English Language Teaching; Vol. 17, No. 12; 2024
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

# Developing an Analytic Rubric for Japanese EFL Learners' Summary Writing

Makiko Kato[1]

[1] Graduate School of Arts and Letters, Tohoku University, Sendai, Japan

Correspondence: Makiko Kato, Graduate School of Arts and Letters, Tohoku University, Auba-ku, Kawauti 27-1 980-8576 Sendai, Japan. E-mail: kato.s.makiko@gmail.com

Received: October 2, 2024 Accepted: November 4, 2024 Online Published: November 6, 2024

doi: 10.5539/elt.v17n12p1 URL: https://doi.org/10.5539/elt.v17n12p1

**Abstract**

This study aimed to develop and revise an analytic rubric based on the results of a survey of the difficulties experienced by raters when assessing the English summaries of Japanese English learners, as reported in a previous study (Kato, in press). In this study, three raters repeatedly discussed and established four categories: "Integration," "Language Use," "Paraphrasing," and "Content Accuracy." They created descriptors for each category from 0 to 5. The three raters evaluated the summaries of 20 Japanese university students for two types of original English texts. The results showed that the inter-rater reliability of Texts A and B was sufficiently high in all categories. Moreover, the correlation between categories was confirmed to measure independent constructs. This indicated that the analytic rubric created, based on Kato's (in press) survey report from the raters, is also useful in terms of reliability and validity. This study proposed an analytic rubric that can serve as a foundation for constructing a more user-friendly rubric for future raters.

**Keywords:** summary writing, analytic rubric, Japanese EFL learners, paraphrasing, copying

## 1. Introduction

### *1.1 Introduce the Problem*

English summarization, a task that integrates various skills and abilities, has attracted increasing attention in recent years as an academic literacy tool that requires mastery, which is difficult to acquire. Test in Practical English Proficiency (EIKEN), known as the most frequently used test of practical English skills among all generations in Japan, introduced an English summary in the writing task for Grades 2 and above from the 2024 academic year. Although the Ministry of Education, Culture, Sports, Science and Technology (MEXT) has reinforced integrated skills in English classes in secondary education for some time, English summarization has attracted more attention regarding practice and research. This study is part of a project to develop an analytic rubric for summary writing. An overview of the project is presented below.

