Title	Assessment of learners' performance on writing tasks : improvements based on a case study
Sub Title	ライティング課題における学習者のパフォーマンス評価 : ケーススタディを基にした改善の試み
Author	嶋田, 和成(Shimada, Kazunari)
Publisher	慶應義塾大学外国語教育研究センター
Publication year	2019
Jtitle	慶應義塾外国語教育研究 (Journal of foreign language education). Vol.16, (2019.) ,p.109- 119
JaLC DOI	
Abstract	
Notes	研究ノート
Genre	Departmental Bulletin Paper
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=AA12043414-20190000-0109

Assessment of Learners' Performance on Writing Tasks: Improvements Based on a Case Study[1]

SHIMADA, Kazunari

Introduction

In assessing learners' language use, two popular evaluation methods have been often used: holistic and analytic evaluation. While holistic evaluation is based on rating the overall performance in learners' speech or writing, analytic evaluation is done by scoring the language productions in terms of some components such as content, organization, vocabulary and grammar of the text. Holistic scoring enables teachers to efficiently and quickly evaluate their students' speech or writing due to the simplicity of the rating scale. The simple and quick method is useful for busy teachers, but is often dependent on their general impression of the students' productions. On the other hand, analytic scoring is regarded as one of the most effective systematic evaluation methods because teachers can objectively grade the students' productions from various aspects and subsequently provide the students with diagnostic feedback for points of improvement. However, the disadvantage of this method is that it is time consuming and requires considerable effort.

Although a comparison between the two evaluation methods has been made by many researchers (e.g., Hamp-Lyons, 1991; Weigle, 2002), it would be difficult to decide which is better, given that each method has its advantages and disadvantages. For example, in English language proficiency tests, the TOEFL iBT® Test adopts a holistic rubric with rating from zero to five, while the Eiken STEP Test uses an analytic rubric. Therefore, this article focuses not on a subject that has already been extensively discussed, but on improving validity, practicality, and reliability of writing assessment using a rubric or rating scale. In other words, the case study in this article is conducted to demonstrate how language teachers can assess their students' writing skills more effectively and efficiently.

What Should Be Assessed in English Writing?

Many researchers and teachers believe that fluency, accuracy, and complexity of language are important aspects to assess second language learners' productive skills (e.g., Housen, Kuiken, & Vedder, 2012; Skehan, 2009). In scoring students' writing, fluency is often defined as the number of words written by them (e.g., Kamimura, 2006), accuracy is measured as to what extent they can use appropriate grammar and vocabulary, and complexity involves a variety of syntactic structures and the average number of words per sentence.

However, it is quite difficult for teachers to properly assess their students' writing skills from these three aspects. The author conducted an analysis of some texts collected in a teacher training class at a private university in Gunma Prefecture in 2018. Six students wrote a paragraph about the best way to learn a foreign language[2] and the author graded the writing. Below are two examples:

(1) The best way to learn a foreign language is listening to English music. That's because, it's difficult for student who aren't good at English to learn it every day. However, people like music, so students can keep to learning motibation.
(2) The best way to learn a foreign language is to use the language every day. I think that it is necessary to use the foreign language positively. So, I can use the language.

In example (1), the student tried to use difficult words such as *motivation* and long and complex sentences including a relative clause and a subordinate clause, but made lexical and grammatical errors in the handwritten opinion. On the other hand, in example (2), the student used easy words and short and simple sentences so as to avoid lexical and grammatical errors. However, the phrase *use the (foreign) language* was repeatedly used. Thus, while example (1) is more complex than example (2), example (1) is less accurate than example (2).

Concerning an analysis of fluency in the two examples, the author used the average number of words per T-unit (Wolfe-Quintero, Inagaki, & Kim, 1998) because the students were asked to write 30-40 words for the writing task. T-unit is a measure to count the number of clauses: the unit indicates one main clause or "one main clause with all the subordinate clauses attached to it" (Hunt, 1965, p.20). The average number of words per T-unit in example (1) was 9.50, while in example (2), it was 10.67. As a result, a significant difference in fluency was not found between the two examples.

Some teachers may value a variety of syntactic structures in example (1), while other

teachers may prefer the error-free writing in example (2). Thus, the assessment of writing skills can strongly depend on determining which aspects of performance are to be prioritized in any given situation. For example, in the case of entrance examinations, writing with no errors is likely to be required.

Table 1 is a summary of the analytic evaluation.

