Journal of Education and Learning; Vol. 14, No. 2; 2025
ISSN 1927-5250 E-ISSN 1927-5269
Published by Canadian Center of Science and Education

Scoring Difficulty in Summary Writing Assessment: Toward the Reconstruction of Analytic Rubric

Makiko Kato[1]

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

Correspondence: Makiko Kato, Auba-ku, Kawauti 27-1 980-8576 Sendai, Japan.

Received: August 3, 2024 Accepted: October 5, 2024 Online Published: October 29, 2024

doi:10.5539/jel.v14n2p74 URL: https://doi.org/10.5539/jel.v14n2p74

Abstract

This study aims to examine whether differences exist in the factors influencing the difficulty of scoring English summaries and determining scores based on the raters' attributes, and to collect candid opinions, considerations, and tentative suggestions for future improvements to the analytic rubric of summary writing for English learners. In this study, seven trained raters with diverse attribute backgrounds evaluated two kinds of English summaries written by Japanese university students using the analytic rubric with three evaluation items. A questionnaire was used to determine which of the three items were difficult to assess and why the raters perceived such difficulty, as well as what backgrounds and factors influenced their scoring decision-making. Moreover, through the raters' most recent experience, candid comments were collected for developing future rubrics. The results showed that whether the evaluators' attributes affected the difficulty of the evaluation was not clear. However, depending on the raters' experience in teaching English/assessing summary writing, requests for improvements in the descriptors of the evaluation items and in the rubric emerged. This study proposes a tentative analytic rubric for summary writing, providing a foundation for constructing a rubric that can be used more easily by future raters. It also highlights the opinions of expert and novice teachers conducting summary evaluations in education.

Keywords: summary writing, analytic rubric, learning English as a foreign language, raters with diverse backgrounds

1. Introduction

Integrated tasks such as summary writing are more complex and demanding than independent tasks (Brown et al., 2005). In Japan's higher education, students are motivated to obtain high scores on the Test of English as a Foreign Language Internet-Based Testing (TOEFL iBT) because numerous Japanese universities have policies such as credits for English language courses, scholarship programs, studying abroad programs, or placement tests (In'nami & Koizumi, 2019; Kato, 2021a). However, acquiring summarization skills is difficult, even for writers whose first language is English (Hirvela & Du, 2013).

In foreign language education, the complexity of evaluating integrated language tasks is well known (Gebril & Plakans, 2014; Shin & Ewert, 2015). Scores on integrated writing tasks have relatively low reliability (Gebril, 2009, 2010; Yamanishi et al., 2019). Therefore, recognizing that summary performance is influenced by various abilities and skills, including reading comprehension, summarization, writing, and metacognitive strategy skills, is necessary. When summary performance is assessed, the rater's reliability must be carefully considered, along with complex mixed factors.

Regarding raters' cognition or strategies, their inner aspect is worth examining to determine the factors influencing the scoring process and score decision as human raters may induce subjectivity in assessments (Suto, 2012). It has been reported that an important issue in the field of language assessment is the impact of raters' attributes, processes, and difficulties faced during scoring on the reliability of their evaluation (e.g., Chang et al., 2015; Milanovic et al., 1996; Suto, 2012). Raters' mental processes during scoring and decision-making depend on expertise and proficiency, which can contribute to optimal scoring reliability (Suto & Nadás, 2008).

2. Literature Review and Research Questions

Summary writing tasks are becoming increasingly important and are included in several high-stake English proficiency tests; however, integrated tasks are "more complex and demanding than traditional stand-alone or independent tasks" (Brown et al., 2005). Kirkland and Saunders (1991) argued that writing a summary is a highly

complex cognitive skill that imposes external and internal constraints. Conversely, Kato (2021a) proposed a working model of the summary-writing skill that displays two elements required for writing summaries in a foreign language: summarizing skills, which contain operating macro-rules and metacognitive skills, and language competence, which contains language ability and metacognitive skills, as described by Bachman and Palmer (1996).

As mentioned above, writing summaries in a foreign language requires several skills and abilities, and this complexity is not only limited to task performance but it also applies to scoring integrated tasks. Difficulties in the assessment of summary tasks occur for reasons such as difficulty in identifying the main ideas (Alderson, 2000) and judging the degrees of content distortion owing to language proficiency and of text-borrowing (Kato, 2021a).

Researchers have developed analytic rubrics to assess summarization skills and conducted studies on the validity and reliability of the rating items (Li, 2014; Li & Wang, 2021; Sawaki, 2019; Yamanishi et al., 2019). However, they have mainly focused on the quantitative analyses of scores and have not addressed raters' cognitive aspects. Although interest in raters' perceptions has increased in writing-evaluation research, many issues remain regarding the role of raters' aspects in summary assessment (Barkaoui, 2011; Cumming et al., 2021, 2002; Lumley, 2002; Milanovic et al., 1996; Weigle, 1994). Despite using the same scoring criteria, raters may give similar ratings for different reasons (Douglas & Selinker, 1992). The same is expected for summary evaluations. Focusing on the fact that collecting qualitative data may help observe the details of the internal and cognitive aspects of summary raters, this study sets the following four research questions (RQ) for reconstructing the analytic rubric for EFL (English as a Foreign Language) learners' summary writing:

RQ1. Which evaluation items do the raters find difficult or easy to assess?

