Effects of Rater Training on Raters’ Severity, Consistency, and Biased Interactions in a Task-Based Writing Assessment

Yoshihito SUGITA
Yamanashi Prefectural University

Abstract

This comparative study examines the effects of a rater-training session in relation to the degree of differences in raters’ severity, consistency, and biased interactions between raters who received training before rating (trained raters) and raters without a training session beforehand (untrained raters). Ten raters (five trained and five untrained) rated a total of 40 scripts (20 scripts for each of accuracy and communicability tasks); the analyses were done using FACETS. In addition, questionnaire responses were analyzed to investigate raters’ views regarding the effectiveness of the training. The results showed that 1) the raters were not equally severe; untrained raters were more severe than trained raters, 2) both trained and untrained raters behaved consistently in scoring; untrained raters were assumed to be less consistent, 3) a clear distinction cannot be made among trained and untrained raters as a group in terms of rater-task interactions, 4) untrained raters might be more biased in rater-subject interactions than trained raters; this is shown by the untrained raters’ inconsistency in scoring. These results imply that rater training more effectively improves raters’ scoring consistency than scoring severity.

1. Introduction

In writing performance assessment (WPA), test takers are asked to write compositions on one or more topics. Then, raters score the compositions. Since such scoring requires subjective evaluations of writing quality, decreasing raters’ scoring variability and increasing their reliability are essential practices. Previous studies on rater variation have illuminated the importance and effect of rater training. These studies offer general agreement on the following characteristics and potential benefits of rater training:

1. Since the phenomenon of rater variation is an inevitable part of the rating process, raters cannot be trained to achieve similar levels of severity (Linacre, 1989; Lumley & McNamara, 1995).
2. Rater training reduces extreme scores in terms of severity and increases the self-consistency of individual raters (Weigle, 1994; Wigglesworth, 1993).
3. Rater training can increase overall consistency through increasing intra-rater consistency (McNamara, 1996). Furthermore, understanding of the definition of the ability measured may be a central aspect of the training process (Weigle, 1994).
These studies' findings imply that training should not function to force raters into agreement with each other (inter-rater reliability), but rather to train raters to be self-consistent (intra-rater reliability). To increase rater self-consistency, rater training should help them understand the assessment criteria for the ability being measured and, in the case of WPA, the characteristics of writing samples at different levels. Based on this view, a rater-training program for a task-based writing performance test (TBWT) consisted of the following stages: First, raters were introduced to the TBWT assessment instruments to promote understanding of the constructs and assessment tasks (1st stage). Second, raters were required to rate a number of writing samples according to the assessment criteria, and were then familiarized with the scoring procedure through rating practice (2nd stage). Third, raters increased experience in rating and thus arrived at a common interpretation of the assessment instruments (3rd stage).

A comparative study was conducted to examine the effects of the programmed training session. There were two steps in analyzing the study. First, to elicit raters' views regarding the effectiveness of the training: The responses from a specifically created questionnaire were analyzed to discover any differences in understanding of the TBWT assessment criteria between trained and untrained raters. Second, to analyze ratings using the computer program "FACETS," which implements the "many-facet Rasch measurement model" (Linacre, 2008); this analysis created a detailed picture of each rater's behavior in terms of severity and consistency.

2. The Study

2.1 Purposes and Research Questions

This comparative study proposes to examine the effects of a rater-training session in relation to the degree of differences in raters' severity, consistency, and biased interactions between trained and untrained raters. The following research questions were addressed:
1) What are trained raters' perceptions of a rater-training session?
2) To what extent do trained and untrained raters differ in the spread of rater severity?
3) To what extent do trained and untrained raters differ in consistency in their judgment?
4) Are trained and untrained raters biased differently with respect to rater-subject interaction?
5) Are trained and untrained raters biased differently with respect to rater-task interaction?

2.2 Context of the Research
2.2.1 The Assessment Instrument

The task-based writing performance test (TBWT) is a scoring procedure for writing-performance assessment (WPA); TBWT was developed on the construct-based processing approach for testing (Bachman, 2002; Sugita, 2008, 2009a). The constructs of the TBWT are assumed to be accuracy and communicability. Accuracy comprises organizational skills and linguistic accuracy. Specifically, organizational skills are defined as the ability to create a logical structure that enables the content to be accurately acquired by the reader.
Linguistic accuracy concerns errors of vocabulary, spelling, punctuation, or grammar. *Communicability* is defined as fluency specialized for writing, in other words, communicative quality and effect. Communicative quality refers to the ability to communicate without causing the reader any difficulty; communicative effect concerns the quantity of ideas necessary to develop the response along with relevance of the content to the proposed task (Sugita, 2009b, 2010).

With regard to processing perspectives (Skehan, 1998), content-based support and form-focused stakes are necessary for *accuracy* tasks. An elicitation task (writing a letter) was chosen, and specific topics of self-introduction were given—in this case, the selected hypothetical situation that the student is going to stay with a host family in Britain. The student is to write a letter of self-introduction based on the specific topic; the intent is for the writer to focus more on writing accuracy than on content. On the other hand, *communicability* tasks need form-oriented support and meaning-focused stakes in order to write with a focus on content. A note-taking task preparing for a discussion was designed because it encouraged students to write their opinions or ideas about the topic, and it emphasized a meaning-focused response (see the specifications in Appendix A).

