Language Testing in Asia

RESEARCH Open Access

Developing a scoring rubric for L2 summary writing: a hybrid approach combining analytic and holistic assessment

Hiroyuki Yamanishi[1*], Masumi Ono[2] and Yuko Hijikata[3]

* Correspondence: hiyamani@kc.chuo-u.ac.jp
[1]Faculty of Science and Engineering, Chuo University, Tokyo, Japan
Full list of author information is available at the end of the article

Abstract

Background: In our research project, we have developed a scoring rubric for a second language (L2) summary writing for English as a foreign language (EFL) students in Japanese universities. This study aimed to examine the applicability of our five-dimensional rubric, which features both analytic and holistic assessments, to classrooms in the EFL context. The examination especially focused on a newly added, optional overall quality dimension and two paraphrasing dimensions: paraphrase (quantity) and paraphrase (quality).

Methods: Six teacher raters evaluated 16 summaries written by Japanese EFL university student writers using our new rubric. The scoring results were quantitatively compared with the scoring results of a commonly used rubric developed by the Educational Testing Service (ETS). For the qualitative examination, the teacher raters' retrospective comments on our five-dimensional rubric were analyzed.

Results: The quantitative examination demonstrated positive results as follows: (a) adding the optional overall quality dimension improved the reliability of our rubric, (b) the overall quality dimension worked well even if it was used alone, (c) our rubric and the ETS holistic rubric overlapped moderately as L2 summary writing assessments, and (d) the two paraphrasing dimensions covered similar but different aspects of paraphrasing. However, the quantitative analysis using the generalizability theory (G theory) simulated that the reliability (generalizability coefficients) of the rubric was not high when the number of raters decreased. The qualitative examination of the raters' perceptions of our rubric generally confirmed the quantitative results.

Conclusion: This study confirmed the applicability of our rubric to EFL classrooms. This new type of rating scale can be characterized as a "hybrid" approach that offers the user a choice of analytic and holistic measures depending on individual purposes, which can enhance teachers' explicit instructions.

Keywords: L2 summary writing, Rubric, Analytic assessment, Holistic assessment, Japanese university students, Generalizability theory, Rater perception, Paraphrasing

Introduction

The importance of summary writing skills is widely acknowledged (e.g., Delaney, 2008; Hidi & Anderson, 1986; Plakans, 2008) because these skills are essential at every level of education. In particular, these skills are crucial for university students because they are often required "to write summaries as stand-alone assignments" (Marshall, 2017, p. 71) or complete other types of assignments that incorporate various kinds of sources into their writing. Summary writing is an integrated writing task, and it is understood as a reading-to-write task. In other words, this involves a series of intricate processes of comprehension, condensation, and production (Kintsch & van Dijk, 1978), which is regarded as "a highly complex, recursive reading-writing activity involving constraints that can impose an overwhelming cognitive load on students" (Kirkland & Saunders, 1991, p. 105). Due to the complexity and higher-order treatments required for summary writing, student writers are likely to encounter difficulties producing effective summaries and developing summary writing skills on their own (Grabe, 2001), and this issue exists in the English language education context. Therefore, scaffolding students' learning processes and providing clear instructions for summarization skills are also necessary in English as a foreign language (EFL) classrooms.

However, language teachers often encounter difficulties in teaching second language (L2) summary writing because of the multidimensional nature of this genre (Yu, 2013) and the limited number of appropriate, practical teaching materials and guidelines such as Marshall (2017) and Oshima, Hogue, and Ravitch (2014). In fact, there is a tendency of insufficient instruction in L2 summary writing in Japanese and Taiwanese educational contexts, which results in self-taught summarization skills (Ono, 2011). Ono (2011) study indicates this problematic situation in EFL writing education that can also be observed in other EFL contexts. Thus, to improve such situations, it is worth developing teaching guidelines for teachers to facilitate and assess L2 summary writing in classrooms. In this study, we focus on the development of tools that can be utilized in L2 summary writing instruction in an EFL classroom context with a particular focus on the assessment of summary writing by Japanese EFL university students.

Paraphrasing and textual borrowing in integrated writing

Integrated writing, including summary writing, is composed of a number of different sub-skills such as reading and writing, depending on task types. Among them, one central skill relevant to summary writing is paraphrasing, which is also often used in academic writing in general. According to Hirvela and Du (2013), summarizing and paraphrasing require different levels of condensation of information. Previous studies indicated that paraphrasing serves a crucial role in summary writing (e.g., Shi, 2012). For example, Johns and Mayes (1990) investigated summary writing operations used by English as a second language (ESL) university students where the performances of high- and low-proficiency students were compared. They demonstrated that the low-proficiency group copied information from the original text more frequently than the high-proficiency group and that students in both groups neither combined information across paragraphs nor invented topic sentences by using their own words. Similarly, Keck (2006, 2014) reported that L2 writers in a US university tended to struggle with paraphrasing by employing insufficient paraphrasing, so-called *Near Copy*, when compared to native speakers of

English who employed effective paraphrasing to a greater degree, namely, *Moderate Revision* and *Substantial Revision*. This copying behavior has also been investigated through studies focusing on textual borrowing. For example, Gebril and Plakans (2016) examined how textual borrowing affects lexical diversity when learners borrow words from sources in integrated reading-based writing tasks. These analyses revealed that borrowing words from the source materials determined the writers' lexical diversity and that lexical diversity significantly differed across the writing scores.

From a sociocultural perspective, paraphrasing behavior may be affected by cultural and linguistic differences. Whether paraphrasing and textual borrowing are influenced by cultural backgrounds was investigated when Chinese graduate students read research papers and paraphrased them (Shi & Dong, 2018). Their results showed that textual borrowing was used more in Chinese, which was the participants' first language (L1), than in English, their L2. Shi and Dong argue that paraphrasing practices have cultural differences in that some paraphrasing practices might be acceptable in Chinese writing but not in English writing. Regarding linguistic differences from English, the Japanese language, which is L1 in our research context, has many differences in terms of orthography, sentence structure, and semantics. In particular, how to replace a certain word with its umbrella term or synonym can be influenced by the learners' L1. Another issue related to paraphrasing is the phenomenon of *patchwriting*, which "is unacceptable paraphrasing, a type of plagiarism" (Marshall, 2017, p. 65). "Patchwriting occurs when a writer copies text from a source and changes only some of the words and grammar" (p. 65). This often happens among novice writers and is seen as a developmental phase of paraphrasing attempts (Pecorari, 2003). Hence, teachers need to be aware that this inadequate manner of paraphrasing occurs regardless of the students' intention and that it takes time until students fully understand and become accustomed to paraphrasing.

Thus, paraphrasing plays a vital role in summary writing; however, this skill is considerably difficult to master and teach due to its complex characteristics and the influence of writers' cultural and linguistic backgrounds.

Assessing L2 summaries holistically and analytically

Teachers face challenges not only in teaching summary writing but also in assessing student summaries. Difficulties in the assessment of summaries arise for many reasons such as difficulty in identifying the main ideas (Alderson, 2000), the intricate operations employed in the summarizing process, and insufficient scoring guidelines for educational purposes. In classroom settings, measures for student writing vary depending on the context (Hamp-Lyons, 1995). Although portfolio-based assessments are favored in some contexts (Black, Daiker, Sommers, & Stygall, 1994), assessments of student writers' summaries usually employ scoring *rubrics*, which are scoring guidelines for different criteria. As Knoch (2009) points out, rubrics or rating scales serve a central role in evaluating integrated tasks, including L2 summary writing. Furthermore, the assessment of written L2 summaries has been a central concern among research in the fields of language testing and writing because of the increasing interest in integrated writing tasks along with task authenticity (Plakans, 2010, 2015; Weigle, 2004) and the importance of integrated skills in educational settings.