RQ2. Why do the raters find them difficult or easy to assess?

RQ3. What affects the raters' evaluation decisions?

RQ4. What do the raters think about revising the rubric?

3. Methodology

3.1 Participants

Twenty Japanese university freshmen and sophomores majoring in education with diverse levels of English proficiency (i.e., Common European Framework of Reference for Languages [CEFR] A1 to C1) wrote summary data. The participants under each CEFR level were four in A1 and B2, five in A2 and B1, and two in C1.

Furthermore, the seven raters recruited to evaluate the summary writings had diverse teaching experiences, occupations, and first languages. Two were non-native speakers of Japanese but had high proficiency in speaking Japanese; three were inactive English teachers at universities, whereas the others were students (Table 1). The author of the present study, who is an English instructor of this study's summary writers, did not participate in the evaluation.

When creating or revising a rubric, raters often include expertise and experience (e.g., Chang et al., 2015; Li et al., 2021; Sawaki, 2020). However, as mentioned in Yamanishi et al. (2019), this study reflects the educational setting in high schools, where beginning teachers, experienced teachers, and assistant language teachers (ALTs) are evaluated together, and the rater's attributes are designed to emphasize diversity.

Table 1. Background of the seven raters

	First language	Occupation	Research major	Teaching experience
Rater 1	English	Associate Prof.	Linguistics	15 years
Rater 2	Chinese/English	Assistant Prof.	Linguistics	5 years
Rater 3	Japanese	Ph.D. student/Part-time teacher at university	Linguistics	2 years
Rater 4	Japanese	MA student	Linguistics	None
Rater 5	Japanese	MA student	Linguistics	None
Rater 6	Japanese	MA student	Linguistics	Under English teacher training
Rater 7	Japanese	MA student	Linguistics	Under English teacher training

3.2 Materials

3.2.1 English texts

Twenty students summarized two types of English reading-comprehension tests, all of which were adopted from EIKEN® Pre first grade (see Appendix A) and second grade (see Appendix B). In Japan, EIKEN® is a famous test of practical English proficiency taken by most Japanese students who study English. The number of words in both texts was approximately 341.5. Students were asked to write English summaries of approximately 120 words.

3.2.2 Rubric

Raters used the analytic rubric for summary writing developed by Li (2014), which consists of four evaluation items. The first item is main idea coverage (MIC), which refers to the number of appropriate main ideas selected. The second item is integration (INT), which includes whether the statements in a summary are written in a logical order and have a global interpretation. The third evaluation item is language use (LU), which measures whether the language used in a summary includes a variety of syntax and vocabulary without errors. The final item is source use (SU), which includes whether the summary is written in the writer's own words and whether the information in the summary is included correctly. Previous studies (e.g., Sawaki, 2003; Yu, 2007) have reported substantially high correlation coefficients for multiple measures in assessing summary writing. Additionally, Kato (2022) reported that the correlation coefficient between MIC and INT using universal scores derived from the D-study of multivariate generalizability theory (Brennan, 2001) was 0.94. As the interpretation of these two items is almost the same, in this study, the MIC was omitted from the original rubric, and three items were used (see Appendix C).

3.2.3 Questionnaires

All raters were asked to answer the question "Did you find it difficult to give scores on INT/LU/SU?" using a Likert scale (1 = *did not hesitate at all*, 2 = *relatively did not hesitate*, 3 = *neither*, 4 = *relatively hesitated*, and 5 = *hesitated a lot*). They were also required to answer the open-ended question regarding the reasons behind the ease or difficulty in scoring each evaluation item. The questions above were for RQ1 and RQ2. The raters were then asked to answer close-ended questions for what affected them in making their evaluation decisions and why they chose their particular answers for open-ended questions. Finally, raters were asked to freely give their frank opinions on the rubric's restructuring and answer close-ended questions about its reconstruction. All the questionnaires are presented in Appendix D.

3.3 Data Collection and Data Analysis

Protocol data for English summaries were collected from students at two Japanese universities with diverse levels of English proficiency and majors. Twenty participants were randomly selected based on their CEFR levels. Before their training sessions, the seven raters were asked to comprehend the English texts and write summaries in English. Additionally, they were asked to segment all English sentences in each text into idea units that follow the segmentation rule posited by Kroll's (1977) and divide each idea unit into three levels: "higher" (containing the most important information), "middle" (containing important information), and "lower" (e.g., transition words and conjunctions). They were also required to understand the content of the analytic rubric.

While the raters' training session was conducted for two hours through Zoom using recordings, discussions and explanations were given in Japanese, in which all raters were fluent. Table 2 illustrates the brief procedure used for data collection.