The underlying constructs served as a useful basis when developing rating scales for *accuracy* and *communicability*. The rating scales were comprised clear descriptions of each construct and of 5-point Likert scales (Appendix B). Moreover, the descriptors for each category display selected written samples as an explanatory part of the scale; hence, busy schoolteachers with limited training in WPA were able to understand the descriptors and work with them consistently. Moreover, raters were asked to rate holistically (an impressionistic-scoring method) each of the participants’ writing proficiency at five levels. In the present study, this method of scoring is called *impression*.

### 2.2.2 Participants

The participants in the research were 10 Japanese teachers of English, 5 from each of the following groups:

**Trained (TRN) Raters:** These raters (hereafter TRN) received a training session before rating. The training provided experience with the rating scales and the elicitation tasks, and thus they learned how to rate the TBWT using accuracy and communicability rating scales. This group consisted of five novice junior high school teachers, all of whom shared similar backgrounds in terms of qualifications of less than five years’ teaching experience. All these raters were female and native speakers of Japanese.

**Untrained (UNTRN) Raters:** These raters (hereafter UNTRN) did not have a training session beforehand, and thus were never exposed to the rating scales and the elicitation tasks. This group consisted of five experienced junior high school teachers, all of whom shared similar backgrounds in terms of qualifications of 10 or more years of teaching. Four of these raters were male and one was female. All were native speakers of Japanese.
There were two reasons for comparing these groups. First, the novice and experienced teacher groups represented the two major backgrounds of typical teacher-raters in Japanese secondary schools. Regarding the influential factors in the shaping of teacher beliefs, both groups value their own experience as learners (Johnson, 1994; Mok, 1994). Experienced teachers, however, place great importance on their own teaching experience (Mok, 1994). Thus, these two groups constituted a representative sample of the population about which the study would generalize. Second, the differences in teaching experience highlight how the training session affects raters’ severity and consistency. Assuming that there are differences in scoring between trained novice teacher-raters and untrained experienced teacher-raters, those differences may reveal the training session’s effects on their approach to rating each elicitation task.

2.2.3 Rater Training Session

Face-to-face training sessions were conducted individually with the five novice teachers. Each session lasted about an hour, and its training protocol is summarized below.

*First Stage:* The rater was asked to read through the scoring guide to understand 1) the theoretical rationales of the test, 2) the assessment task, 3) the implementation method of the testing, and 4) the scoring criteria (rating scales). Then, the rater was asked to read the explanatory segment and writing samples for the accuracy tasks, which exemplify different points on the scale, in order to understand the descriptors of “organizational skills” and “linguistic accuracy.” In the same way, the rater was asked to read the segment with the samples of communicability tasks to understand the descriptors of “communicative quality” and “communicative effect.”

*Second Stage:* The set of samples for accuracy was given to the rater in random order. This sample set provided the opportunity to rate the accuracy tasks themselves and familiarized the rater with the rating scale. Another set of samples was used for communicability tasks. This rating practice was conducted with the two sets. After the first practice was completed and the rater was familiar with both the rating scales, the previous set was given to confirm intra-rater reliability; that is, the rater’s consistency. When the rater displayed individual patterns of ratings that were lower or higher than warranted by the written samples, feedback on the rater’s biased patterns was provided to improve internal consistency in subsequent ratings.

*Final Stage:* The rater was asked to describe her impression of the rating scales. Moreover, the rater was given an opportunity to ask questions about the rating procedures. Since the primary purpose of the training sessions was to increase raters’ internal consistency in assigning ratings, forcing raters into agreement with each other was unnecessary. Instead, it was important to train raters to be self-consistent. Therefore, based on the shared understanding of the rating scales, it was desirable to adopt the practice of familiarizing each rater with the scales and exemplifying certain features of the construct at a point in the scale.
2.3 Methodology

2.3.1 Data Collection

Both TRN and UNTRN raters scored 40 identical scripts (20 scripts for each of two tasks) collected from 20 undergraduate students (6 males and 14 females) who took an English teaching methodology course in the first semester of 2008. The test takers were 14 second-year and 6 third-year students from the faculty of global policy management and communications. All the test takers were Japanese native speakers with an intermediate level of English language proficiency. The TRN raters received the study’s survey questionnaire soon after the rating session finished. The UNTRN raters were asked to complete the questionnaire at home about their experience of using the TBWT scoring guide after they finished rating.

2.3.2 Data Analysis

Both quantitative (frequency counts) and qualitative (thematic analysis) analyses were undertaken on the questionnaire responses. Ratings by the 10 raters were analyzed using FACETS, which provides comparative characteristics. First, rater behavior of both TRN and UNTRN was modeled using FACETS, which provides estimates of the ability of 20 subjects, the difficulty of tasks, and the severity of 10 raters on a common logit scale. In addition to providing logit estimates of the ability, severity, or difficulty of each element of each facet, FACETS also provides statistics indicating the relative spread of these estimates within each facet. In other words, the analysis provides information about the significance of any differences that exist among elements of a facet, for example, differences in severity among raters or ability among subjects.