In performance assessment (e.g., L2 writing assessment), scoring rubrics are generally divided into two types: holistic rubrics and analytic rubrics (Bacha, 2001; Hamp-Lyons, 1995; Hyland, 2003; Weigle, 2002). Holistic assessments provide only an overall score for the performance (Hyland, 2002; Weigle, 2002) and are often used for large-scale assessments such as placement tests or high-stakes examinations. Advantages of holistic assessments are its practicality and cost-effectiveness, as it takes less time for raters to complete the assessment, thereby reducing labor costs, compared to analytic assessments (Bacha, 2001; Hamp-Lyons, 1995; Hyland, 2003; Weigle, 2002). However, a disadvantage of holistic assessments is that they cannot provide informative feedback on the performance, which neither helps teachers identify the weaknesses and strengths of individual students' performance nor provides constructive feedback on the students' performance.

By contrast, analytic assessments require more time for raters to complete the assessments, thereby increasing the labor costs, than holistic assessments because analytic rubrics have several dimensions related to the aspects of tasks or tests assigned. Multiple dimensions in analytic rubrics have descriptors where raters evaluate each dimension and choose a score for each of the dimensions based on the descriptors. This characteristic of analytic assessments enables teacher raters to provide diagnostic and comprehensive feedback on the students' performance and allows them to identify the strengths and weaknesses of individual performance (Hamp-Lyons, 1995) as well as the student writers' learning needs. In summary, holistic scoring rubrics should be chosen if only an overall, summative score of the performance is needed, whereas analytic scoring rubrics are more suitable if both a score for individual aspects of the performance and informative feedback are necessary (Mertler, 2001; Stevens & Levi, 2013). Therefore, analytic rubrics are often preferred in classroom contexts and used for educational purposes rather than testing purposes.

One of the well-known rubrics for L2 summaries is a holistic one developed by the Educational Testing Service (ETS) (Educational Testing Service, 2002). This rubric was one of the pilot rubrics examined in the process of developing the Test of English as a Foreign Language Internet-based Test (TOEFL iBT®). The feature of this rubric is that test takers receive an overall score, ranging from one to five. This means that the descriptors under each score contain various subskills (e.g., organization, sentence formation, use of own language, and language from source text) related to L2 summary writing. This holistic rubric has been used for the evaluation of L2 summaries across countries, not only in classrooms but also for research (e.g., Cumming et al., 2005). For instance, Baba (2009) utilized the ETS holistic rubric when examining Japanese university students' L2 summary writing performance in an EFL context. In line with Baba's study, the ETS rubric is also used in our current study to assess L2 student summaries in Japanese contexts.

Developing scoring rubrics for L2 summary writing

Apart from the ETS rubrics, the development of locally contextualized rubrics for L2 summary writing is becoming important and popular among researchers. This section discusses the features of such studies with a focus on the features of rubrics and individual contexts. In the US university context, Becker (2016) examined the effects of holistic scoring rubrics on student performance by comparing four ESL student groups: (a) those who developed a scoring rubric immediately after they completed the summary writing task, (b) those who used the rubric to score their classmates' products for the summary task, (c) those who only looked at the rubric before they completed the summary task, and (d) the control group. As the first two groups outperformed the latter two groups, it was concluded that involvement in the process of rubric development and/or application is more effective than just reviewing the same rubric. Thus, Becker's study sheds light on the pedagogical value of students developing locally contextualized rubrics, as it can help to improve their summarizing performance. It is, however, noteworthy that, although the rubric used in Becker's study was a 5-point holistic rating scale, different teaching and research contexts may require a different type of rubric.

In the EFL university context, for example, Yu (2007) developed a holistic rubric to evaluate the overall quality of summaries written in Chinese (L1) and English (L2) in a Chinese context. This holistic rubric used "an argumentation method (e.g., D+, D, and D−) to assign scores for each summary" (p. 567), ranging from A+ to F−. In this scoring method, if the score difference between two raters was greater than 3 (e.g., C and B+), a third rater would score the same summary, and if the difference of the three scores was still greater than 3, all three raters would negotiate to assign an agreed score (see Yu, 2007, for details). Although this scoring rubric was holistic in nature, it demonstrated the following four general guidelines: "faithfulness of the source text," "summary and source text relationships scores," "conciseness and coherence," and "rater understanding" (p. 568). In a different Chinese context, an analytic rubric was developed by Li (2014) to investigate the effects of source text genres on summarization performance and the perceptions of student writers. This four-component analytic rubric consisted of "Main Idea Coverage," "Integration," "Language Use," and "Source Use" (p. 79) on a 6-point scale (i.e., 0–5 for each component). Interestingly, the analyses demonstrated contradictory results as students performed better in the expository text summarization compared to the narrative text summarization, while their perceptions indicated that narrative texts were easier to summarize than expository texts.

Unlike the development of these rubrics in Chinese contexts, to our knowledge, only a few studies have developed analytic rubrics for L2 summary writing in Japanese EFL university contexts (e.g., Sawaki, 2019). Although some may question the need for an analytic rubric specifically for the Japanese EFL context, we believe that it is important to meet language teachers' needs and reflect their beliefs as well as the educational policies of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) in Japan on the teaching of integrated language tasks such as L2 summary writing in secondary education (MEXT, 2018). Currently, language teachers in secondary and tertiary education are expected to teach integrated language skills and place emphasis on the integration of skills in language tests. Furthermore, in classroom settings, evaluating learners' paraphrasing skills and providing feedback on how to improve them cannot be accomplished by using a holistic rubric because of the complexity of such skills. Therefore, developing a scoring rubric to assess L2 summaries with a focus on paraphrasing is important for the effective teaching and assessment of integrated skills.

Our project

Based on the context described above, we initiated a research project on the development of rubrics that can improve the teaching and assessment of L2 summary writing. More specifically, the project aims to develop a rubric as a support tool for both teacher raters and student writers to foster their learning, teaching, and assessment of L2 summary writing.

The present study is part of a larger research project, which consists of a series of studies, and Fig. 1 outlines the organization of our research project. As presented in Fig. 1, our project began by examining the ETS (2002) holistic scoring rubric from the perspectives of reliability and validity in study I (Hijikata, Yamanishi, & Ono, 2011). We used the ETS holistic rubric in Japanese EFL university classrooms and found that the use of this holistic rubric lacked diagnostic, detailed, and comprehensible feedback for the students. From the results of study I, we concluded that the rubric was not completely appropriate for Japanese EFL classrooms, and this motivated us to develop a new rubric which informs what a holistic rubric cannot provide. Therefore, our subsequent investigation attempted to develop a rubric to fill this need.

In study II (Hijikata-Someya, Ono, & Yamanishi, 2015), we categorized the reflections of six EFL teacher raters who graded summaries produced by Japanese EFL university students. The difficulties in grading were qualitatively analyzed and coded into content, organization, vocabulary, language, mechanics, paraphrasing, and length. Their reflective comments were used to constitute our provisional rubric with four dimensions: content, paraphrase (quantity), paraphrase (quality), and language use. One way in which this provisional rubric was innovative was the emphasis on the aspect of paraphrasing.

Study III (Yamanishi & Ono, 2018) built on studies I and II and covered the development and refinement of the provisional rubric using the "expert judgment" of three experts in the field of language testing research. Subsequently, the revised rubric contained five dimensions (Appendix 1) because an optional overall quality dimension, which is holistic in nature, was added to the provisional rubric. This addition was based on a suggestion from the experts so that teachers can evaluate (a) the quality of the summary itself from a holistic point of view and (b) whether the summary corresponds to the requirements of the summary writing task. Finally, the current study (study IV) aims to