Table 2. Brief procedure for data collection

First step	*Rater training session (2 hours via Zoom in Japanese)*
	(1) Understanding Li's (2014) analytic rubric.
	(2) Confirming main ideas to write summaries.
	(3) Practicing scoring each evaluation item.
	(4) Discussing scoring gaps among raters (for each of three evaluation items) and deciding the middle ground.
Second step	*Individual session*
	(1) Assigning scores to the remaining summaries
	(2) Answering close-ended questionnaire (for RQ1)
	(3) Answering open-ended questionnaire (for RQ2)
	(4) Answering open- and close-ended questionnaires (for RQ3 and RQ4)

Before beginning data collection, the reading material on Li's analytic rubric was distributed to the raters. First, their understanding of Li's analytic rubric was confirmed. Then, the raters discussed and decided on the main ideas to select when writing summaries. They practiced assigning a score for each evaluation item and discussed scoring gaps for each item to reach the middle ground. The author moderated the discussions to maintain focus and allow the student raters, who were more reserved owing to their status, to speak.

After a two-hour raters' training session, the second step included the individual session. The raters were asked to assess the remaining summaries individually and answer close-ended and open-ended questionnaires on assessing summary writing using the analytic rubric. Regarding the close-ended question for RQ1, when evaluating each summary, the raters were asked to answer which evaluation item was the most difficult to score in all the summaries using a 5-point Likert scale (1 = did not hesitate at all, 2 = relatively did not hesitate, 3 = neither, 4 = relatively hesitated, and 5 = hesitated a lot). For RQ2, they were asked to provide the reason for their answers using an open-ended questionnaire throughout the entire evaluation after completing all the evaluations. To investigate RQ3, raters were required to answer the three close-ended questions about what affected them in making their evaluation decision and comment on why they thought that and why they had done it in that way. Finally, for RQ4, raters were asked to answer the close-ended questions and comment on the rubric's reconstruction. All questionnaire contents are presented in Appendix D.

Concerning data analysis, the results of the close-ended questions are presented using simple descriptive statistics, and the qualitative data obtained from the open-ended questions are reported directly.

4. Results

4.1 Research Question 1

The raters scored two different English summaries written by 20 English language learners. Each rater was asked about the difficulty in assessing each of the three items during each student's assessment. Tables 3 and 4 show the average results of the close-ended questionnaire answered by seven raters for each of the 40 different summary assessments as RQ1.

Table 3. Results of descriptive statistics on close-ended questionnaire for RQ1

	Integration (INT)	Language Use (LU)	Source Use (SU)
	Mean (SD)	Mean (SD)	Mean (SD)
R1-R3 (Expertise)	3.02 (1.95)	2.93 (1.98)	2.92 (1.71)
R4-R7 (Novice)	2.14 (0.88)	2.12 (0.83)	2.28 (1.00)
All (*N*=7)	2.52 (1.37)	2.47 (1.36)	2.56 (1.26)

Note. 1 = *did not hesitate at all*, 2 = *relatively did not hesitate*, 3 = *neither*, 4 = *relatively hesitated*, and 5 = *hesitated a lot*.

The results in Table 3, which indicate the answers from the same attribute groups, reveal that no particular evaluation item was difficult. However, as shown in Table 4, each rater seemed to have a different impression of the evaluation items.

Table 4. Results of descriptive statistics on close-ended questionnaire for RQ1 (individual answers)

	Integration (INT)	Language use (LU)	Source use (SU)
Rater 1 (NES [15])	5.00	5.00	4.53
Rater 2 (CES [5])	1.11	1.06	1.12
Rater 3 (JET [2])	2.94	2.74	3.11
Rater 4 (JS [M1])	1.61	1.53	1.59
Rater 5 (JS [M1])	1.23	1.36	1.31
Rater 6 (JS [B3])	2.59	2.45	2.81
Rater 7 (JS [B3])	3.14	3.14	3.42
Mean	2.52	2.47	2.55

Note. NES = Native English Speaker, CES = Chinese English Speaker, JET = Japanese English Teacher, JS = Japanese Student. The numbers in brackets indicate teaching experience. M1 = Master's first year, B3 = Bachelor's third year. Numbers in parentheses [ ] indicate years of teaching experience.

4.2 Research Question 2

The seven raters' responses to open-ended questionnaire responses to RQ2 are reported below.

Rater 1

Although Rater 1 was a native English speaker and had much experience in assessing summaries written by Japanese university students, he hesitated in deciding the scores for every item for the following reasons (Table 5).

Table 5. Responses from Rater 1

INT
I wanted to ask for help in understanding the descriptors of the assessment items during the individual assessment phase.
LU
The evaluation of LUs was unclear in the case of several copy-pasted phrases in a summary product.
SU
Evaluating the copied sentences/phrases required effort, and it was particularly difficult to assess the content's accuracy. During training, the rater experienced difficulties as the level of acceptance of the copied sentences/phrases differed from that of other raters. I also complained about the difficulty in identifying the main ideas during training.

Rater 2

Rater 2 was a Chinese speaker of English who could speak fluent Japanese and had experience in teaching and learning summary writing for the TOEFL iBT. Of the seven raters, he seemed to have the least hesitation in determining the score for any evaluation item (Table 6).

Table 6. Responses from Rater 2

INT
I was often confused between 2 or 3 points owing to the descriptor's wording.
LU
I was often confused between 3 or 4 points because of the descriptor's expression.
SU
Compared to the other evaluation items, it was easier to decide the SU scores.