In addition, FACETS analysis provides fit statistics for each element; this indicates the degree to which each element behaves according to the model’s predictions. In the case of raters, the fit statistics indicate rater consistency. Thus, a detailed picture is formed for each rater’s behavior in terms of severity and consistency. In the FACETS analysis of the comparative data, three facets were used: Subjects (n = 20), Raters (5 = TRN, 5 = UNTRN), and Tasks (accuracy, communicability, and impression).

3. Results

3.1 Responses to a Questionnaire

Designed for completion in a brief time, the training questionnaire for TRN raters consisted of 10 questions. Most questions were of the multiple-choice variety. The last question invited comments and opinions on the rater-training program. Questions 1 through 3 were answered on a 3-point scale from 1 (No, not useful) to 3 (Yes, useful). Questions 4 through 9 were answered on 4-point scale from 1 (Strongly disagree) to 4 (Strongly agree). As illustrated in Tables 1 and 2, the questionnaire results show that the five TRN raters strongly agreed that the rater-training session was useful. On Question 10, all the raters commented on the benefits of the training: “The practice by using the writing samples familiarized me with
the scoring procedure” (TRN1); “I was not sure of the scoring criteria through reading the scoring guide, but I found the training to be beneficial for understanding the criteria” (TRN2); “The training helped me understand the level of each writing sample,” (TRN3); “I became friendly with the scoring criteria through the manipulating practices” (TRN4); “After the training session, I became more confident with rating” (TRN5). These comments reflect those given in support of the multiple-choice responses to Questions 6 and 9, which sum up the TRN raters’ overall perception of their increased understanding of the TBWT assessment criteria.

Table 1
Results of Scaling in Questionnaire (Q. 1–3) for TRN Raters

<table>
<thead>
<tr>
<th>Questions</th>
<th>Yes</th>
<th>More or less</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Is the introduction useful?</td>
<td>5 (100%)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2. Is the task explanation useful?</td>
<td>5 (100%)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3. Is the scoring procedure useful?</td>
<td>5 (100%)</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2
Results of Scaling in Questionnaire (Q. 4–9) for TRN Raters

<table>
<thead>
<tr>
<th>Questions</th>
<th>Strongly agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. The definition of accuracy is understandable.</td>
<td>2 (40%)</td>
<td>3 (60%)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5. The accuracy scale is easy to evaluate.</td>
<td>0</td>
<td>5 (100%)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6. The training session for accuracy is useful.</td>
<td>5 (100%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7. The definition of communicability is understandable.</td>
<td>2 (40%)</td>
<td>3 (60%)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8. The communicability scale is easy to evaluate.</td>
<td>0</td>
<td>5 (100%)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9. The training session for communicability is useful.</td>
<td>5 (100%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

To confirm the TRN raters’ views on the training’s effectiveness, a questionnaire was also administered by mail to the five UNTRN raters. The nine multiple-choice questions were almost the same as that for the TRN raters, except that Questions 6 and 9 addressed the usefulness of the writing samples in the scoring guide. Moreover, the last question invited comments and opinions on the whole guidebook. As shown in Tables 3 and 4, the questionnaire results reveal that the five UNTRN raters felt that the TBWT scoring guide was fairly useful. However, there seem to be some differences in the responses to Questions 6 and 9. Although all TRN raters strongly agreed that the training session was useful, only one UNTRN rater strongly agreed with the usefulness of the writing samples in the scoring guide. The responses show that although both TRN and UNTRN raters learned about identical writing samples before rating, TRN raters fully admitted the usefulness of the session in which
the samples were used. In addition, the results indicate that TRN raters’ understanding of the TBWT assessment criteria was further promoted through practice on the sample. Overall, this implies that the training session increases raters’ internal consistency in assigning ratings. However, the small number of questionnaire responses limits any claim for generalizing the results.

Table 3
Results of Scaling in Questionnaire (Q. 1–3) for UNTRN Raters

<table>
<thead>
<tr>
<th>Questions</th>
<th>Yes</th>
<th>More or less</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Is the introduction useful?</td>
<td>3 (60%)</td>
<td>2 (40%)</td>
<td>0</td>
</tr>
<tr>
<td>2. Is the task explanation useful?</td>
<td>5 (100%)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3. Is the scoring procedure useful?</td>
<td>5 (100%)</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4
Results of Scaling in Questionnaire (Q. 4–9) for UNTRN Raters

<table>
<thead>
<tr>
<th>Questions</th>
<th>Strongly agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. The definition of accuracy is understandable.</td>
<td>0</td>
<td>5 (100%)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5. The accuracy scale is easy to evaluate.</td>
<td>1 (20%)</td>
<td>4 (80%)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6. The samples for accuracy are useful.</td>
<td>1 (20%)</td>
<td>4 (80%)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7. The definition of communicability is understandable.</td>
<td>0</td>
<td>5 (100%)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8. The communicability scale is easy to evaluate.</td>
<td>0</td>
<td>5 (100%)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9. The samples for communicability are useful.</td>
<td>1 (20%)</td>
<td>4 (80%)</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