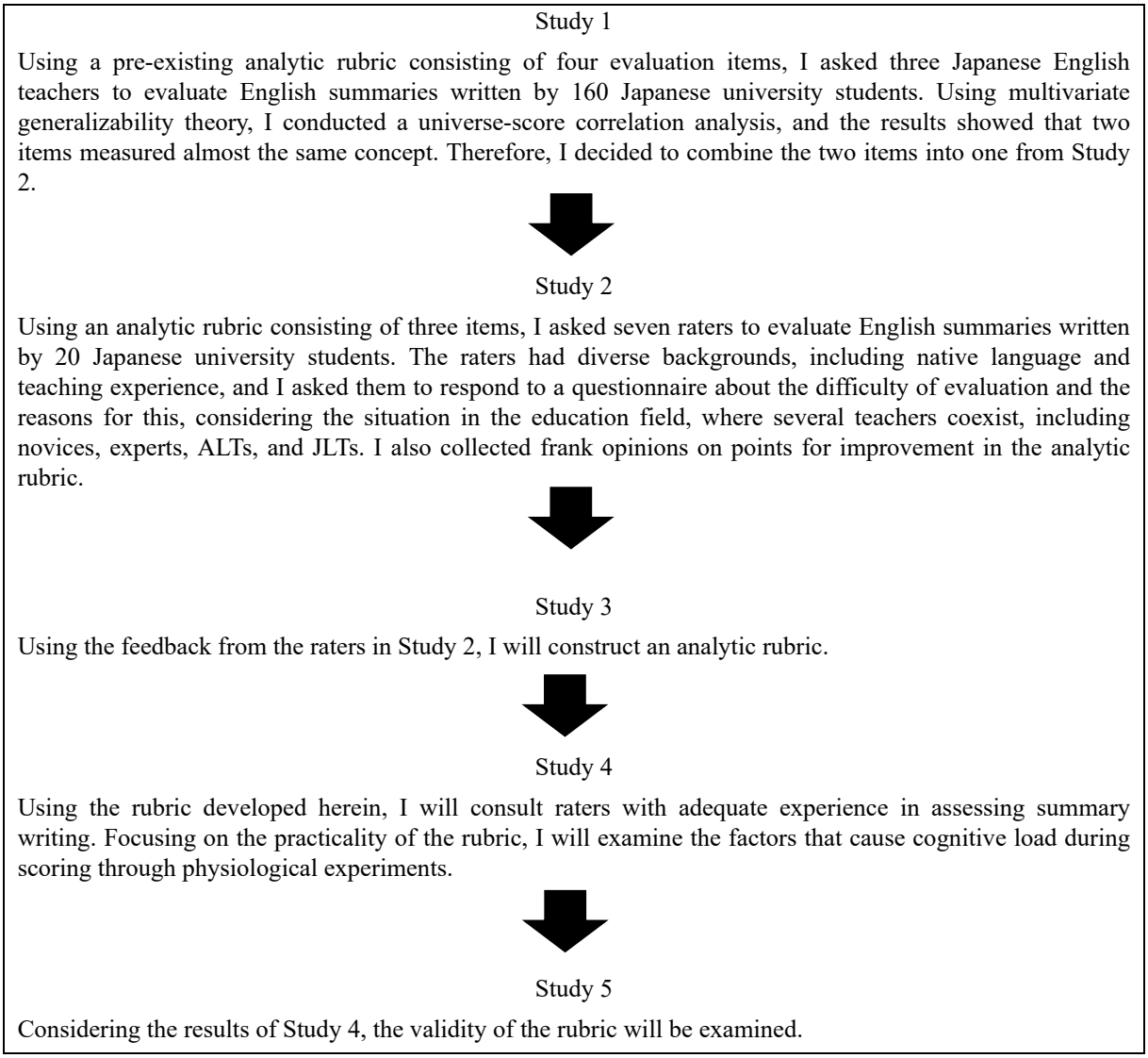


Figure 1. Outline of the Project for the Development of an Analytic Rubric

*1.2 Assessing Summary Writing*

As noted above, one factor that makes it difficult to acquire English summaries is a mixture of various skills and abilities. The acquisition and the complexity of the assessment make it difficult. According to Brown et al. (2005), assessing integrated tasks such as summary writing is more complex and demanding than traditional standalone or independent tasks. The reliability of an assessment has two perspectives: intra- and inter-rater reliability. In either case, to ensure a high level of reliability, sufficient discussions are desirable among raters during training so that the same performance is given the same score regardless of whether the raters are responsible for the evaluation. Another important factor is that the rubric should be well-understood and easy for raters to use.

Various researchers have developed holistic and analytic rubrics for English summaries (Chang et al., 2015; Kubota, 2023; Li, 2014; Sawaki, 2003, 2020; Yamanish et al., 2019; Yu, 2007). A holistic rubric is used in high-stakes tests such as the TOEFL iBT because it reduces the burden on raters regarding labor and time (Yamanishi et al., 2019). However, providing scores for each item is educationally useful for both learners and teachers of English summarization. Against this background, many studies (Currie, 1998; Cheng & Su, 2012; Chou, 2012; Choy & Lee, 2012; Friend, 2001; Johns & Mayes, 1990; Kato, 2018a, 2018b, 2018c, 2021a, 2021b; Kim et al., 2014; Pecorari, 2003; Shi, 2004) have employed analytic rubrics to examine the effectiveness of summary writing instruction for English learners, suggesting that analytic rubrics are in high demand in education.

Kato (2022) adopted Li's (2014) analytic rubric, which has already been validated and cited in many summarization studies on foreign language education, as an English summary assessment tool for Japanese university students. The first item in Li (2014) is the Main Idea Coverage (MIC), which assesses the selection of appropriate main ideas. The second is Integration (INT), which has two phases: whether the statements in a summary are written in logical order and have a global interpretation. The third category is "Language Use (LU)," which evaluates language use mainly through various syntax, grammar, and vocabulary. The last is Source Use (SU), which refers to the accuracy and coherent use of information in the text and the SU component also has two phases: whether the summary is written in the writer's own words and whether the information in the summary is included correctly. In Kato (2022), Multivariate generalizability theory (Brennan, 2001) examined the reliability and practicality of Li's (2014) rubric by investigating the ratings of three preservice English teachers. In the study that employed short English texts, one rater was sufficient to score MIC and INT, but three-four raters were needed to evaluate LU and SU. A strong positive correlation ($r$ = 0.94) was found between MIC and INT. However, a strong negative correlation ($r$ = -0.60) was found for LU and SU. Therefore, MIC and INT almost measure the same construct. Furthermore, beginning-to-intermediate English learners, like those in this study, copied texts when writing summaries and used synonyms or combined parts of texts to avoid copying. This study demonstrated the importance of separating paraphrased items from LU, as discussed in previous studies (Sawaki, 2020; Yamanishi et al., 2019). The study also concluded that the quality of rater training and detailed observation of decision-making for scoring to assess summary writing, especially for non-proficient learners of English, are required because misuse of language use and paraphrased characteristics can also be expected.

*1.3 Scoring Difficulty in Summary Writing and Candid Comments from the Raters*

Rater training is extremely important (Heidari et al., 2022; Lumley, 2002; Shohamy et al., 1992; Weigle, 1994; 1998). However, in actual educational settings, it is common for tests to be marked jointly by ALTs (assistants of language teachers) and JLTs (Japanese language teachers) and by expert and novice teachers (Kato, in press; Yamanishi et al., 2019). With teachers in the field of education swamped with miscellaneous tasks other than teaching, it is almost impossible to provide sufficient time for rater training. Against this background, in Kato (in press), seven raters with different native languages and teaching experience were given careful rater training and asked to evaluate English summaries written by 20 Japanese university students with English proficiency ranging from beginner to intermediate to advanced using an analytical evaluation scale consisting of three evaluation items. Moreover, I surveyed them using a questionnaire to determine the difficulty of evaluation and the reasons for this, as well as to obtain their opinions on how to improve the analytic rubric.