Rater 3

Rater 3 was a Japanese speaker of English with experience in teaching English at a senior high school and university in Japan. He was moderately hesitant in determining the scores for all the evaluation items (Table 7).

Table 7. Responses from Rater 3

INT
Scoring the items was not difficult because I understood the text and its context.
LU
Judging grammatical/lexical accuracy was relatively challenging as I am not a native English speaker.
SU
Content accuracy could be judged because I understood the content and its context; however, evaluating whether the sentences/phrases were copied required effort.

Rater 4

Rater 4 was a Japanese learner of English and a Master's student researching Japanese linguistics. She did not seem hesitant in deciding the scores for any item (Table 8).

Table 8. Responses from Rater 4

INT
It was difficult because we had to compare the text to see if it contained enough content, but we became accustomed to it with practice.
LU
Grading was easy because I simply had to check for grammatical/lexical errors.
SU
I had memorized some sentences and phrases in the text; thus, scoring was not difficult.

Rater 5

Rater 5 was a Japanese learner of English and a Master's student researching psycholinguistics. She did not seem hesitant in deciding the scores for any item (Table 9).

Table 9. Responses from Rater 5

INT
It was difficult to evaluate the summary's quality while simultaneously examining whether it was coherent and contained sufficient or insufficient information to be written in the summary, but I gradually became accustomed to it.
LU
It was not difficult to evaluate English expressions as the difference between those who wrote as much text as possible, those who paraphrased and made errors in English, and those who could paraphrase in perfect English was clear.
SU
It was not difficult to evaluate the texts because we could immediately recognize when a phrase gradually became familiar in the text.

Rater 6

Rater 6 was a Japanese learner of English. She was a junior majoring in linguistics and on a teacher training course to become a high school English teacher. She seemed hesitant in deciding the scores for every item (Table 10).

Table 10. Responses from Rater 6

INT
The raters' training helped me (us) decide on the main idea to some extent, so the evaluation was relatively easy.
LU
It was relatively easy because I only had to judge whether the text contained grammatical/lexical errors.
SU
It was difficult to determine how many words should be replaced by my own to obtain a rating of 5 or, conversely, 1.

Rater 7

Like Rater 6, Rater 7 was a Japanese speaker of English. She was a junior majoring in linguistics and on a teacher training course to become a high school English teacher. She also seemed hesitant in deciding the scores for every item.

Table 11. Responses from Rater 7

INT
This was rather difficult because I simultaneously had to judge whether I got the points and whether the sentence connections were smooth.
LU
In some areas, I was not sure whether the expressions were appropriate because I am not a native English speaker.
SU
It was difficult to determine the degree to which the text was transcribed verbatim.

4.3 Research Question 3

For RQ3, The raters were asked to answer close- and open-ended questionnaires. The results are in Table 12.

Table 12. Results of close- and open-ended questionnaires

	R1	R2	R3	R4	R5	R6	R7
1) Which of the evaluation items do you consider the most important?	SU	INT	INT	INT	INT	INT	INT
Reasons							
R1							
If you do not write in your own words (too much copy and paste), I have no idea if you have understood the content, and I cannot determine your English ability at all. I think the most basic and important question is whether the copy and paste is excessive.							
R2							
The difference in INT scores should be clarified.							

R3
SU is important, but I still thought INT was more important in terms of "integration."
R4
When evaluating a summary, whether it is a copy of the text or not or whether it contains grammatical errors is not important; instead, the relevant aspect is whether the summarizer has captured the main idea of the text and structured the summary accordingly.
R5
I believe that the ability to successfully connect sentences and include information without excesses or deficiencies is the ability to summarize.
R6
I think that the most important thing in summarizing is to ensure that the content is well chosen and structured without contradictions.
R7
Because this item allows me to judge whether I am reading the text accurately.

	R1	R2	R3	R4	R5	R6	R7
2) Which items did you evaluate by looking at the text over and over when evaluating the summary?	INT & SU	INT & LU	INT	INT & SU	INT LU SU	SU	SU

Reasons
R1
Without checking the source text, it is moderately difficult to make an evaluation decision.
R2
Because the descriptor definitions between 2 and 3 points and those between 4 and 3 points are vague.
R3
Because when the order of sentences seemed unnatural, it was necessary to refer to the sentences in the source text to determine if they were phrased awkwardly.
R4
INT: Because unlike other items, the evaluator needed to understand the content of the summary text and compare it to the text.
SU: Because it is necessary to look closely at the text to detect any textual omissions.
R5
INT: Because the coherence of the summary text itself was evaluated.
LU: Because the correct or incorrect use of vocabulary has nothing to do with what is in the text.
SU: Because paraphrasing is using expressions that are not in the text.
R6
Because it was necessary to check the extent to which the text was copied in detail.
R7
Because I thought it was necessary to carefully compare whether the expressions were the same as those in the text.