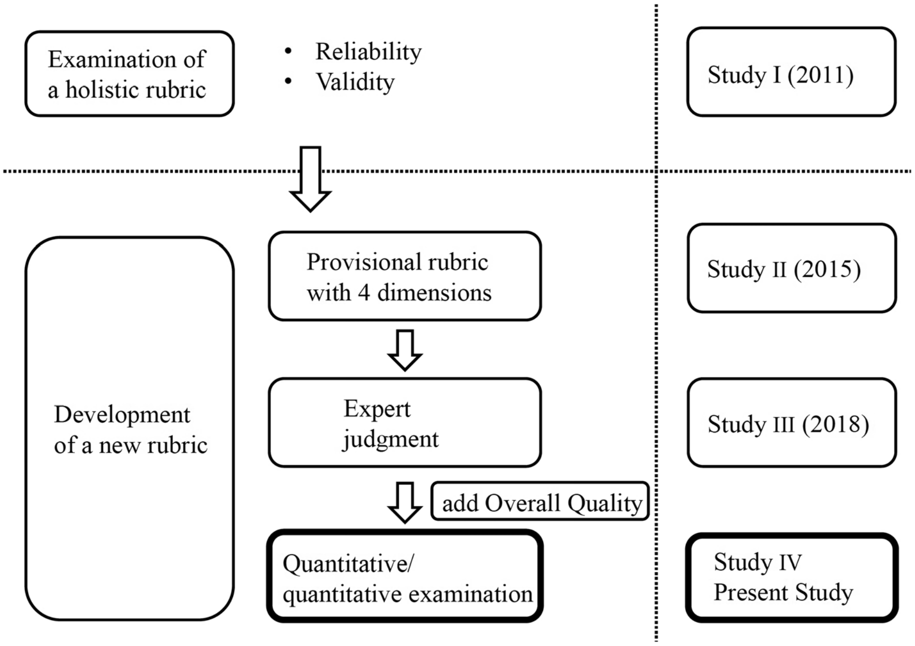
Examination of a holistic rubric
• Reliability
• Validity
Study I (2011)
Development of a new rubric
Provisional rubric with 4 dimensions
Study II (2015)
Expert judgment
Study III (2018)
add Overall Quality
Quantitative/ quantitative examination
Study IV Present Study

Fig. 1 Organization of our research project

examine the applicability of our (revised) rubric to EFL classrooms using both quantitative and qualitative methods.

Purpose of the present study

Our research project aimed to develop a new scoring rubric for L2 summaries in the Japanese EFL context to enhance the learning, teaching, and assessment of L2 summary writing in the classroom, with the following features:

1. The rubric can be used as a scoring scale and as a teaching and achievement guideline for L2 summary-writing instruction;
2. The rubric can be used both analytically and/or holistically, depending on the purpose of the assessment;
3. The rubric should be suitable for the Japanese EFL context, so that it can be used by both native English-speaking (NES) teacher raters and Japanese non-native English-speaking (NNES) teacher raters; and
4. The rubric should not be time-consuming to use, and thus, it consists of simple and concise descriptors.

As Fig. 1 illustrates, in our previous studies (study I through III), we developed a five-dimensional rubric (see Appendix 1)—content, paraphrase (quantity), paraphrase (quality), language use, and overall quality—that includes the features above. For example, for the first feature, our rubric contains two paraphrasing dimensions that intend to guide student writers on how to paraphrase. For the second feature, as a result of expert judgment in study III, we added an optional overall quality dimension that can be used holistically with the descriptor, "As a response to this task, the overall quality of this summary is. . . ." For the third feature, our rubric is written in English and Japanese, so that both NES and Japanese NNES teacher raters, and potentially students, can use it with ease. For the fourth feature, through study III, we simplified the wording of the descriptors of each dimension.

However, the newly added overall quality dimension and the two paraphrasing dimensions have not undergone any validation processes. Thus, the purpose of the present study (study IV) is to conduct a quantitative and qualitative examination of our rubric, especially on the potential of these newly added dimensions.

Methods

Participants

Fifty-one Japanese EFL students at two private universities (universities A and B) in Japan participated in this study. Students from university A were management majors, while students from university B specialized in various fields related to the English language. The students in both groups have studied English for more than 6 years, both before and during university. Both groups of students enrolled in general English language courses for new students in their respective universities. The students' English proficiency level ranged from intermediate to lower-intermediate, which is equivalent to levels B1–A2 in the Common European Framework of Reference for Languages: students from university A had an average TOEIC-IP® score of 532.1 (*SD* = 117.4), while students from university B had an average TOEFL-ITP® score of 420.7 (*SD* = 31.6). Given that there are few differences in the participants' characteristics between the two universities, the participant data were analyzed together without distinguishing between the two groups.

Three NES teachers and three NNES teachers also participated in this study to assess the summaries as raters. NES and Japanese NNES raters were recruited for this study because both types of teachers often teach English courses in EFL contexts in Japan. The six raters varied in terms of educational background and teaching expertise, which reflected the various university contexts, where both novice and experienced teachers are involved in tertiary education. Three NES raters were recruited from the Department of Language and Linguistics of a university in the UK:

- NES 1 (a Ph.D. student and part-time lecturer, 19 years of teaching experience),
- NES 2 (a Ph.D. student and part-time lecturer, 15 years of teaching experience), and
- NES 3 (a Master's student and research associate, 5 years of teaching experience).

The three Japanese NNES raters were recruited from different universities in Japan:

- NNES 1 (an associate professor, 10 years of teaching experience),
- NNES 2 (an associate professor, 8 years of teaching experience), and
- NNES 3 (a lecturer, 4 years of teaching experience).

All of the participants (both the student writers and teacher raters) provided informed consent for participation in this study.

Procedure

The data collection procedures of this study partially overlap the collection procedures of our previous study (study II; Hijikata-Someya et al., 2015). During the classes taught by the authors of the present study, the participants produced a 50- to 60-word written summary of a 199-word comparison/contrast type passage entitled Right Brain/Left Brain, which was adapted from Oshima and Hogue (2007). The summary writing task took approximately 40 min, and the use of dictionaries was allowed. The six raters first scored the 102 summaries, which had been produced by 51 students (each student wrote the summaries twice, before and after an L2 summary writing instruction) using the ETS (2002) holistic rubric (scores ranging from 1 to 5). The raters then judged the difficulty of rating each summary using three indicators: 1 = *easy*, 2 = *moderate*, or 3 = *difficult*. Consequently, 16 out of 102 summaries (15.69%) were judged as difficult to score, because their average grading difficulty from the six raters exceeded 2.0. In Hijikata-Someya et al. (2015), we then used these 16 difficult summaries to identify what aspects of the summaries made them difficult, through a qualitative consideration of the raters' retrospective scoring comments.

In the current study, we asked the same three NES raters and three NNES raters to score the difficult 16 summaries using the five-dimensional rubric (Appendix 1: scores ranging from 1 to 4) that we developed in study III (Yamanishi & Ono 2018). The 16 difficult summaries from Hijikata-Someya et al. (2015) were used again in this study for two reasons. First, the 16 summaries were thought to be suitable for examining our newly developed rubric's potential under a severe condition. In other words, we thought that if the rubric could demonstrate positive results even in severe and challenging conditions, it would be more reasonable to claim the potential of the rubric. Taking the implementation of the rubric in the classroom into consideration, the other reason is that we thought 102 summaries were too many for the raters to score using a multi-dimensional rubric. Sixteen months had passed between when the six raters rated the 16 summaries in the current study and when they first rated the 102 summaries; therefore, we judged the influence of the first scoring on the scoring in this study to be negligible.

Before scoring, as a form of rater training, the raters were provided with a sample of anchor summaries (Appendix 2) that had been scored by us to illustrate a variety of summaries with different scores assigned to each dimension. Therefore, the raters were able to familiarize themselves with the new dimensions of the rubric, assess the severity of the scoring, and understand the appropriate scores for each score band.

We used an open-ended questionnaire, which consisted of the following three questions, to solicit the raters' retrospective comments on our rubric:

1. The revised rubric has five dimensions: content, paraphrase (quantity), paraphrase (quality), language use, and overall quality. What is your opinion on the usage of each of the dimensions?
2. What is your opinion about the levels and descriptors within each dimension?
3. What is your overall impression of the revised rubric?

The questionnaire was in English for NES raters and in Japanese for NNES raters. After the raters completed the evaluation of the summaries and the questionnaire, their data was collected by email. As an alternative to the questionnaire, we considered using face-to-face interviews to receive feedback from the raters. However, because the raters belonged to different universities located in different areas of Japan and the UK, it would have been difficult for us to conduct in-person interviews during the limited data collection period.

