<td>3</td>
<td>Average to Good</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Poor to Fair</td>
<td></td>
</tr>
<tr>
<td>Vocabulary (辞彙)</td>
<td>4</td>
<td>Very good to Excellent</td>
</tr>
<tr>
<td>3</td>
<td>Average to Good</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Poor to Fair</td>
<td></td>
</tr>
<tr>
<td>Language Use (文法・構文)</td>
<td>4</td>
<td>Very good to Excellent</td>
</tr>
<tr>
<td>3</td>
<td>Average to Good</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Poor to Fair</td>
<td></td>
</tr>
<tr>
<td>Mechanics (形態のいかが)</td>
<td>4</td>
<td>Very good to Excellent</td>
</tr>
<tr>
<td>3</td>
<td>Average to Good</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Poor to Fair</td>
<td></td>
</tr>
</tbody>
</table>

Appendix B: The EBB Rating Scale Developed in This Study (Japanese)

1. Content（内容）

   現状とその改善すべき点について述べられている
   
<table>
<thead>
<tr>
<th>現状について述べられている</th>
<th>改善点の重要性が述べられている</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

2. Organization（論理・構成）

   大局的に一貫性が見られる
   
   文と文の間で局的に一貫性が見られる
   
<table>
<thead>
<tr>
<th>文と文の間で局的に一貫性が見られる</th>
<th>ディスコースマーカーを用いながら、順序立てて文章が組み立てられている</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
3. Vocabulary（語彙）

同様の単純な繰り返しが少なく、かつ、必要に応じた言い換えがなされている

No

語彙の間違いが多いが理解可能である

Yes

文脈に合った具体的な語彙を使い分けられている

4. Language Use（言語使用—文法・構文）

文構造が正しく、かつ、単文以外の文も使用している

No

動詞を正しく使用している

Yes

文構造が明確で、読みやすさがある

5. Mechanics（形式）

スペルミスがない、または、ほとんどない

No

文頭や固有名詞など、大文字で書くべきところが、大文字になっている

Yes

カンマが適切に用いられており、文構造が理解しやすい
Appendix C: The EBB Rating Scale Developed in This Study (English)

1. Content

Does the writer describe the situation provided in the prompt, and claim her/his idea to improve the situation?

- No
- Yes

   Does the writer describe the situation provided in the prompt?
   - No
   - Yes

   1

   Does the writer claim importance of the idea?
   - No
   - Yes

   2

2. Organization

Is the essay consistent as a whole?

- No
- Yes

   Is an essay partly consistent?
   - No
   - Yes

   1

   Is an essay organized logically with discourse markers?
   - No
   - Yes

   2

3. Vocabulary

Does the writer use alternative words to avoid repetition, and paraphrase efficiently?

- No
- Yes

   Are a number of vocabulary errors understandable?
   - No
   - Yes

   1

   Does the writer use appropriate vocabulary in sentences?
   - No
   - Yes

   2

   3

   4
4. Language Use

Does the writer use sentence structure correctly and a complex sentence besides a simple sentence?

No

Does the writer use verbs correctly?

No

1

Yes

2

Yes

3

Is a sentence structure clear and readable?

No

4

Yes

5. Mechanics

Are there few spelling mistakes?

No

Does the writer follow capitalization rules?

No

1

Yes

2

Does the writer use commas correctly?

No

3

Yes

4