Table 1

Examples of Analytic Evaluation

	The number of words	Fluency (using T-unit)	Accuracy	Complexity
Example (1)	40	9.50	Lexical error (spelling) Grammatical error (agreement)	Difficult words Including a relative clause and a subordinate clause
Example (2)	33	10.67	No errors	Easy words Short and simple sentences

Using Rubrics for Assessing Learners' Writing Skills

For teachers, rubrics are useful to systematically assess students' writing skills in terms of some aspects including fluency, accuracy, and complexity. One of the most notable rubrics for scoring writing performance is the ESL Composition Profile which was proposed by Jacobs, Zinkgraf, Wormuth, Hartfiel, and Hughey (1981, p.30). The rubric consists of content, organization, vocabulary, language use, and mechanics (i.e., spelling, punctuation, capitalization, and paragraphing) sections to assess English as a second language (ESL) learners' writing skills. While learners' language fluency can be assessed in the content and organization sections, their language accuracy can be assessed in the vocabulary, language use, and mechanics sections. Their language complexity can also be assessed in the language use section. Each section is measured on a four-level scale, with the allotment of points as follows: (a) 13-30 points for content; (b) 7-20 points for organization; (c) 7-20 points for vocabulary; (d) 5-25 points for language use; and (e) 2-5 points for mechanics. In other words, the points are allotted according to the relative importance of the sections in assessing writing skills. Based on Jacobs et al. (1981), several researchers (e.g., Okubo, 2006; Yamanishi, 2004) used the rubric or modified one as a scale for scoring writing performance.

Using rubrics such as the ESL Composition Profile can help teachers assess their students' performance from a variety of aspects in writing, whereas scoring in multiple sections may require considerable time and effort. Additionally, existing rubrics may not meet teachers' needs because the aspects of performance that are prioritized depend on the purpose of the writing in their class.

A Case Study: Assessing Learners' Performance on Integrated Writing Tasks

Positioned against this contextual background, this article explores how language teachers can assess their students' performance in tasks that integrate reading with writing. As Chan, Inoue, and Taylor (2015) point out, there have only been a few studies on the development of rubrics to assess learners' performance of integrated writing tasks. However, the use of integrated tasks is becoming common in language courses, examinations, and standardized tests such as the TOEFL iBT® Test. The purpose of the case study is to consider validity, practicality, and reliability of assessment using an original rubric for integrated writing tasks.

Method

Participants were 51 undergraduates from a private university in Tokyo who took a general English class in the spring semester (from April to July 2018). They were assigned to two classes (i.e., Classes A and B) and were taught by the same instructor. An initial placement test showed that almost all students were at a pre-intermediate or intermediate level of English proficiency. They were from the faculty of letters, and majored in a field of the humanities such as literature, history, philosophy, and psychology.

Both courses provided writing and discussion activities based on reading materials, but the contents differed according to their class. The textbook used for Class A was *Select Readings Intermediate* (Lee & Gundersen, 2011), while the textbook for Class B was *Inside Reading 2* (Zwier, 2012). Both textbooks cover a wide variety of topics such as business, education, psychology, and sociology, with exercises to enhance students' communication and writing skills as well as their reading comprehension. In writing activities, the instructor focused on paraphrasing and summary writing using passages in the textbooks.

After the final class of the semester, the term examination was given to the 51 participants. The students were asked to complete a writing task and answer vocabulary, grammar, and reading comprehension questions on the examination within 60 minutes. While the students in Class A were required to read a paragraph of 95 words about culture shock (see Appendix

A) and create a summary in 15-25 words, the students in Class B were required to read two paragraphs of 145 words total about face-recognition technology and create a summary in 40-50 words. They were not allowed to use dictionaries in the examination. Although the paragraphs had already been read in the classes, the students had not made any summaries before the term examination. Additionally, the paragraphs were excerpted from the textbooks with careful consideration to avoid the use of words and phrases in the other questions on the examination. The students' handwritten products were rated twice by the instructor, using a rubric and a three-week interval to increase the reliability of evaluation. At the beginning of the next semester, marked writing products with feedback comments with were given back to the students (see Appendix B).

Creating an Analytic Rubric

To address the purpose of the study, an analytic rubric was created by the author. First, constructs of the assessment were listed as follows:

1. The written product includes the main idea of the reading text and demonstrates an understanding of it.
2. The written product is based on the text paraphrased in the students' own words, not copied from the reading text.
3. The written product is grammatically correct.
4. The choices of words and phrases in the written product are appropriate.