Analysis

For our quantitative examination, the collected evaluation data of our rubric was analyzed and compared with the results of the ETS holistic rubric. The following analyses were performed:

1. Reliability of our rubric,
2. Correlations within and between the rubrics,
3. Inter-rater reliability for both rubrics, and
4. Generalizability of our rubric.

As the differences in rater background—i.e., NES or NNES rater—were outside the scope of this study, these analyses were conducted using the average mean scores of the six raters.

In this study, as a methodological advantage, we employed the generalizability theory (G theory) to examine the characteristics of our rubric in detail (Bachman, 2004; Brennan, 1992; In'nami & Koizumi, 2016; Lynch & McNamara, 1998; Shavelson &

Webb, 1991). G theory "can decompose the score variances into those affected by the numerous factors and their interactions" (In'nami & Koizumi, 2016, p. 342), and the factors and their interactions are called variance components. The design and data analysis stage of G theory is a generalizability study (G study). The G study design of this study is a two-facet crossed design, and the variance components examined are listed as follows (see also Table 4):

- The variance associated with the object of measurement or the amount of inconsistency across student-writers' summaries is symbolized by *p*;
- The variance across raters or the amount of inconsistency across raters is symbolized by *r*; and
- The variance across items of the rubric or the amount of inconsistency across items is symbolized by *i*.

In addition, their possible interactions are:

- The interactions between summaries and raters, $p \times r$;
- The interactions between summaries and dimensions, $p \times i$;
- The interactions between raters and dimensions, $r \times i$; and
- The unresolved residuals, shown as $p \times r \times i$, that are not accounted for by the other variance components.

G theory also allows us to simulate what the evaluation results would look like if we changed the numbers of raters and/or dimensions of a rubric to determine how to improve future evaluations. This stage is called the decision study, the D study, and this enables the simulation of changes to the generalizability coefficient (*g* coefficient) based on changes to the number of raters and/or dimensions in the rubric. The *g* coefficient (Eρ^2 or *G*) is theoretically and practically equivalent to the reliability coefficient (α) in the classical test theory, and its maximum possible value is 1. The *g* coefficients here were calculated using Eq. 1 for the two-facet crossed design (see Shavelson & Webb, 1991, for more details):

$$G = \frac{p}{p + \frac{p \times r}{Nr} + \frac{p \times i}{Ni} + \frac{p \times r \times i}{Nr \times Ni}} \tag{1}$$

For our qualitative examination, the NES and NNES raters' retrospective comments on our rubric were analyzed descriptively to supplement the findings of the quantitative examination. The rater comments were first categorized into the following four categories: (a) usage of each dimension (16 units), (b) levels and descriptors within each dimension (11 units), (c) overall impression of the rubric (6 units), and (d) concerns about the rubric (4 units). Then, the first two categories were further divided into subcategories representing each dimension. Table 1 shows the distribution of the subcategories for the first two categories in the raters' comments.

As illustrated in Table 1, each dimension was discussed by the NES and NNES raters. Similarly, the rater comments regarding the overall impression of and concerns about the rubric were also analyzed. Their comments are discussed in detail

Table 1 Components of the raters' comments

Dimensions	Usage of each dimension		Levels and descriptors within each dimension	
	NES (*n* = 3)	NNES (*n* = 3)	NES (*n* = 3)	NNES (*n* = 3)
Content	1	2	1	1
Paraphrase (quantity)	3	1	0	2
Paraphrase (quality)	1	1	0	2
Paraphrase (quantity and quality)*	1	2	0	1
Language use	1	1	1	2
Overall quality	1	1	1	0

Note: Each rater could have more than one comment for each category
*This indicates that the quantity and quality of the paraphrase was reflected in a single comment

in the next section. In the discussion, the Japanese NNES raters' comments are English translations.

Results and discussion

Quantitative examinations

Reliability of our rubric

First, we calculated the reliability (internal consistency) of our newest rubric to compare the four-dimensional version, which consisted of content, paraphrase (quantity), paraphrase (quality), and language use, with the five-dimensional version, in which the holistic overall quality was added as an optional dimension. As Table 2 illustrates, while the reliability of the four-dimensional rubric was not high ($\alpha = .48$), adding the overall quality dimension improved the reliability of the rubric ($\alpha = .70$).

The reliability of the five-dimensional version exceeded $\alpha = .60$, which is regarded as moderately high reliability in performance assessments such as L2 writing assessments (Kudo & Negishi, 2002). This result indicates that by adding the overall quality dimension with a simple descriptor (*As a response to this task, the overall quality of this summary is*...), the rubric became more robust.

Correlations within and between the rubrics

To examine the potential of the overall quality dimension when used exclusively as a holistic measure and the two paraphrasing dimensions, correlation analyses were performed (see Fig. 2). As the number of summaries evaluated was not large ($N = 16$), we adopted a non-parametric measure (Spearman's rho).

Figure 2 demonstrates that the correlation coefficient between overall quality and the four-dimensional total score—the sum of content, paraphrase (quantity), paraphrase (quality), and language use dimensions—was positive and very high, $\rho = .91$, $p < .05$. This means that the overall quality dimension has the potential to work well even if teacher raters use it as the sole factor for evaluating student summaries. The correlation coefficient between the two separate paraphrasing dimensions, paraphrase (quantity) and paraphrase (quality), was positive and high, $\rho = .81$, $p < .05$. This suggests that these two dimensions

Table 2 Internal consistency of our four- and five-dimensional rubrics

	Four-dimension	Five-dimension
α	.48	.70

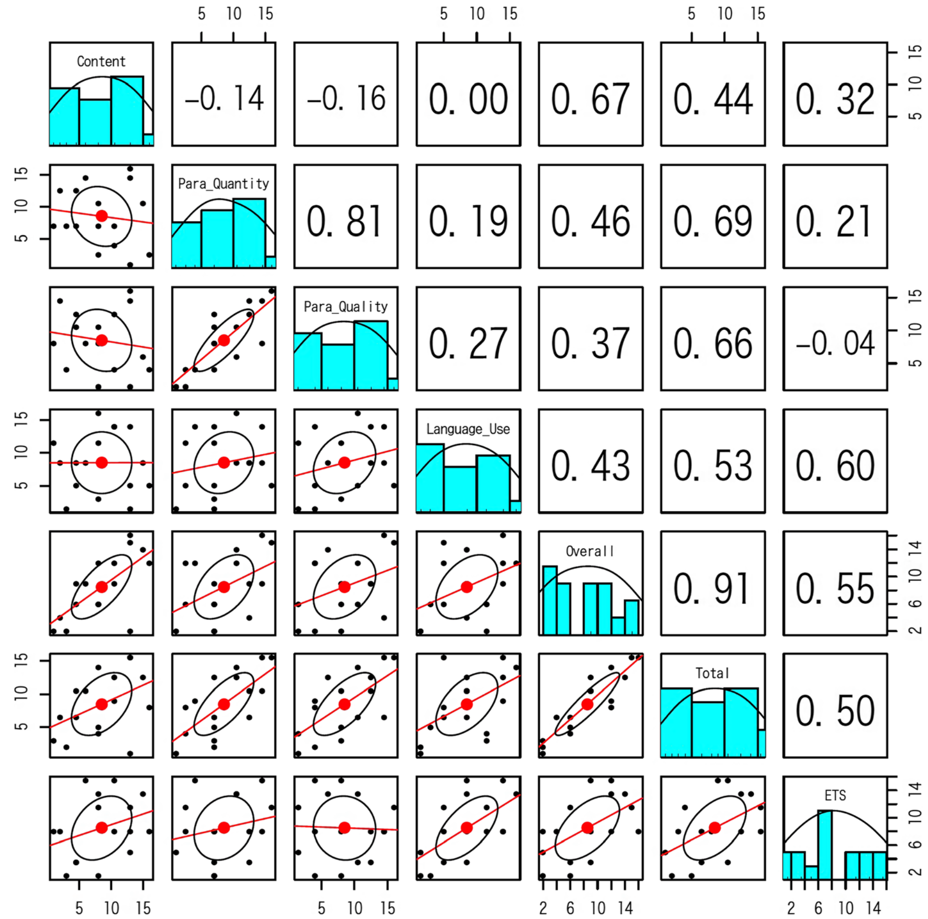

Fig. 2 Correlation matrix and histogram. Scatter plots are shown in the lower left triangle, correlation coefficients are shown in the upper right triangle, and the remainder are histograms

overlap to some extent but cover different aspects of paraphrasing. To investigate the concurrent validity of our rubric, the correlations between our rubric (the total and overall quality dimension) and the commonly used ETS holistic rubric were examined. The correlation coefficients between the ETS holistic rubric and our rubric were positive and moderate, $\rho = .50$, $p < .05$ (the ETS holistic and the total) or $\rho = .55$, $p < .05$ (the ETS holistic and overall quality dimension), proving that these two rubrics evaluate the same construct but focus on slightly different aspects of the L2 summary writing performance.

