Assessing Communicative Writing Proficiency
—An Expert Rater’s Decision Making when Scoring EFL Compositions—

Hideki OSHIMA
Kochi National College of Technology

Abstract

This article investigates how a native speaker of English assesses the communicative writing proficiency of EFL learners. From data collected in an essay writing task in a Japanese high school classroom, this article looks at the assessment behavior of the trained native-speaking rater of English and this rater’s underlying construct of writing assessment. Firstly, two raters were asked to evaluate the overall writing quality of 36 essay compositions with four different writing topics according to a 5-point holistic criterion-referenced scoring scale. Next, one of the raters was asked to assess whether each composition was convincing as a native-like English composition, with a 3-point analytic scale on 4 items: (1) the selection of appropriate words and phrases, (2) the conventions of grammar, punctuation, and spelling, (3) the method of organizing and presenting ideas and (4) the competent use of rhetorical devices to sustain the rater’s attention. Then the rater was asked to present reasons for the analytic assessment with written comments when he felt that the composition was fairly strongly convincing or unconvincing as native-like English. The results of an exploratory factor analysis and a qualitative analysis of the rater’s comments revealed that the rater’s assessment was made on the judgment of both language-specific and writing expertise-specific criteria and exhibited that the rater’s decision-making behavior was strongly influenced by the writing expertise-specific criterion.

1. Introduction

The use of performance-based assessments has gained wide acceptance in the domain of writing studies. In operational scoring, assessment is carried out holistically or analytically according to the criterion-referenced scoring rubric by the intuitive judgment of trained and qualified native or native-like
raters. As Stern (1983) pointed out, the native speaker’s competence, proficiency, or knowledge of the language is a necessary point of reference for the proficiency concept used in language teaching. In the case of writing assessment, it is often taken for granted that the appropriate reference for writing ability lies in the intuitions of native speakers. However, as the reference for assessment, little is known about the underlying construct of such native speakers’ intuitions. This article examines how the native speaker’s intuition explains any differences and variations in operational scoring and what factors are responsible for the rater’s assessment. This article is written from the perspective that writing is an act of communication that involves an interaction between the writer and the reader, here, the rater. Writing is an act to shape ideas and thought appropriately in a context on paper. To satisfy his or her goal the writer generates a text, and in turn, the reader reads through that text to receive information. When assessing the quality of writing, the reader/rater is the final judge of the success or the failure of the text the writer produced. Hence this article argues that the writing proficiency of nonnative EFL learners is judged according to their communicative efforts, namely, whether compositions are convincing, meaning the native English-speaking reader assumes that the text was written by a native English speaker, or unconvincing, meaning the native English-speaking reader is aware that the text may not have been written by a native English speaker.

2. Background of the Study

The performance-based measure of writing skills is perceived to be of ongoing importance to writing assessment. Writing assessment has moved away from traditional multiple-choice indirect assessment to more use of direct performance-based measures. Performance-based assessment of writing is a product-centered measure. The most commonly used formalized approaches to evaluating writing are holistic and analytic scorings. Holistic scoring, which aims to rate the overall writing proficiency, is made by reading a given sample of writing and deciding on a general score based on a scoring rubric, a set of brief writing specifications for grading. Analytic scoring, which scores the various features of a composition separately, seeks to provide precise diagnostic feedback to the learners. Unlike holistic scoring, analytic scoring focuses on assessing specific aspects of writing such as grammar, vocabulary, mechanics, organization, and content according to the separate evaluative criteria of each. Holistic scoring is based on the view that the quality of the whole is more important than the quality of the sum of the parts. Meanwhile, the rationale for analytic scoring is that an ultimate score of writing is evaluated by multiple elements, which will offer a more reliable and valid measure than a single score. Which scoring to use in operational scoring depends on the purpose of the assessment. Hughes (1989) indicates that analytic scoring is appropriate for the diagnostic measurement of individual writing skills, while holistic scoring is suitable for evaluating communicative aspects of writing proficiency as a whole. Holistic scoring has often been used in large-scale standardized writing assessment, while analytic scoring has been common in classroom contexts.