	R1	R2	R3	R4	R5	R6	R7
3) When evaluating the summary, which items did you evaluate by looking at the rubric over and over?	INT & SU	INT & LU	INT & SU	INT & SU	LU	SU	INT

Reasons
R1
I have not seen any of them; the rubric is too vague to be of much help.
R2
Because the descriptor definitions between 2 and 3 points and those between 3 and 4 points are vague.
R3
Both INT and SU seem complicated because they use two criteria to evaluate, and I was often confused regarding how to evaluate them.
R4
Because I went into the evaluation with an understanding of the perspective to be evaluated in each item, but from time to time, I had to check what the score of each level was.
R5
If the vocabulary level of the person in front of me was high, I would have inadvertently given a low score even though they had made no particular mistakes.
R6
Because SU was the most unfamiliar item, I sometimes lost track of what kind of evaluation item it was.
R7
I often looked at the rubric to figure out what score to give if the main points were covered.

	R1	R2	R3	R4	R5	R6	R7
4) When evaluating the summary, did you refer to the segmented idea unit table? (Yes = 1, No = 2)	No	No	No	Yes	No	Yes	Yes

Reasons
R1
The segmented idea unit was not particularly helpful.

R2
It was not necessary.
R3
When I looked at the actual English text, it was hard to see the practicality of the segmented idea unit because few summaries were written in the exact order in which they were written in the segmented idea unit.
R4
The summary needs to be compared with the text to see if it contains the text's contents without excesses or deficiencies.
R5
Because it was generally the same as the part I chose.
R6
Because I thought it would be possible to have an objective standard and make an accurate evaluation rather than think by my own judgment standard alone.
R7
To value objective judgment rather than my own subjective judgment.

4.4 Research Question 4

For RQ4, the raters' questions and answers are in Table 13.

Table 13. Results of close- and open-ended questionnaires on the rubric's reconstruction

Do you think there should be five items, adding "logical" and "global" in INT and "information (of text) accuracy" and "paraphrase" in SU, instead of only three items of INT, LU, and SU?	Mean	SD
	3.57	0.79
Reasons		
R1 (Answer = 4) I think it is best to keep the evaluation criteria as separate as possible.		
R2 (Answer = 4) It is useful if the aspects to be evaluated are several to ensure that a comprehensive judgment can be made.		
R3 (Answer = 4) I feel that if there were five items, each would be easier to intuitively understand and score.		
R4 (Answer = 2) Even if we increase the number of items, I think that the same evaluation would be given among the items that have been integrated this time.		
R5 (Answer = 4) As for "logical" and "global," many sentences were coherent but lacked necessary parts. I think the ability to write sentences and the ability to summarize are two different things.		
R6 (Answer = 3) As this was my first time conducting a summary evaluation, I am not sure of the standard, but I think that if the number of items was reduced to five, the differences between the items would become more difficult to distinguish.		
R7 (Answer = 4) Some summaries were well structured or written in one's own words, but others were different from the main text or exaggerated the content.		
Do you think it is acceptable to use the evaluation criteria to score INT, LU, and SU, like INT for "logical" and "global," and SU for "information (of text) accuracy" and "paraphrase," as shown in the present report?	Mean	SD
	3.00	0.82
Reasons		
R1 (Answer = 3) It is difficult to evaluate logical/global. It would be easier to evaluate if the evaluation items were separated.		
R2 (Answer = 3) What puzzled me was the question of what to do in the case of a student who had summarized the original text but had not understood the content. Therefore, I thought it would be better to separate the logical and global.		
R3 (Answer = 3) I do not have a specific item name in mind, but I thought three items would be fine if the item names were more easily understood.		
R4 (Answer = 4) Even if the items are subdivided, I think that the items integrated in this rubric will be rated the same.		
R5 (Answer = 2) Even if the result is the same evaluation, I think it can be considered a little more in terms of tidying up the evaluator's brain.		

R6 (Answer = 4)
I think the summary has the essentials…However, I thought it would be better to be more specific about what is meant by "accurately written" in the SU item.
R7 (Answer = 2)
In some summaries, the content was different from the text or exaggerated, so I think the evaluation criteria for "information (of text) accuracy" are necessary.

3) Please indicate what items you think should be set in the rubric to be created in the future.
R1
(1) paraphrase, (2) information accuracy (and inclusion), (3) language complexity, and (4) grammatical accuracy
R2
It is better to set logical and global separately.
R3
Word limit.
R4
Rather than setting up a new item, we should set priorities among the items and give more weight to the distribution of points.
R5
I think a "content correctness" item is essential because even if the points to be incorporated are correct, if the content is incorrect, it may be written on a hunch.
R6
Whether or not the number of words is appropriate (I thought that a few items were too short)
R7
What defines a copy of the text as such? (e.g., word-for-word, syntactic structure of the sentence)

4) What points, if any, should we pay attention to in the descriptors (explanatory text) when we develop the analytic rubric for summary writing?
Reasons
R1
I think "approximately," "almost," "moderately," "mostly," "good amount," "very," etc. are forbidden. Similarly, "correct," "appropriate," "rich," and "various" are also considered subtle. It would be easier to understand how many the discrepancies with the content are (e.g., two or more instances of reported information that was factually contradictory to the original material).
R2
As it was explained to me in detail at the raters' training session, it was easy to understand.
R3
I thought it was unnecessary to have a logical section in the INT, since it seems that "logicality" and "illogicality" can be evaluated depending on how well or poorly the "integration" is done.
R4
I think it would be easier to evaluate if it were set what items should be given weightage in the evaluation. It would be easier to do if there were examples of what deserves to be evaluated at that level.
R5
As for SU, I think it would be more accurate as a criterion if you could write the percentage of citations in numbers. Comparatively speaking, LU and SU were easy to evaluate since they were based on a point reduction method, but for INT, it is difficult to tell what a point reduction is if something is not done. I think it would be better with criteria for each rather than writing continuously through each stage.
R6
I still think it would be better to note what is accurate in the SU's "Is it accurately written?"
R7
None in particular.