Inter-rater reliability for both rubrics

For the rater perspective, the inter-rater reliabilities among the six raters (three NES raters and three NNES raters) were examined. First, as Table 3 illustrates, the four dimensions that make up the analytic components of our rubric demonstrated reasonably high inter-rater reliabilities, $\alpha = .66–.75$. Next, the inter-rater reliabilities of the overall quality dimension, the total, and the ETS's holistic rubric (shown in italics in Table 3) were examined.

The results indicate that the overall quality dimension ($\alpha = .73$) produced higher inter-rater reliability than the total ($\alpha = .68$) and the ETS holistic rubric ($\alpha = .59$). This indicates the potential of using the overall quality dimension alone as a holistic assessment.

Table 3 Inter-rater reliability coefficients ($N = 6$)

	Content	Paraphrase (quantity)	Paraphrase (quality)	Language use	Overall quality	Total	ETS
α	.75	.67	.70	.66	.73	.68	.59

Note: Total is the sum of the content, paraphrase (quantity), paraphrase (quality), and language use dimensions

Generalizability of our rubric

Based on the confirmation of the reasonably high reliability ($\alpha = .70$) of our five-dimensional rubric as a performance evaluation tool, we used G theory to examine the characteristics of the rubric in more detail. At the first stage of the G theory analysis, we conducted a generalizability study (G study).

In Table 4, the estimated variance components except for $p \times r \times i$ (residual) are arranged in the order of the magnitude of their effect on the evaluation (from largest to smallest). We examined the three interactions of $p \times i$, $p \times r$, and $i \times r$ and found that the largest interaction was between the summaries and dimensions ($p \times i$, 25.4%). This result is expected because each student writer's L2 summary writing performance assessed through each dimension of our rubric should vary, meaning each dimension was appropriately able to evaluate similar but different aspects of the students' summaries. The second largest interaction was between the summaries and raters ($p \times r$, 19.1%). This indicates that each rater's evaluation was moderately affected by each student writer's varied performance. The last interaction was between the dimensions and raters ($i \times r$), which contributed to a small portion of the variance, 6.3%. This means that each rater's interpretation of each dimension of the rubric did not differ to a great extent, which in turn means that the raters' evaluation of using our rubric was fairly consistent.

The second stage of the G theory analysis was the decision study (D study), the results of which are illustrated in Fig. 3. The default number of raters was six, and the default for the dimensions was five, and we simulated the g coefficients by changing both numbers using Eq. 1. In this study, the simulation changed the numbers from 1 to 10, as shown in Fig. 3. The results of the D study examinations demonstrated that if the number of raters was reduced to three, the g coefficient would be fairly low, but slightly higher than .50. This suggests that there is some room for improvement for when this rubric is implemented in classrooms. Theoretically, the same kind of examination could also be conducted for the number of dimensions in the rubric; however, changing the number of dimensions is not a realistic option because it would require determining the validity of the revised rubrics (e.g., combining two dimensions into one). Thus, the simulation results reflecting the changes to the number of dimensions are shown simply

Table 4 Estimated variance components

Sources of variance	Estimated variance components
$p \times i$	0.184 (25.4%)
$p \times r$	0.138 (19.1%)
p	0.111 (15.4%)
$i \times r$	0.045 (6.3%)
i	0.025 (3.4%)
r	0.000 (0.0%)
$p \times r \times i$ (residual)	0.220 (30.4%)

Note: p is the object of measurement (here, student summaries), i is the items (here, dimensions), and r is the raters. Except for the residual, the variance components are arranged from largest to smallest. The sum of the percentages in parentheses is 100%

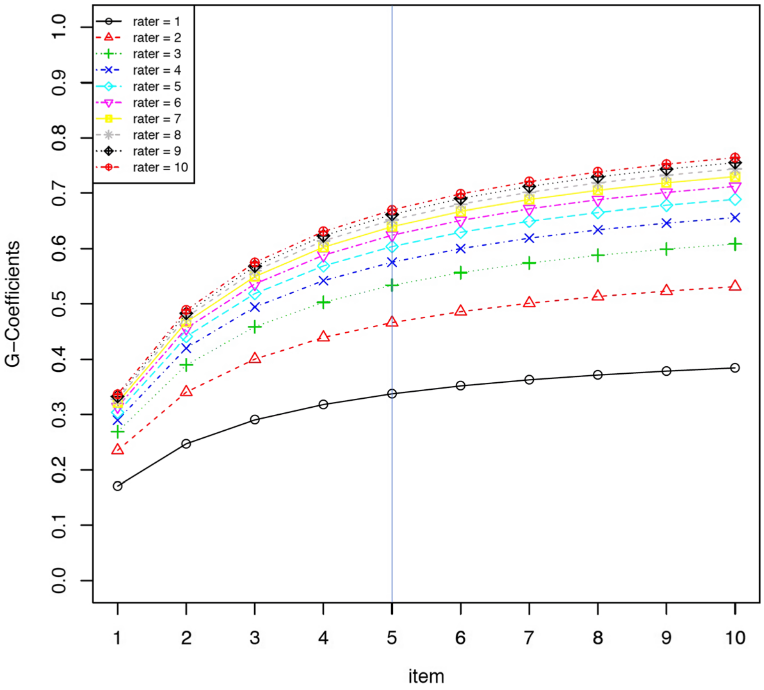

Fig. 3 D study simulation. The results changing the number of raters and dimensions are shown. The original number of raters and dimensions was six and five, respectively

as a reference. However, creating and utilizing D study simulations after conducting an evaluation is preferable as it allows useful, diagnostic information to be obtained that corresponds to its situation and purpose.

Qualitative examinations

Overall impression of the rubric

The analysis of the rater comments demonstrated that the raters' overall impression of the rubric was positive. Both the NES and NNES raters found our rubric easier to use than the ETS (2002) holistic rubric. For instance, the raters NES 1 and NNES 1 reported the following:

> "Compared to the previous holistic scale, the analytic scale was easy to use for the summary evaluation." (NES 1)

> "The new rubric was very easy to use." (NNES 1)

Previously, the NES and NNES raters struggled when using the holistic rubric because it treated several aspects together (Hijikata-Someya et al., 2015). However, our rubric prevented such struggles and helped the raters perform a smooth assessment of the L2 summaries. The practicality of a rubric is important for classroom use; if a rubric is easy to use, it can save language teachers' time and costs (Stevens & Levi, 2013). Additionally, students can be encouraged to use the rubric for learning and peer-assessment purposes (Becker, 2016). Although there has been a claim that holistic scoring is more practical and cost-effective than analytic scoring (Bacha, 2001; Hamp-Lyons, 1995; Hyland, 2003; Weigle, 2002), analytic scoring may better suit integrated writing tasks such as summary writing for educational purposes. Thus, the raters' positive perceptions of our new rubric indicate that the rubric can enhance teachers' practical assessment of and diagnostic feedback on student summary writing performance in EFL classrooms.

Content

The content dimension was perceived positively by the NES and NNES raters. For instance, NNES 3 referred to the content as follows:

> "It was very easy to use. Specifically, the reference to 'secondary information' in the descriptor 4 was helpful when judging whether examples in the summaries were acceptable or not. Also, I didn't need to look back at the descriptors often because each level of the descriptors had a consistent description regarding 'information'." (NNES 3)

The descriptors of our rubric were regarded as clear and distinctive by the NES and NNES raters. As we discussed earlier, because L2 summary writing requires complex information processing to select the main ideas (Brown & Day, 1983), assessing the content of the summaries tends to be challenging for some raters and language teachers who do not necessarily identify the same main ideas to be included in the summaries (Alderson, 2000). Based on these challenges, the content dimension in our rubric is expected to help raters evaluate the selected information in the summaries more effectively.