INT and MIC were measured in a previous study (Kato, 2022); while MIC was omitted, INT was retained. Perhaps one reason the INT descriptor was difficult to understand was that it did not reflect the coverage rate of the main ideas.

Concerning LU, the summaries written by students with low English ability often copied large sections of the text and this should not be assessed. It overlaps with whether the text has been copied or included in the SU information. Therefore, if the original texts were exactly copied, the information in the original text would be incorporated into the summary precisely as it is. However, a separate item should be included to evaluate the quality of the "Paraphrasing." Furthermore, one rater, a native English speaker with extensive experience in teaching English summary writing, highlighted that summary writers made various English errors, suggesting that the evaluation of LU should be broken into smaller categories, such as grammatical accuracy and vocabulary richness.

Regarding SU, it was not easy to judge whether copying had occurred. Therefore, most raters reported that they frequently moved their eyes between the three documents: source text, rubric, and summary. In addition, evaluating whether copying had taken place at SU was extremely emphasized. In this case, it was suggested that the frequency of copying should be expressed as a percentage rather than using adverbs such as "approximately" or "basically." Keck (2006, 2014) provided a detailed definition of the copying rate in summaries. She classified these into the following four types according to the degree to which the information contained in the original text was replicated or paraphrased in the summary: (a) Near-copy, (b) Minimal, (c) Moderate, and (d) Substantial Revisions. Furthermore, Keck uses the framework of Unique links (ULs) and General links (GLs) as criteria for determining the percentage of paraphrasing and defines UL as content words (nouns, adjectives, verbs, adverbs) appearing locally in the text. However, GL is defined as those that appear multiple times over a wide area in the main text. Each paraphrase is classified according to the percentage of the total number of words accounted for by the total number of words in the ULs. If the total number is 50% or more, it is classified as "Near Copy;" for 20-49% or more, it is classified as "Minimal Revision;" for 1-19%, it is classified as "Moderate Revision;" and if

it is not included at all (0%), it is classified as "Substantial Revision." However, some raters commented that while summary writers did not copy the source text but wrote it in their own words, there were often summaries that deviated from the text content or that did not seem to show an understanding of the text. It was also suggested that a detailed evaluation item be added to measure the "Content accuracy" in terms of understanding the text.

Some problems emerge from the perspective of raters who use an analytic rubric to evaluate summaries. Based on the issues reported by the raters, the purpose and research plan of this study are as follows:

## 2. Purpose of the Study

As explained in Figure 1, this study aims to construct an analytic rubric that incorporates the opinions of raters with diverse backgrounds found in actual educational settings and organizes the perspectives required for evaluation after grasping the characteristics of English summaries written by students with various English abilities. In this study, I revised the rubric by referring to the opinions of the raters gathered in a previous study by Kato (in press) and through discussions between the three raters.

## 3. Procedure

The following four steps in Vercellotti and McCormick (2021) were used as references for revising the rubric.

Table 1. Procedure of Constructing a Rubric

| Procedure of Constructing a Rubric |
|---|
| Step 1: Categories: |
| Identify categories that reflect separate skills of the stated learning objectives. |
| Step 2: Levels of performance: |
| Describe the expected levels of performance in each category appropriate for the context. |
| Step 3: Pre-Use Review: |
| Review the rubric for validity (categories are aligned with the assessment's stated learning objectives), reliability (performances can be consistently scored with the descriptions), practicality (a rubric is easy to use), and the beneficial consequences of using the rubric. |
| Step 4: Post-Use Evaluation: |
| Check that the scores are meaningful and based on the descriptions; the categories are independent, the descriptions are level-appropriate, and the rubric is easy to use. |

The main detailed process is divided into two phases: "Constructing a Rubric" and "Using and Confirming the Rubric," and is explained in the following subsections.

### *3.1 Participants*

Twenty Japanese first- and second-year university students majoring in education with diverse levels of English proficiency (Common European Framework of Reference for Languages [CEFR] A1 to C1) wrote summaries. Four participants were under CEFR levels in A1 and B2, five in A2 and B1, and two in C1. For the rating, the author, another Japanese rater, and a native English-speaking rater revised the rubric through discussions, whether they used the revised rubric to evaluate the English summaries created by the students.

### *3.2 Phase 1: Constructing a Rubric*

Twenty Japanese first- and second-year university students majoring in education with diverse levels of English.