5) Please feel free to write your impressions after completing this summary evaluation.
Reasons
R1
Once again, I found the rubric difficult and the summary writing especially difficult to evaluate.
R2
What puzzled me was the issue of what to do in the case of students who summarize the original text but clearly do not understand the content. Therefore, I thought it would be better to separate the logical and global.
R3
Judging grammar is still difficult for me, a native speaker, and I am not confident enough in my grade.
R4
In evaluating summaries, the aspects to look at are many, such as grammatical accuracy, vocabulary used, and uniqueness of sentence structure, in addition to whether the content of the summary matches the text. I realized that even summary sentences that receive the same score can be completely different in character.

R5	It was difficult because my knowledge of English is limited; therefore, I could not determine whether this expression was correct or natural based on standards of English correctness.
R6	This was the first time for me to perform this type of summary evaluation work, and I thought it was difficult because judgments often became subjective. Additionally, since the evaluators were not fully aware of the rubric, I was worried whether I was evaluating using the same criteria as everyone else.
R7	I found it difficult to do a summary evaluation of students. First, it was difficult to give a score of 1 or 5. I thought about and could not answer whether minimum and maximum scores should not be raised unless they are perfect, or whether they should be given to make a difference because "the whole group is at this level."

Note. For questions 1 and 2, 1 = *not at all disagree*, 2 = *rather disagree*, 3 = *neither agree nor disagree*, 4 = *somewhat agree*, 5 = *totally agree*

5. Discussion and Conclusion

Table 3 shows that overall and by attribute, the rater did not find any of the items particularly difficult or easy. However, Table 4, which focuses more on individual impressions, shows for Rater 1, a native English speaker and expert teacher with 15 years of English teaching experience, all items were difficult to assess. Specifically, Evaluator 1 stated that the definitions of the descriptive items for each of the INT scores were difficult to understand. In this regard, the rater constantly shifted his/her eyes from one source document to another and to the original text, rubric, and summary as he/she evaluated the INTs with respect to his/her actions when scoring. Additionally, Rater 1 said, "I did not look at any of them. The rubric is vague and not particularly helpful." However, in the actual evaluation, the rater stated that it was difficult to judge because the adverbs used in the descriptions (e.g., "very much," "somewhat") were subjective. Furthermore, as for what items need to be added to improve the summative evaluation rubric, the rater did not mention the need for items that fall under INT.

Regarding LU, the rater was often hesitant to evaluate it in the first place as he seemed to have doubts about evaluating LU for summaries with a large number of copies of the original text. Therefore, the rater suggested that the items "paraphrasing," "language (syntactic structure) complexity," and "grammatical accuracy" are necessary for the analytic rubric to be developed in the future. In other words, since this is only a summary assessment for EFL learners, multiple items on "language use" would allow to assess "language use" as accurately as possible, which is affected by English proficiency.

Regardless of the extent to which SU is discussed in the raters' training, the judgment of the copying percentage varies by rater, and the evaluation of LU is meaningless if the writer does not write English sentences in his/her own words as a precondition for the summary evaluation. Therefore, for Rater 1, the most important item in the summary evaluation is the SU, and this should be included in the rubric to be created in the future (Table 12).

Rater 2, who, like Rater 1, has native-level English proficiency and English teaching experience, did not seem to experience any particular difficulty in evaluating this rubric. However, the rater often felt confused between INT and LU scores, indicating that these items should be included in the rubric. Rater 2 mentioned that the evaluation training was especially thorough and that they were asked several questions. They tried to be as objective as possible about the evaluation concept, which they had been looking at subjectively.

Rater 3 found all the items somewhat difficult; for INT, he was able to remember the context of the text to some extent, and he had to review the rubric several times during the evaluation owing to the two phases of the text. For LU, he suggested that it is important to rely on native speakers or automatic evaluation to judge the details of grammatical correctness and syntactic complexity as the rater is not a native English speaker. However, the rater expressed that it is difficult to judge and grade the copying percentage. For future improvements to the rubric, in addition to setting a word limit, the rater suggested that the descriptor for each rubric score be accompanied by a summary example of when the corresponding score could be assigned.