3.2 FACETS Summary

Figure 1 summarizes the facets and their elements. They are positioned on a common logit scale that appears as "measure" in the first column. The second column shows the severity variation among raters. The most severe rater (ID: UNTRN3) is at the top, and the least severe rater (ID: TRN2) is at the bottom. The third column shows the ability variation among the 20 subjects. The subjects are ranked with high ability at the top (ID: 9) and low ability at the bottom (ID: 4 & 11). The fourth column shows that the difficulty variation among the tasks was equivalent. The last three columns graphically describe the three rating scales. Each of the two tasks and the total impression has its own scale. The most likely scale score for each ability level is shown: S1 for accuracy, S2 for communicability, and S3 for impressionistic scoring.

As the figure indicates, subject estimates range from a high of about 6 logits to a low of −6 logits, indicating a spread of 12 logits in terms of students’ ability. Subject separation value was 9.22, meaning that populations such as the students in this study may spread through about
9 levels. The reliability index was 0.99; this demonstrates the possibility of achieving reliable student ability scores. The column for raters shows that the three TRN and UNTRN raters tended to cluster between (−1) and (−2). One TRN and two UNTRN raters were more severe and one TRN rater was more lenient than the rest. Thus, the figure indicates that the raters were not at the same level of severity.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Raters</th>
<th>Subjects</th>
<th>Tasks</th>
<th>S.1</th>
<th>S.2</th>
<th>S.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td></td>
<td></td>
<td>(5)</td>
<td>(5)</td>
<td>(5)</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td>(5)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>UNTRN 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>UNTRN2 TRN3</td>
<td></td>
<td></td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td>TRN4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-2</td>
<td>UNTRN4 TRN1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-2</td>
<td>UNTRN5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-3</td>
<td>TRN2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 1. FACETS summary**

### 3.3 FACETS Analysis

1) To what extent do TRN and UNTRN raters differ in the spread of rater severity?

Table 5 provides information on the characteristics of raters. From the left, each column shows rater IDs, fair average scores, rater severity, error and infit mean square values. Raters
are presented in descending order of severity. The severity span between the most severe rater (ID: UNTRN3) and the most lenient rater (ID: TRN2) was 3.11; the difference, based on fair average scores in the first column, is 0.86 of one grade in the scale. In addition, summary statistics for the entire sample of raters and for the two groups of raters (TRN and UNTRN) are provided. In Table 5, the UNTRN raters are found to be more severe than the TRN raters. This suggests that the UNTRN raters tended to apply stricter standards overall to the written samples than did the TRN raters. Moreover, this tendency of the UNTRN raters is reflected in the mean severity estimates (−0.92 logits for UNTRN raters, −1.51 logits for TRN raters).

<table>
<thead>
<tr>
<th>Rater</th>
<th>Fair-M average</th>
<th>Severity (logits)</th>
<th>Infit (mean square)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNTRN3</td>
<td>2.86</td>
<td>0.53</td>
<td>0.25</td>
</tr>
<tr>
<td>UNTRN2</td>
<td>3.05</td>
<td>−0.20</td>
<td>0.25</td>
</tr>
<tr>
<td>TRN3</td>
<td>3.08</td>
<td>−0.32</td>
<td>0.25</td>
</tr>
<tr>
<td>TRN4</td>
<td>3.27</td>
<td>−1.06</td>
<td>0.25</td>
</tr>
<tr>
<td>UNTRN4</td>
<td>3.30</td>
<td>−1.19</td>
<td>0.25</td>
</tr>
<tr>
<td>TRN1</td>
<td>3.38</td>
<td>−1.44</td>
<td>0.25</td>
</tr>
<tr>
<td>UNTRN5</td>
<td>3.45</td>
<td>−1.69</td>
<td>0.25</td>
</tr>
<tr>
<td>UNTRN1</td>
<td>3.57</td>
<td>−2.06</td>
<td>0.25</td>
</tr>
<tr>
<td>TRN5</td>
<td>3.59</td>
<td>−2.13</td>
<td>0.25</td>
</tr>
<tr>
<td>TRN2</td>
<td>3.72</td>
<td>−2.58</td>
<td>0.26</td>
</tr>
<tr>
<td>Mean-All</td>
<td>3.33</td>
<td>−1.21</td>
<td>0.25</td>
</tr>
<tr>
<td>SD</td>
<td>0.26</td>
<td>0.93</td>
<td>0.00</td>
</tr>
<tr>
<td>Mean-TRN</td>
<td>3.41</td>
<td>−1.51</td>
<td>0.25</td>
</tr>
<tr>
<td>SD</td>
<td>0.23</td>
<td>0.79</td>
<td>0.00</td>
</tr>
<tr>
<td>Mean-UNTRN</td>
<td>3.25</td>
<td>−0.92</td>
<td>0.25</td>
</tr>
<tr>
<td>SD</td>
<td>0.26</td>
<td>0.96</td>
<td>0.00</td>
</tr>
</tbody>
</table>