In most standardized test situations, performance-based assessment is conducted by judging an essay composition on a certain topic. In scoring, the trained competent rater assigns a grade to each
composition, applying specific evaluative criteria and following the procedures set out in a scoring guide. The rater is usually a native or a native-like speaker of English and an expert of EFL/ESL writing. Each rater makes decisions about the final placement of each composition or each component feature of a composition within a relatively limited time, relying on an established baseline characterized by impression marking. Such assessment, therefore, appears to be highly subjective and even unreliable, but in reality the reliability between qualified raters is fairly high with a range of coefficient from .75 to .90 even though scores are not identical, as verified in a considerable number of studies (see examples in Stansfield & Ross (1988); Eliot (1990); Shohamy et al. (1992)). These studies indicate that the trained raters consistently make valid and reliable judgments by responding to surface features of learner’s writing, and this tendency is especially remarkable in holistic scoring.

With regard to the rater’s decision making on scaling, several studies have been undertaken to identify their characteristics. Astika (1993) investigated what aspects affected comprehensibility, acceptability, and irritation on the part of the raters in analytic scaling based on the ESL Composition Profile by Jacobs et al. (1981). Among the five measure components of content, organization, vocabulary, language use, and mechanics, Astika found that vocabulary accounted for the largest amount of variance in total scores: 83.75% contribution to total scores in the stepwise regression analysis and added that vocabulary, language use, and content were highly correlated and it became difficult to interpret what was causing the variation in total scores. Finally, Astika concluded the study with discrepancy: the raters paid more attention to vocabulary component, mainly wording and word choice in this case, than the other analytic measuring components in the study; however, other factors such as content, organization, language use, and mechanics should also be considered as the predictors of the subjects’ writing proficiency scores. This inconsistency in interpretation could be caused by the apriorism in the study that the sum of analytic components will be a valid score of the writer’s overall writing quality. In point of fact, the totals of the various scoring components of analytic scoring do not necessarily explain learner’s writing proficiency as a whole. As White (1985) pointed out, currently there is little evidence to support the assumption that writing quality is the result of the accumulation of a series of sub skills. On the issue of the rater’s decision making, Astika’s study seems to call for further research from different perspectives.

As for the raters’ decision-making behaviors, Cumming (1990) studied the method of grading when expert and novice raters analytically scaled four levels of ESL compositions grouped by ESL proficiency and the learners’ mother tongue writing skills. The statistical analyses showed that both raters produced scores conforming closely to grouping by proficiency and writing skill: the raters assessed the compositions by distinguishing qualities of language proficiency and writing skill concurrently, implicitly attributing their scorings to the two different factors. The raters’ verbal reports further revealed that there was substantial difference in the quality of decision-making behaviors between the expert and the novice raters: the expert raters utilized a wide range of criteria, knowledge and strategies to read and judge compositions, while the novice raters used very limited sources and strategies of judgments based often on inexplicit criteria. In the study, Cumming did not give any definitions of the analytic categories, language use, rhetorical organization, and substantive content to the raters. The raters were only asked to assess the effectiveness
of the compositions according to the scaling categories and the raters themselves were allowed to fix the basis for their scaling. Cumming concluded that even the decision-making processes of the experienced raters varied from rater to rater in the absence of pre-specified criteria and procedures for evaluating. The study suggests that analytic scales have the advantage of drawing raters' attention to specific aspects of learners' compositions, as well as appropriate evaluation strategies and criteria. That is, analytic criteria will have a controlling influence on the impact of raters' personal experience, variations, and expectations.