Raters 4 and 5 did not seem to find all the evaluation items difficult to score. They considered INT more important than any other item, and they often went back to the rubric and source text to ensure that the summary included enough main ideas to evaluate the INT. Rater 5 found that it would be helpful to subdivide the items, so that they could organize the evaluation in their mind as an easy-to-use rubric for future evaluations. Rater 4 suggested prioritizing the items and assigning more weight to those items, rather than assigning the same score to all items. Like Rater 3, Rater 4 also suggested the usefulness of providing item descriptors and accompanying them with examples. Further, Rater 5 noted that it would be more helpful and easier for raters to understand if SU descriptors gave approximate numerical values for the copying percentage rather than using adverbs to describe them.

Rater 6 and 7, who were undergraduate students interested in becoming English teachers, provided particularly useful comments. Unlike Rater 4 and 5, they seemed anxious about scoring all the assessment items, and although the INT assessment was easier than the others owing to their practice and familiarity with the assessment in the raters' training, they both stated that they had repeatedly reviewed and scored the segmented idea unit table, which was discussed as an indicator in the raters' training, to eliminate subjectivity in the assessment. The LU was not as difficult for these two as they both have relatively high English proficiency, but the SU was the most difficult to evaluate. The two respondents seemed to have read the text several times during the SU evaluation to ensure that they could paraphrase the summary. They also discussed the ambiguity of the definition of "content accuracy" as an SU item, although they also mentioned that the copying percentage was unclear. They reported that, in some summaries, the content deviated slightly from the source text, and other summaries had completely different interpretations. As indicated in (5) of Table 13, regarding his impression after completing the summary evaluation, Rater 2 was puzzled as to how to evaluate a summary that did not seem to understand the source text. The SU items tend to focus on the copying percentage as the evaluation, since paraphrasing is an important part of the definition of summarization, but the SU descriptors need to be devised to consider the importance of "content (information) accuracy."

This study highlighted the candid opinions of both expert and novice teachers in the field of education when conducting summary evaluations. As the requests for improvement in the description of evaluation items and the rubric have been many, a tentative analytic rubric for summary writing is expected to be proposed in the future, and this study will serve as a basis for constructing a rubric that is easier for more evaluators to use.

Finally, I will mention Q2. (Which items did you evaluate by looking at the text repeatedly when evaluating the summary?) and Q3. (When evaluating the summary, which items did you evaluate by looking at the rubric over and over?) of RQ3. As shown in Table 12, most of the raters responded that they had to move their eyes the most when evaluating INT and SU. However, it should be noted that, as a limitation of this study, the raters were not asked about the reasons for their behavior. Some interesting remaining issues include how eye movement is measured and whether the need to move one's gaze adds cognitive load to the evaluation process.

Acknowledgments

This study was supported by JSPS KAKENHI Grant Number JP 22K13174. I would like to express my deepest thanks to seven raters and twenty summary writers. I would also like to acknowledge two anonymous reviewers for their invaluable comments.

Authors' contributions

Dr. Makiko Kato was responsible for whole of this study.

Funding

This study was supported by JSPS KAKENHI Grant Number JP 22K13174.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Informed consent

Obtained.

Ethics approval

The Publication Ethics Committee of the Canadian Center of Science and Education.

The journal's policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Data sharing statement

No additional data are available.

References

Alderson, J. C. (2000). *Assessing reading*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732935

Ascención Delaney, Y. (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

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing*, *28*(1), 51–75. https://doi.org/10.1177/0265532210376379

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

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

Chang, S., Inoue, C., & Taylor, L. (2015). Developing rubrics to assess the reading-into-writing skills: A case study. *Assessing Writing*, *26*, 20–47. https://doi.org/10.1016/j.asw.2015.07.004

Cumming, A. H., Kantor, R., & Powers, D. E. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework*. TOEFL Monograph No. MS-22. Educational Testing Service.

Cumming, A. H., Kantor, R., & Powers, D. (2002). Decision making while rating ESL/EFL composition tasks: A descriptive framework. *Modern Language Journal*, *86*, 67–96. https://doi.org/10.1111/1540-4781.00137

Douglas, D., & Selinker, L. (1992). Analyzing oral proficiency test performance in general and specific-purpose contexts. *System*, *20*(3), 317–328. https://doi.org/10.1016/0346-251X(92)90043-3

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Journal of Language Testing*, *26*, 507–531. https://doi.org/10.1177/0265532209340188

Gebril, A. (2010). Bringing reading-to-write and writing-only assessment tasks together: A Generalizability analysis. *Assessing Writing*, *15*(2), 100–117. https://doi.org/10.1016/j.asw.2010.05.002

Gebril, A., & Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. *Assessing Writing*, *21*, 56–73. https://doi.org/10.1016/j.asw.2014.03.002

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*(2), 87–98. https://doi.org/10.1016/j.jeap.2012.11.005

In'nami, Y., & Koizumi, R. (2019). Using EIKEN, TOEFL, and TOEIC to award EFL course credits in Japanese universities. *Language Assessment Quarterly*, *14*(3), 274–293. https://doi.org/10.1080/15434303.2016.1262375

Kato, M. (2021). *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. (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

Kirkland, M., & Saunders, M. (1991). Maximizing student performance in summary writing: Managing cognitive load. *TESOL Quarterly*, *25*(1), 105–121. https://doi.org/10.2307/3587030

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

Li, J. (2021). Examining EFL learners' source text use in summary writing. *Language Teaching Research*. https://doi.org/10.1177/13621688211055887

Li, J., & Wang, Q. (2021). Development and validation of a rating scale for summarization as an integrated task.