#### 3.2.1 Construction Framework

Referring to the comments of the raters in Kato (in press), three raters revised the rubric through a three-hour discussion. Table 1 shows that the first step was to identify the categories reflecting the learning objectives. Thus, the definition of summary writing conveyed to student summary writers becomes a skill for learning objectives. Particularly, it is "selecting the necessary information, summarizing it in your own words within the word limit." Regarding LU, "it is desirable to be able to construct a sophisticated syntactic structure and choose vocabulary." Rubrics with many categories are difficult to use in practice because it is difficult to assess all the language skills listed in it, and burdensome for raters (Schreiber et al., 2012). Therefore, generally, four categories are the maximum that a rater can handle (Green, 2014; Popham, 1997; Vercellotti & McCor mick, 2021). Therefore, considering the burden on the raters, I set the number of categories to four: "Integration," "Language Use," "Paraphrasing," and "Content Accuracy."

As shown in Table 2, the next step is to describe the descriptors for each category set in the rubric according to performance level. The first is the number of levels. Previous studies (Brookhart, 2018; Suskie, 2009) generally state that three–five levels are appropriate; too many levels will be difficult to assess, time-consuming, and will affect the reliability of matching the performance of learners to the level that most closely resembles it. The rubric levels used in this study ranged from 0–5 points. Second, as was also mentioned in the comments of the raters in Kato (in press), using adverbs (almost, basically) in the descriptors leaves the scoring judgment up to the rater's subjectivity; therefore, using adverbs should be avoided. Vercellotti & McCormick (2021) also discuss the above. As it is impossible to make "few" estimates consistent between and within raters, it is best to avoid using quantifiers (few, some, many) to distinguish performance levels in English.

3.2.2 Integration

I developed an analytic rubric based on the number of categories and levels and the points to be careful about when writing descriptors. Regarding each item, referring to the opinions of the raters in Kato (in press), most raters thought it would be wise to retain the category "Integration." However, the descriptors need to be revised. In fact, when assessing summaries, it is important to select the appropriate number of main ideas, and the rubrics developed in previous studies (Li, 2014; Rivard, 2001; Sawaki, 2003) naturally include evaluation items related to the main ideas. Nevertheless, as reported in previous studies (Kato 2022), the categories "Integration" and "Main idea coverage" overlap and measure the same concept. Categories in an analytic rubric should be independent (Vercellotti & McCormick, 2021) and should assess different aspects of learners' skills (Popham, 1997; Van Moere, 2014). Therefore, the descriptors for "Integration" should also include a reference to the extent to which the main ideas are included in the summary. Moreover, regarding the main ideas to be selected, when using long English texts as sources, the idea units are divided in advance in rater training, and those to be included are chosen through discussions (Kato, 2021a; Kato, in press; Sawaki, 2003; 2020; Sawaki et al., 2024). In addition, to evaluate "Integration," it is necessary to evaluate whether the main ideas have been integrated and whether the information has been condensed and logically reconstructed. Considering these two perspectives, we decided on the following score distribution as a rough guide for the current discussion:

Essentially, if the summary contains 50% or more of the main idea, it receives four or five points. However, if there are problems with the chronological order or points that lack logic, it receives four points. If the information is logically condensed and reconstructed, it receives five points. Next, if the summary includes 20-50% of the main idea, it receives two or three points. However, if there are problems with the chronological order or points that lack logic, it receives two points. If the information is logically condensed and reconstructed, it receives three points. If the summary contained less than 20% of the main idea, it received 0 or 1 point because the information in the summary was insufficient. However, if the main idea were not included, the score would naturally be zero. During rater training, summaries that receive one or zero points are likely to have been read poorly; therefore, the score for "Content Accuracy" is also low.

3.2.3 Language Use

Regarding "LU," it is not logical to assign a high score to the LU score of a copied version. The condition under which the text is rewritten is a prerequisite for awarding a high score. Next, Vann et al.'s (1984) "Errors Listed in Order of Increasing Seriousness" is useful for considering the gravity of language errors. As shown in Table 2, the gravity of the language errors was classified into 12 categories. After discussing the examples of each error, we decided to classify them into four levels, from "no language errors at all" to "relatively many errors similar to (1)-(4)," "relatively many errors like (5)-(8)," and "relatively many errors similar to (9)-(12)," with the score decreasing from the top. If the rewriting was unclear, it was assigned a score of 1, and if it was almost completely blank, it was assigned 0.

Table 2. Errors Listed in Order of Increasing Seriousness (cited from Vann et al., 1984)