In the study of Conner and Carrell (1993), verbal-report techniques following Cumming's coding scheme and statements about essays were used to investigate the criteria that the expert raters were actually drawing on when they were doing holistic scoring. Verbal reports exhibited that the raters' decisions were related to the elements of content, development, and language, and little concern was shown about text organization. The raters also spent considerable amounts of effort in thinking about marking factors such as overall impression, a point of reference in scoring, establishing personal judgment strategies and criteria, and summarizing judgments concurrently. Conner and Carrell emphasized that the selection, training, and calibrations for the raters were important consideration in arriving at consistent as well as reliable holistic assessment. As illustrated in Cumming's study, without any guidelines, scorings tend to be highly individualistic and truly impressionistic. Explicit guidelines, specific scoring procedures, and trained raters will lead to more appropriate and rigorous direct performance-based writing assessment. Thereby such assessment would become more valid and more reliable.

Thus far this chapter has reviewed current concerns and research relating to performance-based writing assessment. New challenges to such assessment have highlighted a considerable number of implications for performance-based assessment of writing. In particular, the following suggestions derived from the literature should be kept in mind for further research.

(1) Holistic scoring provides more reliable scores in standardized writing proficiency tests.
(2) Overall quality of writing is not necessarily explained by the sum of the writing sub-skills.
(3) When compared with novice raters, trained raters utilize extensive rating resources and strategies and their scaling is highly consistent and reliable.
(4) Raters assess compositions implicitly distinguishing writing skill from language proficiency.
(5) Analytic criteria of scoring draw the rater's attention to specific aspects of compositions, and will guide the rater to more appropriate and informed judgments.
(6) The analytic categories of content, language, organization, vocabulary, and mechanics would be related to the elements that explain the rater's decision making when scoring writing, although their underlying interrelationship is still unknown.

Now the question of the exact attributes of the rater's intuitive judgments and the basis for writing evaluation remains unresolved. The next step forward in the study of performance-based measure is to reach more precise understanding of what constitutes the rater's implicit evaluative criteria in decision making when scoring essay compositions.
3. The Study

3.1 Purpose of the Study

The purpose of the present study is to investigate what factors underlie the rater’s decision making when scoring essay compositions. This research study was conducted in order to clarify the trained native-speaking rater’s underlying evaluative criteria and basis for writing assessment.

3.2 Research Questions

In the present study the following research questions were addressed.

1. What kind of underlying factors are related to the rater’s decision making when scoring essay compositions?
2. What elements does the native English-speaking rater view as convincing or unconvincing when assessing essay compositions?

3.3 Participants and Tasks

A total of nine second-grade Japanese high school students participated in the study. The participants were enrolled in an EFL class when the study was carried out. Four different essay writing tasks were assigned to each participant at intervals in order to avoid a possible rating effect and control context variables. The tasks involved two types of stimulus, of the kind used in the Test of Written English (TWE), namely, a compare and contrast type and a take-up-a-position and argue type. The tasks were provided in the following order: Task 1 (a compare and contrast type); Task 2 (a take-up-a-position and argue type); Task 3 (a compare and contrast type); Task 4 (a take-up-a-position and argue type). Intervals were placed between the tasks to maintain consistency and keep away a halo-effect: one-month interval between Tasks 1 and 2 and between Tasks 3 and 4; and six-month interval between Tasks 2 and 3. The essay writing tasks were carefully designed, piloted and developed with the help of the EFL/ESL writing specialist’s consultation. For the participants, twenty minutes were allocated to plan and write each composition. The essay writing prompts are presented in Appendix A.