While constructs 1 and 2 involve the content and organization of the written product, as well as reading comprehension, constructs 3 and 4 are intended to assess students' writing accuracy. Writing fluency was not incorporated as a construct because the number of words was limited. Writing complexity was also not considered as a construct because the case study placed more importance on paraphrasing of the reading text than the use of a variety of syntactic structures. As mentioned in the second section of this article, the constructs of assessment should correspond to aspects of writing performance that are prioritized.

Next, criteria and levels of performance were defined based on the four constructs of the assessment. The case study focused on three criteria: (a) content and organization, (b) grammar, and (c) vocabulary. Constructs 1 and 2 were combined into criterion (a) in light of the simplicity of the rating scale. Similarly, the students' writing performance was graded at three levels, not at four or five levels commonly used in the rubrics outlined above.

Finally, descriptors and scoring points were given to the analytic rubric (see Table 2).[3] Some descriptors were based on Weigle's (2004) analytic rubric. Scores for the content and organization section were weighted twice as heavily as the other because the section contained two constructs of assessment.

Table 2

An Analytic Rubric for Integrated Writing Tasks

Criteria	Score	Description
Content Organization	6	The response includes the main idea of the reading text. The response is based on the text paraphrased in students' own words, not copied from the reading text.
	4	The response includes the main idea of the reading text. The response considerably relies on words and phrases of the reading text.
	2	The response is off the main idea of the reading text.
Grammar	3	The response includes few, or no, errors in grammar. Any errors are minor, such as agreement, number, and articles.
	2	The response includes a few errors in grammar. Some of them are major errors such as word order and may cause comprehension problems.
	1	The response includes many errors in grammar. The response includes few errors in grammar, but considerably relies on words and phrases of the reading text.
Vocabulary	3	The choices of words and phrases in the response are appropriate. There are few spelling errors.
	2	Some choices of words and phrases in the response are inappropriate and may cause comprehension problems.
	1	Many choices of words and phrases in the response are inappropriate and cause comprehension problems. The response considerably relies on words and phrases of the reading text.

Results and Discussion

The author first rated the 51 students' writing products according to the analytic rubric, and repeated the same procedure at an interval of three weeks. Three students were eliminated from the data analysis because the number of written words was too low. The descriptive statistics for the rating scale are shown in Tables 3 and 4.

Table 3

Descriptive Statistics for the Rating Scale in Class A (n = 25)

	Rating 1				Rating 2 (delayed)			
	M	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*
Content	4.32	1.70	2	6	3.84	1.72	2	6
Grammar	2.24	.44	2	3	2.52	.51	2	3
Vocabulary	2.60	.50	2	3	2.60	.50	2	3
Total	9.16	2.21	6	12	8.96	2.01	6	12

Table 4

Descriptive Statistics for the Rating Scale in Class B (n = 23)

	Rating 1				Rating 2 (delayed)			
	M	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*
Content	4.52	1.62	2	6	4.87	1.58	2	6
Grammar	2.13	.76	1	3	2.26	.75	1	3
Vocabulary	2.26	.81	1	3	2.39	.72	1	3
Total	7.89	4.08	4	12	8.42	3.81	6	12

Table 3 shows that the standard deviations for the content and organization section are larger than those for the other sections both in rating 1 and in rating 2. Table 4 also shows similar descriptive statistics. The results indicate that there may be differences in the ability to understand and paraphrase texts, although the weighted points of the contents and organization section affect the standard deviations. Some examples are shown below.

(3) Tamara Blakemore experienced culture trauma when she arrived at Boston College. She discovered firsthand there is a sea of difference.

(4) Blakemore, an Australian exchange student, was surprised to discover that real American lifestyle is different from which she thought before she went to America.

While the summary in example (3) relies on direct copying from the source, that in example (4) demonstrates that the writer described the situation in her own words. Therefore, the gap

between the two examples' scores was large (see Table 5).

Table 5

Examples of Using the Analytic Rubric

	The number of words	Scores				Comments
		Content	Grammar	Vocab.	Total	
Example (3)	20	4	1	1	6	Relying on direct copying from the reading text
Example (4)	24	6	3	3	12	Using the writers' own words

In order to test the reliability of the ratings, the simple agreement rates between the two ratings were calculated. In class A data, the simple agreement rate of the overall scores was 73.3%. The simple agreement rate of the scoring for the content and organization, grammar, and vocabulary sections was 72.0%, 64.0%, and 84.0%, respectively. In class B data, the simple agreement rate of the overall scores was 57.3%. The simple agreement rate of the scoring for the content and organization, grammar, and vocabulary sections was 48.0%, 60.0%, and 65.0%, respectively. Thus, the reliability of the ratings for class A was moderate, but that for class B was lower. In particular, the agreement rate indicates that the scale of content and organization section may lack reliability. Additionally, the consistency of scoring in the weighted section may be affected by differences between source texts.