| Errors Listed in Order of Increasing Seriousness | |
|---|---|
| (1) Spelling-1 (spelling varieties that differ from standard American spelling) | Light |
| 1. These programmes (errors were not highlighted in the original questionnaire) are not acceptable. |  |
| 2. One example of osmosis is when the water goes thru the soil to the roots of the plants and from there to the leaves. | |
| (2) Articles | |
| 1. When two metals are combined, the new product is called _____ alloy. | |
| 2. At certain times of a year, geese are plentiful in Iowa. | |
| (3) Comma Splice (connecting two complete sentences with a comma) | |
| 1. The current was swift ___ he was unable to swim. | |
| 2. The official television in Venezuela is like Channel 11 in Iowa ___ programming. | |
| (4) Spelling-2 (errors such as deletion and substitution) | |
| 1. Osmosis is the process by wich a weak solution diffuses through a semi-permeable membrane to a stronger solution. | |
| 2. They learned a valueable lesson. | |
| (5) Prepositions | |
| 1. Inorganic acids are obtained in nonorganic matter. | |
| 2. An example for this process is when water goes into the plant's roots from the soil. | |
| (6) Pronoun Agreement | |
| 1. Each of us should say what they believe. | |
| 2. When one cannot swim, you fear the water. | |
| (7) Subject-Verb Agreement | |
| 1. The article is about an experiment that try to explain how osmosis happens. | |
| 2. Osmosis are the process in which water passes across a semi-permeable membrane. | |
| (8) Word Choice | |
| 1. Acids are qualified into two big groups. | |
| 2. The Nobel Prize winner this year was interesting in this subject. | |
| (9) Relative Clauses | |
| 1. In his free time he wrote his autobiography which he did not finish it. | |
| 2. Acids are divided into two groups: those that always contain the element carbon whose can be found in growing things and those that do not contain the element carbon. | |
| (10) Tense | |
| 1. Citric acid that is founded in lemons and oranges is an organic acid. | |
| 2. We have made distinguished two kinds of metals. | |
| (11) It-Deletion | |
| 1. ___ is necessary that a reservoir is higher than the town. |  |
| 2. When we combine zinc with copper ____ is called an alloy. | |
| (12) Word Order | |
| 1. Not only it allows people to travel cheaply, but to enjoy themselves as well. | |
| 2. Stops happening the process when concentrations on both sides are equal. | Serious |

3.2.4 Paraphrasing

Referring to the definition of copying introduced by Keck (2006), if the text was almost entirely in the student's own words, even if there were language errors, it was worth five points. If only 1-19% of the text is copied, even if there are some language errors, it is worth 4 points; 3 points if there is a language error but 20-49% of the text is copied; 2 points if 50-80% of the text is copied; 1 point if 80% or more of the text is copied; and 0 points if the entire text is copied or the answer sheet is left blank.

3.2.5 Content Accuracy

Finally, the evaluation of "Content Accuracy," which is also related to the level of understanding of the source text, should be affected by the quality LU and the selection of main ideas. If all the sentences in the summary are true to the content of the source text, it is worth 5 points; four points when 80% of the sentences are true to the content of the source text; three points if 50-80% of the sentences are true to the content of the source text; two points if 20-49% of the sentences are true to the content of the source text; one point if 20% of the sentences are true to the content of the source text; zero points if the summary is blank or the content of the summary is completely different from the content of the source text.

*3.3 Phase 2: Using and Confirming the Rubric*

3.3.1 Materials

(1) English Source Texts

The English levels of the 20 students (CEFR A1–C1) varied. However, considering that EIKEN Grade 2 is equivalent to the high school graduation level, we selected two source texts from previous questions for the long-text comprehension section of the pre-first grade (Appendix A, B). The number of words in each text and the ease of reading are presented as follows:

Table 3. Readability of the Source Texts

| | Source text A | Source text B |
|---|---|---|
| Flesch Reading Ease | 33.08 | 60.13 |
| Flesch-Kincaid Grade Level | 14.86 | 9.61 |
| Count of Words | 325 | 362 |
| Count of Paragraphs | 3 | 4 |
| Counts of Sentences | 13 | 19 |
| Average (Sentence per Paragraph) | 4.33 | 4.75 |
| Average (Words per Sentence) | 25.00 | 19.05 |

The number of words in both source texts is approximately the same; however, as indicated by the Flesch Reading Ease, the academic levels differ. In fact, Source Texts A and B are at the university level and the eighth to ninth grade (second to third year of junior high school) level, respectively, in English-speaking countries. Both texts concerned the environment, so the summary writing in this study was not expected to be difficult for the participants.

(2) Rubric

As mentioned previously, the rubric created through discussions among the three raters in this study is presented in Appendix C.

3.3.2 Procedure of Using and Checking the Rubric

The raters used the newly constructed rubric to evaluate participants' performances, referring to the Idea Unit Segmentation List created by Kato (in press). During the training of the seven raters (i.e., Kato, in press), the raters segmented all the sentences in the source text into idea units, according to the segmentation rules proposed by Kroll (1977). Each idea unit was divided into three levels: higher (containing the most important information), middle (containing important information), and lower (e.g., transition words and conjunctions). It was agreed that the idea units that were unanimously classified as "higher level" should be included in the summary. As shown in Step 3 in Table 1, the following methods will be used to verify the validity and reliability of the rubric in this study. Herein, I will not be discussing practicality.

Reliability refers to the property of whether the same measurement values can be obtained for the same measurement target. In other words, the inter-rater reliability in terms of consistency (i.e., the extent to which the

scores of the three raters agree) will be examined. For validity verification, I will check whether each category measures an independent construct by examining the correlation between categories.

Ideally, it would be beneficial to measure the correlation between peer and expert evaluations. However, as shown in Study 4 in Table 1, our rubric's final validity verification will be conducted after its practicality has been established through physiological experiments with the cooperation of more experienced raters. Thus, in this study, I will focus on reliability verification and omit detailed validity verification.