3.4 Scoring Procedures

The aim of this study is primarily to assess the communicative value of EFL learners’ compositions and clarify the underlying factors that are related to such assessment. First, the study needed to evaluate the overall quality of compositions and then investigate the factors explaining the underlying construct of writing assessment. Hence the holistic measure was adopted to assess the writing quality and the analytic measure and the written comments on rating were employed to analyze the rater’s decision making in scoring. At first, the communicative quality of the compositions was holistically measured by two trained raters according to a 5-point criterion-referenced scale (see Appendix B). Then, in order to investigate the scoring behavior of convincing and unconvincing, one of the raters was asked to reevaluate the same compositions according to a 3-point scale on four measures for (1) the selection of appropriate words and
phrases (*appropriate language use*), (2) the conventions of grammar, punctuation, and spelling (*language conventions*), (3) the method of organizing and presenting ideas (*ideas and content*), and (4) the competent use of rhetorical devices to sustain the rater's attention (*development of the topic*). The measures were selected on the implications obtained in the previous chapter. Finally, the same rater was asked to give explicit statements explaining his assessment and whether he felt fairly *convinced* or *unconvinced* about each category as a native speaker. The raters were both EFL/ESL writing specialists of native speakers of English who had been engaged in standardized writing proficiency tests over years in the United States.

### 3.5 Analysis

The research questions were addressed using two methods:

1. an exploratory factor analysis to explore the underlying factors in the four components of *appropriate language use* (wording, vocabulary, and phrases), *language conventions* (grammar, punctuation, and spelling), *ideas and content* (organizing and presenting ideas), and *development of topic* (rhetorical tactics and development devices);

2. a qualitative analysis of the written comments on composition, in response to the question of what elements were considered to be fairly *convincing* or *unconvincing*.

In addition, holistic scores were statistically analyzed to examine the task validity and the intra-rater reliability.

### 4. Results and Discussion

The findings revealed that the two trained raters measured the compositions with high rate of consistency and reliability (see Table 1). The intra-rater reliability for holistic scoring based on Spearman-Brown prophecy formula was very high: \(r = .9658\).

<table>
<thead>
<tr>
<th>Task</th>
<th>N</th>
<th>Average</th>
<th>SD</th>
<th>Median</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>9</td>
<td>3.222</td>
<td>0.833</td>
<td>3</td>
<td>.9561*</td>
</tr>
<tr>
<td>Task 2</td>
<td>9</td>
<td>3.000</td>
<td>1.000</td>
<td>3</td>
<td>.9466*</td>
</tr>
<tr>
<td>Task 3</td>
<td>9</td>
<td>3.222</td>
<td>0.667</td>
<td>3</td>
<td>.9449*</td>
</tr>
<tr>
<td>Task 4</td>
<td>9</td>
<td>2.667</td>
<td>1.118</td>
<td>3</td>
<td>.9338*</td>
</tr>
</tbody>
</table>

*p < .000

Table 2 indicates that acceptably positive correlations were found between the tasks and the range of the correlation slightly varied from task to task. This variance verifies that each task reflected different aspects of each participant's writing proficiency and how the raters evaluated various types of compositions. As a result, a total of 36 compositions were selected as samples.
Table 2. Spearman’s rank correlation matrix for the essay composition tasks

<table>
<thead>
<tr>
<th>Task</th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Task 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task 2</td>
<td>.561 (p&lt;.116)</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task 3</td>
<td>.609 (p&lt;.082)</td>
<td>.574 (p&lt;.103)</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Task 4</td>
<td>.758 (p&lt;.012)</td>
<td>.761 (p&lt;.017)</td>
<td>.674 (p&lt;.046)</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Tables 3 and 4 present the results of the exploratory factor analysis, using Varimax rotation, showing whether the compositions were convincing or not by native-speaker’s standards. Three factors were extracted as the latent predictors that explain the underlying relationship among analytic rating categories and the evaluative basis for rating compositions. The cumulative contribution of convincing or unconvincing English factors was 94.9% in Table 3 and 97.1% in Table 4. The results had statistically significant value to demonstrate the rater’s underlying construct for writing evaluation.

As for the convincing factors (Table 3), Factor 1 had the greatest effect (.951) on ideas and content, which served as the underlying basis by which to measure the ability to organize thoughts and ideas that engaged the rater in writing. Factor 2 showed an acceptable effect (.424) on development of topic, although Factor 2 received weak loadings from ideas and content (.152) and language conventions (.127). The factor concerned the evaluation of the ability to develop the essay topic with proper fluency and elaboration. Factor 3 had the greatest effect (.982) on language conventions, which contributed as the criterion to measure the ability to produce lexically and syntactically well-formed language.