All raters: Separation: 3.56 Reliability = 0.93; fixed (all same) chi-square: 137.1, df: 9; significance: 0.00

TRN raters: Separation: 2.99 Reliability = 0.90; fixed (all same) chi-square: 49.7, df: 4; significance: 0.00

UNTRN raters: Separation: 3.70 Reliability = 0.93; fixed (all same) chi-square: 73.8, df: 4; significance: 0.00

The FACETS analysis indicates a number of differences in severity among raters. These are the Separation Index, Reliability, and the Fixed (all same) chi-square. The Separation Index of Table 5 is the ratio of the true standard deviation of raters to the root mean-square standard.
error. If the raters are equally severe, the standard deviation of the rater difficulty estimates should be equal to or smaller than the mean estimation error of the entire data set. However, the Rater Separation Index for the entire sample of raters is 3.56, indicating that the variance among raters is about three times as large as the error of estimates.

The Reliability statistics provided by the FACETS analysis indicate the degree to which the analysis reliably distinguishes between different levels of difficulty or severity among the different raters. For raters, low reliability is desirable because, ideally, the different raters would be equally severe. In this case, however, the reliability is 0.93 for all raters, indicating that the analysis is fairly reliably separating raters into different levels of severity. Finally, the Fixed chi-square tests the null hypothesis that all the elements of the facet are equal. The chi-square of 137.1 with 9 df. is significant at $p = 0.00$, indicating that the null hypothesis must be rejected; in other words, the raters are not equally severe.

In comparing the range of severities for the two groups of raters, Table 5 shows that the UNTRN raters as a group vary much more in severity than the TRN raters, with a standard deviation of 0.96 logits, compared with 0.79 for the TRN raters. Moreover, this variability is reflected in the Separation Indices for the two groups: 3.70 for the UNTRN raters, and 2.99 for the TRN raters. However, it should be noted that both TRN and UNTRN raters differ significantly in their severity.

2) To what extent do TRN and UNTRN raters differ in rater consistency?

Furthermore, the FACETS analysis provides two measures of fit or consistency: the infit and the outfit. The infit is the weighted mean-squared residual that is sensitive to unexpected responses near the point where decisions are being made. The outfit is the unweighted mean-squared residual and is sensitive to extreme scores (Linacre, 2002). For the purpose of this study, only the infit statistics are reported. Fit statistics of $M + 2SD$ or greater indicate too much unpredictability in raters' scores, and a value of $M - 2SD$ or less indicates overfit or not sufficient variation in scores. Applying these standards to Table 5, all raters fall within the range of two standard deviations around the mean (0.98 ± [2 × 0.26]), indicating that no raters were identified as misfitting. In other words, both TRN and UNTRN raters behaved consistently in scoring.

In terms of consistency, the infit statistics also show slight group differences. The standardized infit has an expected mean of 0 and a $SD$ of 1. For the TRN raters, the mean is 0.86 and the $SD$ is 0.22; these figures for the UNTRN raters are 1.10 and 0.23, respectively. A distinction between TRN and UNTRN raters was made as a group, and thus the UNTRN raters were supposed to be less consistent.

3) Are TRN and UNTRN raters biased differently with respect to rater-subject interaction?

In Table 6, the first column shows the ability estimate range, and the second column shows the number of subjects within the particular range of ability estimate. In the range of 3.00 or higher, the total number of subjects was three, but five biased interactions were
observed. One TRN and one UNTRN rater harshly scored Subject 18. Although one UNTRN rater leniently scored Subject 12, another UNTRN rater harshly scored the same subject. In the range between −2.99 and 2.99, there were nine biased interactions. One TRN and one UNTRN rater harshly scored Subject 1. It should be noted that three UNTRN raters were significantly biased with Subject 3: two UNTRN raters harshly scored the subject, and one UNTRN rater leniently scored the same subject. In the range of −3.00 or lower, there were only three biased interactions among TRN raters, whose scoring of these subjects was lenient.

Table 6
Frequency of Rater-Subject Bias Interaction

<table>
<thead>
<tr>
<th>Ability</th>
<th>N</th>
<th>Harsh (Raters)</th>
<th>Lenient (Raters)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>TRN UNTRN</td>
<td>TRN UNTRN</td>
</tr>
<tr>
<td>3.00 higher</td>
<td>5</td>
<td>1 3</td>
<td>1</td>
</tr>
<tr>
<td>−2.99−2.99</td>
<td>9</td>
<td>3 4</td>
<td>2</td>
</tr>
<tr>
<td>−3.00 lower</td>
<td>3</td>
<td></td>
<td>3</td>
</tr>
</tbody>
</table>

Note: 17 out of 200 pairs (8.5%) show significant interaction.