The three raters' inter-rater reliability for the English summaries of the two types of English texts was calculated using the Cronbach's alpha coefficient, as shown in Table 4.

Table 4. Results of Cronbach's-Alpha for Inter-Rater Reliability

| | Text A | Text B |
|---|---|---|
| Integration (INT) | 0.85 | 0.87 |
| Language Use (LU) | 0.83 | 0.85 |
| Paraphrasing (PARA) | 0.84 | 0.79 |
| Content Accuracy (CA) | 0.83 | 0.83 |

Because all items demonstrated a value of 0.7 or more, sufficient reliability was obtained. In other words, no differences existed in the evaluations of the three raters.

As shown in Table 3, inter-rater reliability between the texts was confirmed; therefore, in the correlation analysis, the category scores for the summaries of the two texts were combined to identify the correlation between the categories. The results of the Pearson's correlation analysis are shown in Table 5, and we observed no significant correlation between the categories. However, a slight correlation was observed between "Integration" and "Content Accuracy."

Table 5. Results of Pearson's Correlation Analysis for Each Category (Combined two texts)

| | Integration | Language Use | Paraphrasing | Content Accuracy |
|---|---|---|---|---|
| Integration | - | | | |
| Language Use | -.01 | - | | |
| Paraphrasing | -.12 | -.02 | - | |
| Content Accuracy | .27 | .07 | -.19 | - |

## 4. Discussion and Conclusion

The reliability verification results showed that there were no differences in the three raters' evaluations, and the correlation analysis results confirmed the independence of all categories; therefore, the raters did not evaluate or measure the same constructs. However, the correlation results demonstrated a weak positive correlation only between "Integration" and "Content Accuracy" ($r = .27$, $p = .09$). Simply put, "the more important the information (i.e., the main ideas) included by writers in their summary, the more accurate the content will be." However, as the writers of the summaries had a mixture of lower- and upper-intermediate English abilities, it became necessary to carefully consider whether the content of the English sentences they wrote in their own words made sense and whether it was faithful to the content of the text. Regarding the former, we found no correlation between "Paraphrasing" and "Language Use" ($r = -.02$, $p = .89$). This implies that just because the writers were writing in their own words did not necessarily indicate that they were using the correct language. As for the latter, we observed no correlation between "Paraphrasing" and "Content Accuracy" ($r = -.19$, $p = .23$). This indicates that the faithfulness of the English written in their own words to the content of the text cannot be determined. Finally, when we examined the relationship between the language used in the summary (i.e., "Language Use") and whether the English was faithful to the content of the text (i.e., "Content Accuracy"), there was no correlation ($r = .07$, $p = .65$). Therefore, just because the writers could accurately write English did not mean that they could determine whether it was true to the content of the text. As content accuracy was related to whether one could identify the main ideas, there was a correlation between "Content Accuracy" and "Integration."

However, to more comprehensively examine whether the writers use accurate language and whether they are writing in their own words, it is necessary to increase the number of samples and consider them according to their English ability. In this study, I developed an analytic rubric in stages for the summaries written by English learners. The categories and proposed descriptors were based on the opinions of raters reported in a previous

study by Kato (in press). Although the independence between the evaluation categories has been confirmed and the reliability among the three raters in this study was sufficient, the raters had engaged in revising the rubric and comprehended its content; therefore, the following three specific issues remain unaddressed. First, the results can be generalized to different raters, and the rubric can be applied to evaluate summaries of source texts at various levels. Second, feedback from students who used the proposed rubric in this study and experts are also necessary to examine its usefulness. Finally, more experts should be asked to use this analytic rubric and compare its usefulness with other analytic rubrics developed in EFL environments.

Nevertheless, this study clarifies the rationale for developing an analytic rubric and substantiates the research from both theoretical and practical perspectives. Furthermore, considering the introduction of the summary task in the EIKEN and the importance of summary instruction and evaluation in the field of education, the rubric was designed with feedback from raters with diverse backgrounds in actual educational settings. Thus, the rubric may serve as a foundation for further examination of the remaining issues.

**Acknowledgments**

This study was supported by JSPS KAKENHI Grant Number JP 22K13174. I would like to acknowledge two anonymous reviewers for their invaluable comments.

**References**

Brennan, R. L. (2001). *Generalizability theory*. Springer-Verlag, Inc. https://doi.org/10.1007/978-1-4757-3456-0

Brookhart, S. M. (2018). Learning is the primary source of coherence in assessment. *Educational Measurement: Issues and Practice, 37*(1), 35-38. https://doi.org/10.1111/emip.12190

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for Academic-Purposes speaking tasks (TOEFL Monographs Series # MS29)*. Educational Testing Services. https://doi.org/10.1002/j.2333-8504.2005.tb01982.x

Currie, P. (1998). Staying out of trouble: Apparent plagiarism and academic survival. *Journal of Second Language Writing, 1*(7), 1-18. https://doi.org/10.1016/S1060-3743(98)90003-0

Chen, Y. S., & Su, S. W. (2012). A genre-based approach to teaching EFL summary writing. *ELT Journal, 66*(2), 184-192. https://doi.org/10.1093/elt/ccr061

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

Chou, M. (2012). Implementing keyword and question generation approaches in teaching EFL summary writing. *English Language Teaching, 5*(12), 36-41. https://doi.org/10.5539/elt.v5n12p36

Choy, S., & Lee, M. (2012). Effects of teaching paraphrasing skills to students learning summary writing in ESL. *Journal of Teaching and Learning, 8*(2), 77-89. https://doi.org/10.22329/jtl.v8i2.3145

Friend, R. (2001). Teaching summarization as a content area reading strategy. *Journal of Adolescent & Adult Literacy, 44*(4), 320-329.