Table 3. Results of the exploratory factor analysis: convincing factors (N=36)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ideas and content</td>
<td>.951</td>
<td>.152</td>
<td>-.110</td>
</tr>
<tr>
<td>Development of topic</td>
<td>-.814</td>
<td>.424</td>
<td>-.261</td>
</tr>
<tr>
<td>Language conventions</td>
<td>.025</td>
<td>.127</td>
<td>.982</td>
</tr>
<tr>
<td>Appropriate language use</td>
<td>.027</td>
<td>-.971</td>
<td>-.149</td>
</tr>
<tr>
<td>Eigenvalue</td>
<td>1.568</td>
<td>1.161</td>
<td>1.067</td>
</tr>
<tr>
<td>Contribution (Cumulative)%</td>
<td>39.2 (39.2)</td>
<td>29.0 (68.2)</td>
<td>26.7 (94.9)</td>
</tr>
</tbody>
</table>

As for the unconvincing factors (Table 4), Factor 1 greatly influenced development of topic, whose factor loading was .779. Factor 1 was characterized as the criterion to evaluate the ability to present the topic clearly and smoothly. Factor 2 obtained loadings from ideas and content (.171), development of topic (.589), and appropriate language use (.125). The factor was relevant to the criterion to measure the ability to elaborate on the topic in a proper manner. The qualitative difference between Factors 1 and 2 was mainly due to the extent of the factor loadings on ideas and content and language conventions.
Table 4. Results of the exploratory factor analysis: unconvincing factors (N=36)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ideas and content</td>
<td>-.951</td>
<td>.171</td>
<td>-.175</td>
</tr>
<tr>
<td>Development of topic</td>
<td>.779</td>
<td>.589</td>
<td>-.041</td>
</tr>
<tr>
<td>Language conventions</td>
<td>.038</td>
<td>-.973</td>
<td>-.148</td>
</tr>
<tr>
<td>Appropriate language use</td>
<td>.118</td>
<td>.125</td>
<td>.983</td>
</tr>
<tr>
<td>Eigenvalue</td>
<td>1.525</td>
<td>1.339</td>
<td>1.020</td>
</tr>
<tr>
<td>Contribution (Cumulative) %</td>
<td>38.1 (38.1)</td>
<td>33.5 (71.6)</td>
<td>25.5 (97.1)</td>
</tr>
</tbody>
</table>

Factor 3 showed the greatest effect (.983) on appropriate language use, which involved the evaluative basis for measuring the ability to use language appropriately in writing.

The results of the exploratory factor analysis showed that the extracted factors were classified into two categories and each of them differently affected the rater's decision making. Factors 1 and 2 in Table 3 and Factors 1 and 2 in Table 4 were the factors pertaining to the writing expertise-specific category. Factor 3 in Table 3 and Factor 3 in Table 4 were the factors pertaining to the language-specific category. The results demonstrated that the factors pertaining to the writing expertise-specific category were closely related with the rater's decision-making behavior in rating the compositions. The cumulative contributions of the factors were 68.2% in Table 3 and 71.6% in Table 4. The results provided sufficient information to enable us to detect the underlying construct of the rater's decision making.

The rater's statements in response to the question of what elements were considered to be fairly convincing or unconvincing by a native speaker's standards were summarized as follows. These statements were qualitatively classified into two categories: writing expertise-specific and language-specific statements.

1. Rater's statements on convincing English writing

   Writing expertise-specific statements: the ideas are clearly described and stated, showing good examples and interesting facts; the things written are well explained and discussed, giving clear reasons to support opinions; the essay is appropriately organized with rhetorical patterns such as cause and effect, and comparison and contrast; the ideas presented are well connected to the topic.