The three cases in which TRN raters scored the subjects leniently comprised 43% of the seven rater-subject bias interactions (3/7 = 0.43). On the other hand, the UNTRN raters harshly scored the subjects in seven cases, which was 70% of the 10 rater-subject bias interactions (7/10 = 0.70). These results may indicate that the TRN raters were more lenient, and the UNTRN raters were more severe in scoring. Another finding is difference in the total number of subjects in the biased interactions. As for TRN raters, seven biased interactions were observed, but the total number of subjects was six. Although the two raters were biased with Subject 5, both of them leniently scored the subject. As for UNTRN raters, 10 biased interactions were observed, but the total number of subjects was seven. The two raters were biased with Subject 12; however, one of them scored the subject leniently, and the other, harshly. Moreover, three raters were biased with Subject 3, one of them leniently scored and the others scored the subject harshly. These results indicate that the total number of subjects in biased interactions was almost the same in both rater groups. However, UNTRN raters might be more biased than TRN raters; this is suggested by the UNTRN raters’ inconsistency in scoring.

4) Are TRN and UNTRN raters biased differently with respect to rater-task interaction?

The analysis of the two test tasks and overall impression in Table 7 shows that no significant variation in difficulty exists among them. Raters are considered to be self-consistent in scoring; the tasks do not appear to separate the students to a significant degree, meaning that the difficulty of the two tasks and the overall impression are equivalent. An estimate of the item discrimination was computed according to the Generalized Partial Credit Model approach. The expected value was 1.0; however, discriminations in the range of 0.5 to 1.5 provided a
reasonable fit to the Rasch model (Linacre, 2007).

Table 7

<table>
<thead>
<tr>
<th></th>
<th>Difficulty Error</th>
<th>Infit Mean Square</th>
<th>Estimate of Discrimination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>.02 .14</td>
<td>1.08</td>
<td>.92</td>
</tr>
<tr>
<td>Communicability</td>
<td>.17 .13</td>
<td>1.08</td>
<td>.91</td>
</tr>
<tr>
<td>Impression</td>
<td>-.1 .14</td>
<td>.76</td>
<td>1.26</td>
</tr>
</tbody>
</table>

Note. Reliability of separation index = 0.13; fixed (all same) chi-square: 3.5, df: 2; significance: p = 0.17

The results of the bias analysis in terms of the interaction between raters and tasks show that overall, there were 30 rater-task interactions (10 raters × 3 tasks), but only one interaction among TRN raters was significantly biased, TRN2 × Accuracy. The bias analysis indicates that the rater tended to award severe ratings to all the subjects on the accuracy task. The mean of the infit mean square value was 0.9, and its standard deviation was 0.3. Thus, fit values above 1.5 logits suggest misfit (0.9 + [0.3 × 2]). The value by UNTRN1 on communicability was 1.8; this indicates that the rater was inconsistent in evaluating the task in the identified patterns of bias across all subjects.

4. Discussion

4.1 Implications for Raters’ Severity

Study results indicate an influence by errors of measurement resulting from variation in rater severity. The rater-subject bias analysis shows particular raters’ tendency toward excessive severity or leniency toward particular subjects. Although the TRN raters exhibited a tendency toward leniency, the UNTRN raters awarded unexpectedly severe ratings to the subjects. According to McNamara (1996), “raters typically show highly reliable differences in severity, indicating that elimination of rater differences in severity is an unachievable goal” (p. 140). Indeed, previous empirical studies have shown that the relative severity of individual raters could not be reduced to acceptable levels even after rigorous rater training or among thoroughly trained raters (Lunz et al., 1990; Weigle, 1998). In this study, both TRN and UNTRN raters differed significantly in their severity, implying that rater training was not successful in inducing raters to give identical scores. These findings justify the use of FACETS analysis to compensate for inter-rater differences.

4.2 Implications for Raters’ Consistency

Furthermore, the bias analysis indicates possibly problematic rater behavior: UNTRN
raters showed inconsistency among raters in scoring Subjects 3 and 12; TRN raters exhibited consistency toward biased interaction with Subject 5. Despite the acceptable overall fit values for both TRN and UNTRN raters (Table 5), since bias analysis examines individual raters, it is assumed that TRN raters behaved more consistently in scoring. Therefore, there are differences in consistency between the TRN and UNTRN rater groups, indicating that rater training is effective in improving raters’ scoring consistency. If the TRN raters behave consistently, the rater-training session significantly succeeds in helping novice teachers score more predictably based on clear understanding of a definition of the ability being measured. As confirmed in previous studies, rater training may have limitations in reducing the differences or biases caused by variation among raters. However, shared understanding of the construct of writing ability could effectively be promoted by training sessions that familiarize raters with the procedure through practice on samples and increased experience in rating. This view of rater-training function accepts the notion that rater training more effectively improves rater consistency than severity in scoring.

4.3 Implications for Task Difficulty

Although a task was one source of score variance in the TBWT, the variance was negligible in terms of difficulty. Out of the 30 total interactions, only one interaction showed significant bias. The biased interaction occurred with an accuracy task scored by TRN2, indicating that the rater tended to award severe ratings on the task to all the subjects. The result confirms that this study’s assessment tasks validly inferred students’ writing performance, thus implying that task development by construct-based processing approach could be a reasonably solid basis for estimating students’ writing ability.