Paraphrase (quantity) and paraphrase (quality)

Paraphrase (quantity) and paraphrase (quality) were also regarded as reasonable by the NES and NNES raters. Both groups of raters provided positive opinions towards the two separate paraphrasing dimensions and understood the purpose of having both dimensions.

> "It is good that the rubric distinguishes the writers' effort to paraphrase from the appropriateness of paraphrasing in dealing with both the quantity and quality of paraphrasing." (NNES 1)

> "I like the idea of measuring originality on two dimensions, e.g., quantity and quality." (NES 2)

The establishment of paraphrase (quantity) and paraphrase (quality) in the rubric seemed to work well in the EFL context and was perceived positively for educational purposes as it met the teacher raters' needs and demands. When they used the holistic rubric previously, they suggested that an ideal rubric for written summaries should explicitly deal with the important and difficult paraphrase dimension (Hijikata-Someya et al., 2015). Thus, these teacher raters' opinions were reflected in our rubric through the use of explicit, self-explanatory descriptors for the paraphrase dimensions.

However, for paraphrase (quality), NNES 3 found it difficult to evaluate whether "more than four words in a row were copied from the original text," which was written in the score bands 1 and 2 in paraphrase (quality). In a similar vein, NNES 3 suggested the importance of teacher instruction in paraphrasing as follows:

> "To begin the summary task or as general writing instruction in class, teachers should tell students not to copy more than four consecutive words from the source text." (NNES 3)

NES 1 also commented on the innovative feature of the paraphrase (quantity) dimension in the rubric:

> "It is good that the descriptors of Paraphrase (Quantity) are clear since [the] percentage of paraphrase at each level is specified." (NES 1)

Instead of raters needing to calculate the exact percentage of paraphrases employed in the summaries, we provided the percentages shown in the descriptors as an approximate estimation of the paraphrasing to be employed. The use of labels for paraphrasing attempts such as *Near Copy*, *Moderate Revision*, and *Substantial Revision* (Keck, 2006, 2014) in the rubric seems to work well in some contexts, but these labels might be interpreted differently by individual teacher raters and student writers. Thus, our rubric was determined to be self-explanatory in terms of paraphrase (quantity) based on the listed percentages of paraphrasing provided for each level.

Overall quality

The newly added overall quality dimension as a holistic measure was viewed as effective and useful by the NES and NNES raters, as demonstrated by the following comments:

> "As this is a holistic assessment of the summary, it could be used to check the consistency of the analytical scores." (NES 2)

> "Very easy to use. I felt that this aspect prevents summaries that only actively paraphrase from getting a high mark." (NNES 3)

Although the basic feature of our rubric was analytic, the overall quality dimension served as a holistic assessment of the summaries, when considering all dimensions as a whole. In line with this, Marshall (2017) states the importance of the summary to represent "a sense of the complete original text" (p. 71). Thus, the addition of this holistic dimension allows raters to employ both analytic and holistic views of the summaries and maintain consistency between the two assessment methods.

However, NES 2 suggested the potential need for descriptors of overall quality concerning each of the four score bands as follows:

> "As there are many contributing factors to a successful summary, there is a risk that the interpretation of poor / fair / good / very good will differ between [the] raters (effecting [sic] inter-rater reliability). Perhaps there could be descriptors for the different levels of 'Overall Quality'?" (NES 2)

This suggestion should be considered to improve the rubric because the overall quality dimension can be used alone. One way to improve this would be to add descriptors explaining each score band, but this may impair the dimension's simple and holistic nature. Another option would be to develop a scoring method based on individual teacher raters' satisfactory standards that correspond to varied teaching contexts (e.g., Sawaki, 2019). For example, we could set the standard/reference point based on the teachers' needs and students' expertise as a score of 3 (good). If the quality of a summary is above the standard, it would be marked as 4 (very good), and if it is below the standard, it would be marked as 2 (fair) or 1 (poor). This option regards the overall quality dimension as a holistic criterion-referenced measure (the criterion score here is 3) to effectively maintain the holistic nature of this dimension.

Concerns about the rubric

NNES 2 highlighted concerns about different scoring weights between the paraphrase and content dimensions:

> "The aspect of paraphrase in the rubric seems reasonable since quantity and quality are treated separately whereas I felt that it may be questionable in terms of the balance between Paraphrase and Content in a total score . . . . I personally think that, in my case, the priority in the summary evaluation tends to be accurate reading comprehension rather than paraphrasing skills and language use." (NNES 2)

This opinion is understandable because paraphrase is weighted more than the other dimensions due to the existence of paraphrase (quantity) and paraphrase (quality) in the rubric. In addition, the priority of the skills in the summary writing task depends on the purpose of the task and assessment, that is, whether it is used as a reading comprehension task or an integrated writing task. Hijikata-Someya et al. (2015) revealed that language teachers particularly struggled to assess and teach paraphrasing attempts in L2 summary writing. Therefore, this newly developed rubric emphasizes the importance of paraphrasing by including two paraphrase dimensions and placing more scoring weight on paraphrase than the other dimensions.

Conclusion

In this study, as the final step of our research project on assessing L2 summary writing (see Fig. 1), we have gained insight into the potential use and function of our rubric through quantitative and qualitative examinations. With regard to the quantitative examination, the results of comparing four and five dimensions clarified that the optional overall quality dimension should be included because it improves the reliability of the rubric significantly, from $\alpha = .48$ to .70. We then examined the correlations between our rubric and the ETS (2002) holistic rubric. The results revealed that the newly added overall quality dimension could work well even if used alone, and our rubric and the ETS holistic rubric had a positive, moderate correlation for L2 summary writing assessments.

The examination of our rubric using G theory explained the nature of the evaluation results in detail through the G study examination. However, the D study simulation based on changing the number of raters indicated that the reliability (generalizability coefficients) of our rubric was not high enough when the number of raters decreased.

This result suggests that we need further research to seek ways to help raters evaluate summaries using the rubric. One way to improve this might be to provide teacher raters with opportunities to practice evaluating student L2 summaries using the rubric in a rater training session. When administering these training sessions, making good use of G and D study examinations would be very helpful in specifying the characteristics of the evaluation results to improve future evaluation.

The qualitative examinations supplemented the findings from our quantitative examination and demonstrated the potential of the overall quality dimension. Ideally, all the five dimensions of our rubric should be employed because of their high reliability; however, in educational contexts, the use of the overall quality dimension is a valid option. For example, the overall quality dimension can be used alone as a holistic assessment, depending on the purpose of the assessment and teaching in individual teaching contexts. In other words, if only holistic assessment results are needed, teacher raters might use this dimension exclusively; if they can conduct both holistic and analytic assessments, that is preferable. It should be noted that when the overall quality dimension is used alone for the assessment of students' L2 summaries, teacher raters should review and understand the other four dimensions and their descriptors in advance. Nevertheless, we do not exclude the usage of the four-dimensional rubric alone as a rigid and independent analytic assessment. In essence, this flexible combination of holistic and analytic assessments is expected to enhance the effective and efficient evaluation and teaching of L2 summary writing in various educational settings, based on the purpose of summary writing tasks and educational levels. Even within the Japanese EFL context, teachers' needs and students' expertise may vary considerably, which creates the demand for a flexible assessment tool such as the one we have developed.

The correlation coefficient of the two separate paraphrasing dimensions, paraphrase (quantity) and paraphrase (quality), was positive and high, indicating that they overlap to some extent but cover different and important aspects of paraphrasing. The teacher raters' comments confirmed this; they appreciated the distinction of these two aspects of paraphrasing, which met their needs and demands. Teacher raters can emphasize both the quantitative and qualitative aspects of paraphrasing when they teach L2 summary writing by using this rubric, and the features of our rubric can also be helpful for student writers to understand the importance of paraphrasing. Student writers can understand how much paraphrasing is expected from the self-explanatory nature of the paraphrase (quantity) dimension, while the paraphrase (quality) dimension may encourage them to actively and appropriately paraphrase to a greater degree.