Green, A. (2014). *Exploring language assessment and testing*. Routledge. https://doi.org/10.4324/9781315889627

Heidari, N., Ghanbari, N., & Abbasi, A. (2022). Raters' perceptions of rating scales criteria and its effect on the process and outcome of their rating. *Language Testing in Asia, 12*. https://doi.org/10.1186/s40468-022-00168-3

Johns, A., & Mayes, P. (1990). Analysis of summary protocols of university ESL students. *Applied Linguistics, 11*(3), 253-271. https://doi.org/10.1093/applin/11.3.253

Kato, M. (2018a). Exploring the Transfer Relationship of Summarizing Skills in L1 and L2. *English Language Teaching, 11*(10), 75-87. https://doi.org/10.5539/elt.v11n10p75

Kato, M. (2018b). Providing Comprehension Clues in L1 to Japanese EFL Summary Writers: Do they help? *International Journal of Applied Linguistics and English Literature, 7*(5), 12-21. https://doi.org/10.7575/aiac.ijalel.v.7n.5p.12

Kato, M. (2018c). Good and Poor Summary Writers' Strategies: The Case of Japanese High School EFL Learners. *Journal of Language Teaching and Research,* 9(6), 1199-1208. https://doi.org/10.17507/jltr.0906.09

Kato, M. (2021a). *Examining the Effects of Explicit and Implicit Instructions on Summary Writing for Japanese University Students Learning English as a Foreign Language with Low English Language Proficiency* [Unpublished doctoral dissertation]. Sophia University.

Kato, M. (2021b). Summarization in English as a Foreign Language: A Study Comparing L2 Summary Performances to Summarizer's L2 Vocabulary Size and L1 Summarizing Skill. *English Language Teaching, 14*(5), 77-88. https://doi.org/10.5539/elt.v14n5p77

Kato, M. (2022). Examining the Dependability and Practicality of Analytic Rubric of Summary Writing Using Multivariate Generalizability Theory: Focusing on Japanese University Students with Lower-Intermediate Proficiency in English. *English Language Teaching, 15*(9), 82-94. https://doi.org/10.5539/elt.v15n9p82

Kato, M. (in press). Scoring Difficulty in Assessing Summary Writing: Toward the Reconstruction of Analytic Rubric. *Journal of Education and Learning, 14*(2). https://doi.org/10.5539/jel.v14n2p74

Keck, C. (2006). The use of paraphrase in summary writing: A comparison of L1 and L2 writers. *Journal of Second Language Writing, 15*(4), 261-278. https://doi.org/10.1016/j.jslw.2006.09.006

Keck, C. (2014). Copying, paraphrasing, and academic writing development: A re-examination of L1 and L2 summarization practices. *Journal of Second Language Writing, 25*, 4-22. https://doi.org/10.1016/j.jslw.2014.05.005

Kim, M., William, J. C., & Jindarat, D.V. (2014). Summary writing in a Thai EFL university context. *Journal of Second Language Writing, 24*, 20-32. https://doi.org/10.1016/j.jslw.2014.03.001

Kubota, K. (2023). Developing and Revising a Rating Scale for Reading-to-Write Tasks in Classroom Assessments. *TELES Journal, 43*, 43-54. https://doi.org/10.57539/telesjournal.43.0_43

Kroll, B. (1977). Combining ideas in written and spoken English: a look at subordination and coordination. In E. O. Keenan & T. L. Bennett (Eds.), *Discourse across time and space* (pp. 69-108). University of Southern California.

Li, J. (2014). Examining genre effects on test taker's summary writing performance. *Assessing Writing, 22*, 75-90. https://doi.org/10.1016/j.asw.2014.08.003

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246-276. https://doi.org/10.1191/0265532202lt230oa

Pecorari, D. (2003). Good and original: Plagiarism and patchwriting in academic second-language writing. *Journal of Second Language Writing, 12*(4), 317-345. https://doi.org/10.1016/j.jslw.2003.08.004

Popham, W. J. (1997). What's wrong–and what's right–with rubrics. *Educational Leadership, 55*(2), 72-75.

Rivard, L. P. (2001). Summary writing: A multi-grade study of French-immersion and francophone secondary students. *Language, Culture and Curriculum, 12*(2), 171-186. https://doi.org/10.1080/07908310108666620

Sawaki, Y. (2003). *A comparison of summarization and free recall as reading comprehension tasks in web-based assessment of Japanese as a foreign language* [Unpublished doctoral dissertation]. University of California, Los Angeles.