Conclusion

This article attempted to explore how analytic rubric scales can help language teachers assess their students' performance on integrated writing tasks. In the case study, the original rubric for the assessment was created and tested in terms of its validity, practicality, and reliability.

Concerning validity, listing aspects which should be assessed in tasks is important. The case study in this article focused on four constructs to assess students' integrated skills of English. Therefore, although the validity of the rubric itself could be measured, the descriptors may be complicated in the content and organization section because these two constructs were combined into one criterion.

The rubric's practicality was somewhat attested though its simple criteria and score ranges.

However, its reliability was called into question in this limited case study. In order to further test and develop the rubric for wider use, more research will be needed in terms of both inter- and intra-rater reliability. Additionally, as only a limited number of reading materials were involved in this case study, analytic scores using other prompts for summary writing should be examined.

Despite these limitations, the attempt in the present study can be seen as a step towards improving assessment of learners' performance on integrated writing tasks. To refine the analytic rubric, components such as criteria and descriptors should be modified and validated through empirical validation data.

Acknowledgements

I would like to express my great appreciation for the constructive comments of an anonymous reviewer, which improved the quality of the manuscript.

Notes

[1] An earlier version of this paper was presented as part of a symposium at Expo-lingva edukado 2018, Tokyo, Japan, 4 March, 2018. I would like to thank Yukinari Shimoyama, Takane Yamaguchi, and Kazuya Kito for their help in the symposium.

[2] The writing task was based on the teacher employment examination conducted by the Gunma Prefectural Education Board in 2016.

[3] Although the number of words of the reading text and the summary in Class A was fewer than that of Class B, the same scoring framework was used in the analysis in terms of comparison between the two ratings.

References

Chan, S., Inoue, C., & Taylor, L. (2015). Developing rubrics to assess the reading-into-writing skills: A case study. *Assessing Writing*, *26,* 20-37. doi:10.1016/j.asw.2015.07.004

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

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency: Definitions, measurement and research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp.1-20). Amsterdam, the Netherlands: John Benjamins.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels* (NCTE Research Report No. 3). Champaign, IL: National Council of Teachers of English.

Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.

Kamimura, T. (2006). Effects of peer feedback on EFL student writers as different levels of English proficiency: A Japanese context. *TESL Canada Journal, 23*(2), 12-39. doi:10.18806/tesl.v23i2.53

Lee, L., & Gundersen, E. (2011). *Select readings intermediate: Teacher-approved readings for today's students* (2nd ed.). New York, NY: Oxford University Press.

Okubo, N. (2006). Shido to hyoka no ittaika wo mezashita shinraisei no takai eisakubun hyoka kijunhyo no sakusei: Tahenryo ippanka kanousei riron wo mochiite [Developing a reliable rubric to assess English composition in aiming at a closer connection between teaching and assessment: Using multivariate generalizability theory]. *STEP Bulletin, 18*, 14-29.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*, 510-532. doi:10.1093/applin/amp047

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

Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of

English. *Assessing Writing, 9*, 27-55. doi:10.1016/j.asw.2004.01.002

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy & complexity*. Honolulu, HI: University of Hawai'i Press.

Yamanishi, H. (2004). Kokosei no jiyu eisakubun wa donoyoni hyoka sarete irunoka: Bunsekiteki hyoka shakudo to sogoteki hyoka shakudo no hikaku wo toshite no kento [How are high school students' free compositions evaluated by teachers and teacher candidates?: A comparative analysis between analytic and holistic rating scales]. *JALT Journal, 26*, 189-205.

Zwier, L. J. (2012). *Inside reading 2: The academic word list in context* (2nd ed.). New York, NY: Oxford University Press.

Appendix A

A Reading Text for a Writing Task in Class A

Saying Tamara Blackmore experienced culture shock when she arrived here last September is an understatement. It was more like culture trauma for this adventurous student who left Melbourne's Monash University to spend her junior year at Boston College (BC). Blackmore, 20, was joined at BC by 50 other exchange students from around the world. Like the thousands of exchange students who enroll in American colleges each year, Blackmore discovered firsthand there is a sea of difference between reading about and experiencing America firsthand. She felt the difference as soon as she stepped off the plane.

(Lee & Gundersen, 2011, p.53)

Appendix B

An Example of a Student's Writing in Class B