On the other hand, although tasks in the form of five rating categories were found to show acceptable fit to the Rasch model, the fit value of UNTRN 1 on the communicability task reveals that the rater was not consistent in scoring the task across all subjects. This finding was supported by evidence of biased rater-subject interaction discussed in the previous section. The training session may encourage novice teachers to clearly understand a definition of the ability being measured by the test and the five rating categories. The assessment tasks and scales were shown to be mostly comprehensible and usable by trained raters. As a result, rater training affected consistency in scoring. An experienced teacher-rater (UNTRN 1) was, however, supposed to understand the definition of the writing ability based on reading the scoring guide; however, the rater was not provided any opportunity to promote understanding, practice scoring on the samples, and increase experience. Consequently, UNTRN 1 may score the communicability task inconsistently across all subjects even though he is an experienced teacher. In addition, this finding suggests that rater training is more likely to affect raters’ consistency.
5. Conclusion

The questionnaire responses imply that the training session increases the raters’ self-consistency. The analyses of comparative data reveal that all raters as a group differ significantly from one another in terms of severity. Additionally, the UNTRN raters tended to be more severe than the TRN raters. Despite the experience of receiving rater training, all raters displayed acceptable levels of self-consistency (intra-rater reliability). According to the bias analyses, the total number of subjects in biased interactions was almost the same between rater groups, UNTRN raters might be more biased than TRN raters, as shown by the UNTRN raters’ inconsistency in scoring. In terms of rater-task interactions, a distinction between the TRN and UNTRN groups cannot be made, but one of the UNTRN raters was not consistent in evaluating the task in the identified patterns of bias across all subjects. These findings indicate that rater training is more effective in improving raters’ consistency than in improving their severity in scoring, replicating findings obtained in previous studies (McNamara, 1996; Weigle, 1994; Wigglesworth, 1993).

However, since this study’s scope was restricted, the results should not be overgeneralized. Several limitations should be addressed in future research. First, the present study included a small rater sample of novice and experienced teachers. Since experienced teachers’ beliefs may be very resistant to change (e.g., Tsui, 2003), future studies should include a number of raters who have similar teaching and rating experiences. Second, the present study presented only one each of accuracy and communicability tasks. The number of test takers was also small. Given this study’s limited amount of data, it is difficult to speculate about what may occur in a given situation. Many more similar analyses, perhaps on a larger scale, will be necessary in order to predict what may occur.

Acknowledgements

An earlier version of this paper was presented at the 16th Conference of Pan-Pacific Association of Applied Linguistics in Hong Kong on August 9, 2011. The participation in the conference was supported in part by a Grant-in-Aid for Scientific Research for 2010–2012 (No. 22520572) from the Japan Society for the Promotion of Science. I am deeply grateful to two anonymous reviewers whose suggestions have improved this article.

References


Johnson, K. E. (1994). The emerging beliefs and instructional practice of pre-service English second language teachers. Teaching and Teacher Education, 10 (4),
439-452.

**Appendix A: 課価タスク**

1. ライティング・タスク1

   (1) タスクの説明(rubric)
   
   この課題では、指定された内容を伝えうる英語の手紙をどのくらい正しい英文で書くことができるかをテストします。最初に1〜5のそれぞれの質問に対する自分の答を
考えます。次に、その内容をもとに英語の手紙の形式に合わせて、正しい英文で自己紹介の手紙を書きなさい。解答のための時間は20分間で、語数は100〜120語としなさい。なお、採点は次の3つの観点から行います。
① ～5の内容をすべて含み、自己紹介を目的とした手紙文としての適切さ
② 語的側面（文法・語彙・スペル等）の正確さ
③文章の形式的側面（構成・展開法）の適切さ

(2) タスクの指示
You are going to stay with Parker Family in Britain this summer. Write a 100–120 word letter introducing yourself to your host family. Before writing, think of the following topics.

- Your name and age
- Your job, profession, or majors in school
- Your family and pets
- Your interests and hobbies
- Your favorite places, foods, and activities
- Your experience in traveling abroad
- Some things you want to do while you are in England

2. ライティング・タスク2

(1) タスクの説明(rubric)
この課題では、限られた時間の中で英文によるメモを書き、英語でどのくらい情報伝えることができるかをテストします。最初に与えられた“Discussion Topic”に対する自分の答えをできるだけ多く考えます。次に、その内容を(例)にならい簡潔に、わかりやすい英語で書き表すようにします。解答のための時間は10分間で、語数に制限はありません。1つでも多くの考えを英語で書き表すようにしてください。なお、採点は次の3つの観点から行います。
① 題の意図をよく理解して書かれた内容になっているか
② 分の考えや意図がわかりやすく伝えられているか
③ 内容を効率的に伝えるために適切な語彙や表現を使用しているか

(2) タスクの指示
You are going to have a discussion on the following topic, “Why do you study English?”
To prepare for the discussion, think of many possible answers to the question. Then make notes about it the following table.