   Language-specific statements: the words and sentences are well formed and arranged; a variety of sentences and words are used to communicate own ideas.

2. Rater's statements on unconvincing English writing

   Writing expertise-specific statements: the writing tends to be general or incomplete and needs more specific information and examples; several different ideas are irrelevantly presented; the writer's ideas are not clear and unnecessary information is contained; the reasons for the writer's statements are unclear.

   Language-specific statements: the writer needs to choose more specific words and expressions; it is difficult to understand how the ideas are related to each other; paragraphs would help readers
understand the ideas better.
The results showed that the rater's statements highly pertinently conformed to the factor loading tendencies presented in Tables 3 and 4.

The present study revealed that the rater evaluated the compositions on the basis of the two underlying criteria: one was writing expertise-specific criterion, which was explained by Factors 1 and 2 in Tables 3 and 4 and the rater's writing expertise-specific statements; the other was language-specific criterion, which was explained by Factor 3 in Tables 3 and 4 and the rater's language-specific statements. These criteria strongly affected the rater's judgment of whether the compositions were convincing or unconvincing. The study also attested Cumming's research implication that the rater assesses compositions implicitly distinguishing writing skill from language proficiency. The present findings explicitly indicated the latent criteria underlying the rater's intuitive evaluation of compositions.

5. Closing Thoughts

This study provides an insight into the causal structure underlying the observed evaluative behaviors of a trained and experienced native English-speaking rater. The exploratory factor analysis and the rater's statements demonstrated that the rater utilized two qualitatively different evaluative criteria as the basis for decision making in scoring compositions. The results found that the rater's assessment was strongly influenced by writing expertise-specific criterion, which was the evaluative basis for writing-specific knowledge and skills. However, the implications of the study are restricted in some respects. The study focused on the analysis of a single rater's evaluative construct in rating compositions and limited generalizability. In addition, the writing tasks provided in the study were of restricted types and stimuli. The study calls for further research for more precise and rigid examination of the raters' intuitive evaluative criterion/criteria and for the understanding of raters' decision-making behaviors.

References


Appendix A: Essay Prompts
1 Of the following two pictures, which one do you want to have? Showing the good point and the bad point, write your own opinion.
2 Now family members cannot spend their time together. Do you agree or disagree with it? Write your own opinion, giving some reasons.
3 Comparing the pictures, write your own opinion. Which one do you select if you can stay there?
4 Japanese high school students have to study hard. Do you agree or disagree with it? Write your own opinion, giving some reasons.

Appendix B: Rating Scales of Holistic Scoring (Summary)
5 An essay at this level effectively addresses the task. It is well organized and developed, illustrating ideas and specific information. Sentences are logically connected. It shows appropriate syntactic variety and word choice.
4 An essay at this level addresses the task adequately. It is properly organized and developed. Sentences are consistent. It demonstrates word and syntactic variety.
3 An essay at this level addresses the task adequately, but it reveals some inadequate organization or development. It uses some details to illustrate ideas and specific information. It demonstrates some word and syntactic variety.
2 An essay at this level develops the task in a minimal way and reveals inadequate organization and development. It uses a small variety of words and sentences. It tends to be general and incomplete.
1 An essay at this level is very short with little detail. It may be undeveloped or incoherent. Sentences are very simple and incomplete.

Appendix C: Sample Composition (the Original Text in Task 3)
I want to go to the mountains. I think that we had better spend time to relax. Because we hardly have a time to relax in our recently life. So I want to relax with my family in the beautiful mountains like a picture. But if I go to the amusement park which has the wheel, the couster, and many things to enjoy ourselves, I want to go with my friends. But now I thinking that both the amusement park and the mountains make us change our mind and enjoy ourselves. (Holistic grade: 4/4) (Analytic grade: convincing Ideas and content 3; Development of topic 2; Language conventions 2; Appropriate Language use 2; unconvincing Ideas and content 1; Development of topic 1; Language Conventions 3; Appropriate language use 2)