Sawaki, Y. (2020). Developing Summary Content Scoring Criteria for University L2 Writing Instruction in Japan. In G. J. Ockey and B. A. Green (Eds.), *Another Generation of Fundamental Considerations in Language Assessment*. (pp. 153-171). Springer. https://doi.org/10.1007/978-981-15-8952-2_10

Sawaki, Y., Ishii, Y., Yamada, H., & Tokunaga, T. (2024). Developing and validating an online module for formative assessment of summary writing with automated content feedback for EFL academic writing instruction. *Language Testing in Asia, 14*(50). https://doi.org/10.1186/s40468-024-00325-w

Schreiber, L. M., Paul, G. D., & Shibley, L. R. (2012). The development and test of a public speaking competence rubric. *Communication Education, 61*(3), 205-233. https://doi.org/10.1080/03634523.2012.670709

Shi, L. (2004). Textual borrowing in second-language writing. *Written Communication, 21*, 171-200. https://doi.org/10.1177/0741088303262846

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal, 76*(1), 27-33. https://doi.org/10.1111/j.1540-4781.1992.tb02574.x

Suskie, L. (2009). *Assessing student learning: A common-sense guide.* Jossey-Bass.

Vann, R. J., Meyer, D. E., & Lorenz, F. O. (1984). Error gravity: A study of faculty opinion of ESL errors. *TESOL Quarterly, 18*(3), 427-440. https://doi.org/10.2307/3586713

Vercellotti, L. Mary., & McCormick, E. Dawn. (2021). Constructing Analytic Rubrics for Assessing Open-Ended Tasks in the Language Classroom. *TESL-EJ, 24*(4).

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197-223. https://doi.org/10.1177/026553229401100206

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287. https://doi.org/10.1177/026553229801500205

Yamanishi, H., Ono, M., & Hijikata, Y. (2019). Developing a scoring rubric for L2 summary writing: a hybrid approach combining analytic and holistic assessment. *Language Testing in Asia, 9*(13). https://doi.org/10.1186/s40468-019-0087-6

Yu, G. (2007). Students' voices in the evaluation of their written summaries: Empowerment and democracy for test takers? *Language Testing, 24*(4), 539–572.

**Appendix A**

**English Source Text A**

| New York City vs. the Car |
| --- |
| Over the years, concerns about heavy traffic and the effect of cars on air quality have led New York City officials to make various attempts to reduce traffic on the island of Manhattan. The first was a 1971 proposal by then mayor John Lindsay to ban automobiles from entering a crowded area of Manhattan's business district during the daytime. The plan, promoted as a solution to the air pollution problem, was widely applauded by citizens.…… |

**Appendix B**

**English Source Text B**

| In the Air |
| --- |
| In many of the world's large cities, air pollution is a serious problem. According to experts, about 7 million people around the world die every year from the effects of dirty air. London, the capital city of the United Kingdom, is one example of a city with this problem. London is said to have some of the worst levels of air pollution in the world. …….. |

**Appendix C**

**Analytic Rubric for EFL Learners' Summary Writing**

| | Integration | Language Use | Paraphrasing | Content accuracy |
|---|---|---|---|---|
| 5pt | The main idea is covered by over 50%, and the information is logically condensed and reconstructed. | Sentences in the summary have been rewritten with (almost) no language errors. | Even with language errors, the summary is written almost entirely in your own words. | All the sentences in the summary are true to the content of the text. |
| 4pt | The summary contains over 50% of the main idea, but there are points where the chronological order is incorrect, or the logic is lacking. | Sentences in the summary have been rewritten, but the language errors are limited to British English spelling mistakes, lack of articles, or punctuation. | Even with language errors, only 1-19% of the sentences in the summary are copied. | Around 80% of the sentences in the summary are true to the content of the text. |
| 3pt | The information is logically condensed and reconstructed, including the main idea of 20-50%. | Sentences in the summary have been rewritten, but there are language errors in prepositions, pronoun agreement, subject-verb agreement, or vocabulary selection. | Even with language errors, only 20-49% of the sentences in the summary are copied. | Approximately 50-80% of the sentences in the summary are true to the content of the text. |
| 2pt | The summary contains the main idea to a degree of 20-50%, but there are points where the chronological order is incorrect, or the logic is lacking. | Sentences in the summary have been rewritten, but the language errors are related to relative pronouns, tense, or omission. | Approximately 50% - 80% of the sentences in the summary are copied. | Around 20-49% of the sentences in the summary are true to the content of the text. |
| 1pt | The summary contains less than 20% of the main idea. | The summary has almost no rewrites. | Over 80% of the sentences in the summary have been copied. | Around 20% of the sentences in the summary are true to the content of the text. |
| 0pt | The summary does not contain the main idea. | The summary is almost blank. | The sentences in the summary are entirely copied, or the summary is blank. | The summary is blank, or the content of the summary is completely different from that of the text. |