Discussion Topic: Why do you study English?
- (例)To travel abroad

- 76 -
Appendix B: 評価対象となる言語能力特性と評価基準
1. タスク1の評価基準

<table>
<thead>
<tr>
<th>評価対象となる言語能力特性</th>
<th>評価基準</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. 准則1の評価基準</td>
<td>きわめてあてはまる</td>
</tr>
<tr>
<td></td>
<td>B + (4) かなりあてはまる</td>
</tr>
<tr>
<td></td>
<td>B(3) わりとあてはまる</td>
</tr>
<tr>
<td></td>
<td>B- (2) 少しあてはまる</td>
</tr>
<tr>
<td></td>
<td>C(1) あてはまらない</td>
</tr>
</tbody>
</table>

この答えには、上記の評価基準が

A(5) きわめてあてはまる

B + (4) かなりあてはまる

B(3) わりとあてはまる

B- (2) 少しあてはまる

C(1) あてはまらない

「A(5) きわめてあてはまる」

[解説]

・文章の構成および展開がうまくできている
・論理展開の方法が適切で説得力がある
・部分的に誤りはあるが、語彙使用が適切である
・主語と動詞の一致、時制、単数・複数、語順および語法、冠詞、名詞、前置詞の使用にほとんど誤りがない
・スペル、句読法、大文字使用、段落分けの仕方にほとんど誤りがない

- 77 -
「B+(4) かなりあてはまる」

[解説]
- 文章の構成および展開ができている
- 論理展開の方法が概ね適切で全体的に理解できる
- 語彙使用が不適切で部分的に意味が分かりにくくなっているところがある
- 主語と動詞の一致、時制、単数・複数、語順および語法、冠詞、代名詞、前置詞の使用にやや誤りがある
- スペル、句読法、大文字使用、段落分けの仕方にやや誤りがある

「B(3) わりとあてはまる」

[解説]
- 文章の構成および展開がほぼできている
- 論理展開の方法は必ずしも適切ではないが、全体的にほぼ理解できる
- 語彙使用が不適切で意味が分かりにくくなっているところがある
- 主語と動詞の一致、時制、単数・複数、語順および語法、冠詞、代名詞、前置詞の使用に誤りがある
- スペル、句読法、大文字使用、段落分けの仕方に誤りがある

「B(2) 少しあてへまる」

[解説]
- 文章の構成および展開が不十分である
- 論理展開の方法が不適切で理解しにくい
- 語彙使用が明らかに不適切で、意味を取り違えたり、意味が分かりにくいところがある
- 主語と動詞の一一致、時制、単数・複数、語順および語法、冠詞、代名詞、前置詞の使用に誤りが多い
- スペル、句読法、大文字使用、段落分けの仕方に誤りが多い

「C(1) あてはまらない」

[解説]
- 文章の構成および展開ができていない
- 論理展開の方法や語彙使用についての知識がほとんどなく、理解しにくい
- 主語と動詞の一一致、時制、単数・複数、語順および語法、冠詞、代名詞、前置詞の使用に誤りが非常に多い
- スペル、句読法、大文字使用、段落分けの仕方に誤りが非常に多い
2. タスク２の評価基準

評価対象となる言語能力特性: Communicability (意味・内容の伝達を重視し、言語による効率的な情報伝達を行うことのできる能力)

<table>
<thead>
<tr>
<th>伝達内容の質 (Communicative quality): 書かれている内容を、読み手が支障なく、明瞭に理解することができる</th>
<th>情報伝達の効果 (Communicative effect):</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>・言語使用能力が確かなものであることがある</td>
<td>・与えられた課題に対して適切かつ十分に、関連性が明確である考えが効果的に示されている</td>
<td></td>
</tr>
<tr>
<td>・自分の考えを表現したり、意図を伝えることのできるすぐれた構文力・語彙力がある</td>
<td>・与えられた課題に対してそつ無く回答している</td>
<td>・課題に対する関連性が十分にある考えが数多く提示され、効果的に内容が伝えられている</td>
</tr>
</tbody>
</table>

この答えには、上記の評価基準が
A(5) きわめてあてはまる
B+(4) かなりあてはまる
B(3) わりとあてはまる
B−(2) 少しあてはまる
C(1) あてはまらない

「A(5) きわめてあてはまる」

【解説】
・与えられた課題に対してそつ無く回答している
・読み手に対して非常に明確に内容が伝わる
・言語使用能力が確かなものであることがわかる
・自分の考えを表現したり、意図を伝えることのできるすぐれた語彙力がある

「B+(4) かなりあてはまる」

【解説】
・与えられた課題に対して部分的に回答している
・読み手に対して十分に内容が伝わる
・十分な言語使用能力があるが、不確かに思われることももある
・自分の考えや意味を十分に伝えることのできる語彙力がある
「B(3) わりてはまる」

[解説]
・与えられた課題の一部に回答している
・読み手に対してほぼ内容が伝えられている
・言語使用能力はあるが、不確かな場合はところも多い
・自分の考えや意味をほぼ伝えることができるのでう足が安ある

「B(2) 少しであはまる」

[解説]
・与えられた課題のごく一部について回答している
・読み手に対して十分に内容が伝わらない
・言語使用能力が不足している
・足が安が欠如であり、自分の考えや意味を十分伝えることができる

「C(1) あてはまらない」

[解説]
・与えられた課題にほとんど回答できていない
・読み手に対してほとんど内容が伝わらない
・言語使用能力や足が安が不足し、自分の考えや意味をほとんど伝えることができない