Finally, we address the limitations that will be considered in our future studies. As a methodological limitation, the number of raters was not large, and their teaching backgrounds were not strictly controlled. Further investigation into the raters' attributes and backgrounds that could affect the evaluation results may be necessary. Another limitation is related to task selection: only a single summary writing task based on a comparison/contrast type of passage was employed. If more than one type of text or genre had been used, it would have been possible to discuss the appropriateness of the rubric from a broader perspective. Similarly, only a relatively short passage (i.e., 199 words) was used for the summary writing task, which means that there was no way to compare task difficulty. A comparison of both short and long passages for summary writing tasks could be ideal to ensure the effectiveness of the rubric. Furthermore, the summaries produced in this study were relatively short (i.e., 50–60 words); therefore, most of the summaries were written in one paragraph. If a long passage is used for a summary writing task and a longer summary is produced, new dimensions such as the organization of the summary may need to be added to the rubric.

Despite these limitations, our newly developed rubric is innovative as it embodies the teacher raters' voices and experiences and is characterized as a "hybrid" of analytic and holistic assessments within a single rubric. Similar to a hybrid car using a conventional engine and an electric motor separately or simultaneously depending upon the situation and purpose, our new rubric offers flexibility for specific individual purposes. We hope that this rubric is helpful in teaching and assessing L2 summary writing in Japan and potentially in other EFL contexts.

Appendix 1

Table 5 The five-dimensional rubric

Dimension	Level		Criteria	観点	レベル		基準
CONTENT	4	very good	Can grasp all of the main ideas. Can develop the main point substantially by occasionally using secondary information.	内容	4	大変良い	要点を全て把握することができる。時々，二次的情報を使いながら，論旨を十分に展開することができる。
	3	good	Can grasp most of the main ideas. Includes somewhat incorrect information or information beyond the original text, but it does not substantially deviate from the main point.		3	良い	大体の要点を把握することができる。多少誤った情報や原文から飛躍した情報が含まれるが，論旨から大きく逸脱するほどではない。
	2	fair	Can grasp only limited main ideas. Cannot demonstrate an adequate development of the main point. Noticeably includes incorrect information or information beyond the original text.		2	あまり良くない	限られた要点しか把握することができない。論旨を十分に展開することができない。誤った情報や原文に書かれていない情報が目立つ。
	1	poor	Cannot identify main ideas. Cannot grasp main ideas correctly.		1	良くない	要点を選ぶことができない。要点を正確に把握することができない。
PARAPHRASE (Quantity)	4	very good	Can paraphrase 80% or more of the expressions included in the summary in one's own words.	言い換え（量）	4	大変良い	要約を構成する80%以上の表現を自分の言葉（原文外の言葉）で言い換えることができる。
	3	good	Can paraphrase from 50% to less than 80% of the expressions included in the summary in one's own words.		3	良い	要約を構成する50%以上80%未満の表現を自分の言葉で言い換えることができる。
	2	fair	Can paraphrase only from 25% to less than 50% of the expressions included in the summary in one's own words.		2	あまり良くない	要約を構成する25%以上50%未満の表現しか自分の言葉で言い換えることができない。
	1	poor	Can paraphrase only less than 25% of the expressions included in the summary in one's own words.		1	良くない	要約を構成する25%未満の表現しか自分の言葉で言い換えることができない。
PARAPHRASE (Quality)	4	very good	Can actively attempt to paraphrase. Can demonstrate effective paraphrases where both sentence construction and vocabulary choice are different from the original text.	言い換え（質）	4	大変良い	積極的に言い換えを試みることができる。文構造および語彙選択が共に原文とは異なる形で効果的な言い換えを提示することができる。
	3	good	Can actively attempt to paraphrase. Can paraphrase using vocabulary different from the original text. Seldom changes sentence construction from the original text.		3	良い	積極的に言い換えを試みており，原文とは異なる語彙を使用して言い換えることができる。原文の文構造はほとんど変えていない。
	2	fair	Includes few expressions consisting of more than 4 words in a row copied from the original text. Can only demonstrate paraphrases using vocabulary from the original text. Deletes expressions partially or changes word order.		2	あまり良くない	5語以上連続して原文から抽出する表現はほとんど見られないが，原文で用いられている語彙をつなぎ合わせるだけの言い換えしかできていない。一部を削除したり，語順を言い換えたりしている。
	1	poor	Includes a number of expressions consisting of more than 4 words in a row copied from the original text. Cannot demonstrate effective paraphrases.		1	良くない	連続して5語以上原文からそのまま借用する表現が多い。効果的な言い換えができていない。
LANGUAGE USE	4	very good	Can demonstrate a sophisticated range of vocabulary with effective word/idiom choice and usage. Can demonstrate effective and complex sentence construction with few grammatical errors.	言語使用	4	大変良い	語彙範囲が洗練され，語/慣用句の選択と使用が効果的である。文法的な誤りがほとんどなく，効果的で複雑な文構造を提示することができる。
	3	good	Can demonstrate an adequate range of vocabulary with good word/idiom choice and usage. Can demonstrate simple but effective sentence construction. Includes minor and occasional errors.		3	良い	語彙範囲が十分で，語/慣用句の選択と使用が辛うじて適切とみなせる要約を提示することができる。単純であるが効果的な文構造を提示することができる。軽微な誤りが時折含まれる。
	2	fair	Can demonstrate only a limited range of vocabulary, word/idiom choice and usage. Can demonstrate simple sentence construction. Meaning is obscure due to frequent major errors.		2	あまり良くない	語彙範囲と，語/慣用句の選択と使用が限られたものしか提示できない。頻繁に起こる誤りのため，意味が不明瞭である。提示された文構造は単純である。頻繁に起こる重大な誤りのため，意味が不明瞭である。
	1	poor	Can demonstrate little knowledge of vocabulary, idioms, and word form. Can demonstrate little knowledge of sentence construction rules and English writing conventions. Meaning is obscure due to a number of minor and major errors.		1	良くない	語彙，慣用句，語形に関する知識がほとんど提示できない。文構造の規則と英語のライティングの慣習に関する知識がほとんど提示できていない。多くの誤りが含まれるため、意味が不明瞭である。
OVERALL QUALITY	4	very good	As a response to this task, the overall quality of this summary is...	全体的な要約の質	4	大変良い	本課題に対するこの要約の質を総合的に判断すると…
	3	good			3	良い	
	2	fair			2	あまり良くない	
	1	poor			1	良くない	

Appendix 2

Table 6 Sample summaries and their scores

Sample summary 1						
Dimension	Content	Paraphrase (quantity)	Paraphrase (quality)	Language use	Overall quality	Total
Score	3	4	4	3	3	17
There are some different ways of cerebral activity in the right and left sides of the brain. The former is more intuitively and sentimentally, on the other hand, the latter is more logical and reasonable. Even if one side is stronger than the other side, both sides can work regularly. (Summary_a10)						
Sample summary 2						
Dimension	Content	Paraphrase (quantity)	Paraphrase (quality)	Language use	Overall quality	Total
Score	3	3	3	3	3	15
The left and right sides of your brain process information in different ways. The left side is more logical. On the other hand, the right side uses the five senses more. So a left-brained and a right-brained person think in different ways. Though, usually people's both sides of their brain work together. (Summary_A03)						
Sample summary 3						
Dimension	Content	Paraphrase (quantity)	Paraphrase (quality)	Language use	Overall quality	Total
Score	2	3	2	2	2	11
Our brain thinks in different process. We use the right brain in visual and intuitive and sensual. However, we use the left side in logical and verbal. So, it is different from a left-brained person to a right-brained person. But, we do not use one side only. So we always use both sides. (Summary_a09)						

Note: A full score is 20

Abbreviations

EFL: English as a foreign language; ESL: English as a second language; ETS: Educational Testing Service; L1: First language; L2: Second language; MEXT: Ministry of Education, Culture, Sports, Science and Technology; NES: Native English speaker/speaking; NNES: Non-Native English speaker/speaking; TOEFL-ITP: Test of English as a Foreign Language—Institutional Testing Program; TOEIC-IP: Test of English for International Communication—Institutional Program

Acknowledgements

The authors thank all the students and the raters who participated in this study.

Authors' contributions

All authors read and approved the final manuscript.

Funding

The project was funded by grants from the Japan Society for the Promotion of Science, KAKENHI (#23520725 and #26580121).

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Faculty of Science and Engineering, Chuo University, Tokyo, Japan. [2]Faculty of Law, Keio University, Kanagawa, Japan. [3]Faculty of Humanities and Social Sciences, University of Tsukuba, Ibaraki, Japan.

Received: 7 April 2019 Accepted: 4 June 2019
Published online: 24 July 2019

References

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Baba, K. (2009). Aspects of lexical proficiency in writing summaries in a foreign language. *Journal of Second Language Writing, 18*, 191–208. https://doi.org/10.1016/j.jslw.2009.05.003.

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System, 29*, 371–383. https://doi.org/10.1016/S0346-251X(01)00025-2.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Becker, A. (2016). Student-generated scoring rubrics: Examining their formative value for improving ESL students' writing performance. *Assessing Writing, 29*, 15–24. https://doi.org/10.1016/j.asw.2016.05.002.

Black, L., Daiker, D. A., Sommers, J., & Stygall, G. (1994). *New directions in portfolio assessment*. Portsmouth: Boynton/Cook Heinemann.

Brennan, R. L. (1992). *Elements of generalizability theory (Rev. ed.)*. Iowa City: ACT Publications.

Brown, A. L., & Day, J. D. (1983). Macrorules for summarizing texts: The development of expertise. *Journal of Verbal Learning and Verbal Behavior, 22*, 1–14. https://doi.org/10.1016/S0022-5371(83)80002-4.

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing, 10*, 5–43. https://doi.org/10.1016/j.asw.2005.02.001.

Delaney, Y. A. (2008). Investigating the reading-to-write construct. *Journal of English for Academic Purposes, 7*, 140–150. https://doi.org/10.1016/j.jeap.2008.04.001.

Educational Testing Service. (2002). *LanguEdge courseware: Handbook for scoring speaking and writing*. Princeton: Educational Testing Service.

Gebril, A., & Plakans, L. (2016). Source-based tasks in academic writing assessment: Lexical diversity, textual borrowing and proficiency. *Journal of English for Academic Purposes, 24*, 78–88. https://doi.org/10.1016/j.jeap.2016.10.001.

Grabe, W. (2001). Reading-writing relations: Theoretical perspectives and instructional practices. In D. Belcher & A. Hirvela (Eds.), *Linking literacies: Perceptions on L2 reading-writing connections* (pp. 15–47). Ann Arbor: University of Michigan Press.

Hamp-Lyons, L. (1995). Rating nonnative writing: The trouble with holistic scoring. *TESOL Quarterly, 29*, 759–765. https://doi.org/10.2307/3588173.

Hidi, S., & Anderson, V. (1986). Producing written summaries: Task demands, cognitive operations and implications for instruction. *Review of Educational Research, 56*, 473–493. https://www.jstor.org/stable/1170342.

Hijikata-Someya, Y., Ono, M., & Yamanishi, H. (2015). Evaluation by native and non-native English teacher-raters of Japanese students' summaries. *English Language Teaching, 8*(7), 1–12. https://doi.org/10.5539/elt.v8n7p1.

Hijikata, Y., Yamanishi, H., & Ono, M. (2011). The evaluation of L2 summary writing: Reliability of a holistic rubric. Paper presented at the 10th Symposium on Second Language Writing in 2011. Taipei, Taiwan: Howard International House.

Hirvela, A., & Du, Q. (2013). "Why am I paraphrasing?": Undergraduate ESL writers' engagement with source-based academic writing and reading. *Journal of English for Academic Purposes, 12*, 87–98. https://doi.org/10.1016/j.jeap.2012.11.005.

Hyland, K. (2002). *Teaching and researching writing*. Harlow: Pearson Education Limited.

Hyland, K. (2003). *Second language writing*. Cambridge: Cambridge University Press.

In'nami, Y., & Koizumi, R. (2016). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language Testing, 33*, 341–366. https://doi.org/10.1177/0265532215587390.

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

Keck, C. (2006). The use of paraphrase in summary writing: A comparison of L1 and L2 writers. *Journal of Second Language Writing, 15*, 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.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*, 363–394. http://dx.doi.org/10.1.1.468.1535

Kirkland, M. R., & Saunders, M. A. P. (1991). Maximizing student performance in summary writing: Managing cognitive load. *TESOL Quarterly, 25*, 105–121. https://doi.org/10.2307/3587030.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. Language Testing, 26, 275–304. https://doi.org/10.1177/0265532208101008.

Kudo, Y., & Negishi, M. (2002). Interrater reliability of free composition ratings by different methods. *Annual Review of English Language Education in Japan (ARELE), 13*, 91–100.

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

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*, 158–180. https://doi.org/10.1177/026553229801500202.

Marshall, S. (2017). *Advance in academic writing: Integrating research, critical thinking, academic reading and writing*. Montréal: Pearson.

Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation, 7*(25). http://www.pareonline.net/getvn.asp?v=7&n=25.

MEXT. (2018). Koutou gakkou gakushuu shidou youryou (The course of study for secondary education). http://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2018/07/11/1384661_6_1_2.pdf

Ono, M. (2011) Japanese and Taiwanese university students' summaries: A comparison of perceptions of summary writing. *Journal of Academic Writing, 1*, 191–205. https://doi.org/10.18552/joaw.v1i1.14

Oshima, A., & Hogue, A. (2007). *Introduction to academic writing*. White Plains: Pearson Education.

Oshima, A., Hogue, A., & Ravitch, L. (2014). *Longman academic writing series, level 4: Essays*. White Plains: Pearson Education.

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

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing, 13*, 111–129. https://doi.org/10.1016/j.asw.2008.07.001.

Plakans, L. (2010). Independent vs. integrated writing tasks: A comparison of task representation. *TESOL Quarterly, 44*, 185–194. https://www.jstor.org/stable/27785076 .

Plakans, L. (2015). Integrated second language writing assessment: Why? What? How? *Language and Linguistics Compass, 9*, 159–167. https://doi.org/10.1111/lnc3.12124.

Sawaki, Y. (2019). Issues of summary writing instruction and assessment in academic writing classes. In *Paper presented at the 48th Research Colloquium of the Japan Language Testing Association*. Japan: Waseda University.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks: Sage.

Shi, L. (2012). Rewriting and paraphrasing source texts in second language writing. *Journal of Second Language Writing, 21*, 134–148. https://doi.org/10.1016/j.jslw.2012.03.003.

Shi, L., & Dong, Y. (2018). Chinese graduate students paraphrasing in English and Chinese contexts. *Journal of English for Academic Purposes, 34*, 46–56. https://doi.org/10.1016/j.jeap.2018.03.002.

Stevens, D. D., & Levi, A. (2013). *Introduction to rubrics: An assessment tool to save grading time, convey effective feedback, and promote student learning* (2nd ed.). Sterling: Stylus Publications.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: 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. https://doi.org/10.1016/j.asw.2004.01.002.

Yamanishi, H., & Ono, M. (2018). Refining a provisional analytic rubric for L2 summary writing using expert judgment. *Language Education & Technology, 55*, 23–48. https://iss.ndl.go.jp/books/R000000004-I029417776-00.

Yu, G. (2007). Students' voices in the evaluation of their written summaries: Empowerment and democracy for test takers? *Language Testing, 24*, 539–572. https://doi.org/10.1177/0265532207080780.

Yu, G. (2013). From integrative to integrated language assessment: Are we there yet? *Language Assessment Quarterly, 10*, 110–114. https://doi.org/10.1080/15434303.2013.766744.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.