Asian-Pacific Journal of Second and Foreign Language Education, *6*(11). https://doi.org/10.1186/s40862-021-00113-6

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

Milanovic, M., Saville, H., & Shuhong, S. (1996). A study of the decision-making behavior of composition markers. In M. Milanovic, & N. Saville (Eds.), *Studies in language testing 3: Performance testing, cognition and assessment* (pp. 92–111). Cambridge University Press.

Plakans, L. (2009a). Discourse synthesis in integrated second language writing assessment. *Language Testing*, *26*(4), 561–587. https://doi.org/10.1177/0265532209340192

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 & B. A. Green (Eds.), *Another Generation of Fundamental Considerations in Language Assessment* (pp. 153–171). Springer. doi: https://doi.org/10.1007/978-981-15-8952-2_10

Shin, S. Y., & Ewert, D. (2015). What accounts for integrated reading-to-write task scores? *Language Testing*, *32*(2), 259–281. https://doi.org/10.1177/0265532214560257

Suto, I. (2012). A critical review of some qualitative research methods used to explore rater cognition. *Educational Measurement: Issue and Practice*, *31*(3), 21–30. https://doi.org/10.1111/j.1745-3992.2012.00240.x

Suto, W. M. I., & Nadás, R. (2008). What determines GCSE marking accuracy? An exploration of expertise among maths and physics markers. *Research Papers in Education*, *23*(4), 477–497. https://doi.org/10.1080/02671520701755499

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. (2002). *Assessing writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

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. https://doi.org/10.1177/0265532207080780

Appendix A

English Source Text 1

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. …….

Appendix B

English Source Text 2

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

Appendix C

Analytic Rubric for Summary Writing (Modified from Li, 2014)

Integration (INT)		
	Logical	***Global***
5 points	The order is illogical	Very nicely put together.
4 points	(The written summary	Well organized.
3 points	must meet this condition to	Reasonably summarized (integrated).
2 points	get 1 to 5 points.)	There is little reordering or integration, and the text lacks coherence as a summary of the text content.
		Lacks coherence as a summary of the textual content with little or no reordering or integration.
1 point		Inappropriate merging (e.g., English is partially merged, but the meaning is not understood) without removing redundant parts of the text, and there is no convergence as a summary of the text content.
0 points	No integration or merging at all.	
Language Use (LU)		
5 points	Well structured, with appropriate vocabulary selection.	
4 points	The text is well structured, with mostly appropriate vocabulary selection (no significant grammatical or lexical errors).	
3 points	Some grammatical or lexical errors that interfere with semantic interpretation.	
2 points	Inappropriate word choice, word formation, and sentence structure are common.	
1 point	Significant and frequent errors in sentence structure and usage. (Shows poor understanding of vocabulary and grammar.)	
0 points	Many grammatical errors that are unintelligible OR nothing is written.	
Source Use (SU)		
	Information accuracy	***Paraphrase***
5 points	It is written accurately.	Mainly written by the summarizer, who changed the word and sentence structure.
4 points	(The written summary must	The summarizer changed the structure of words and sentences, mostly by himself/herself.
3 points	meet this condition to get 2	Basically written by the summarizer, changing the structure of words and sentences.
2 points	to 5 points.)	Some of the words and sentence structures were changed by the summarizer.
1 point	Mainly copied from the original.	
0 points	Completely copies the original.	

Appendix D

Questionnaires

1. When evaluating each summary one at a time, how did you feel about the scoring of each evaluation item?

Please put the numbers referring to your opinions on the right side of the evaluation sheet.

(e.g., 1 = did not hesitate at all, 2 = relatively did not hesitate, 3 = neither, 4 = relatively hesitated, 5 = hesitated a lot).

ID	Student summaries	Scores			Questionnaire answers		
		INT	LU	SU	INT	LU	SU
1	New Yok City have seeking a way to reduce traffic. In 1971, the mayor John Lindsay stoped entering cars on Manhattan's ……..						

2. Please feel free to answer the reason why you felt the way about each item throughout the entire evaluation after completing all the evaluations. (Ex: SU was difficult to evaluate because it was difficult to determine the percentage of text copying.)

3. Please answer the following questions regarding what affected you to have made your evaluation decision, and please comment on why you thought/did so.

1) Which of the evaluation items do you consider the most important?

2) Which items did you evaluate by looking at the text over and over again when evaluating the summary?

3) When evaluating the summary, which items did you evaluate by looking at the rubric over and over?

4) When evaluating the summary, did you refer to the segmented idea unit table? (Yes or No)

4. Please answer the following question regarding the rubric's reconstruction and comment about it freely.

1) Do you think the items should be five, adding “logical” and “global” included in INT and “information (of text) accuracy” and “paraphrase” included in SU, instead of only three items of INT, LU, and SU?

2) Do you think it is acceptable to use the evaluation criteria to score INT, LU, and SU, like INT for “logical” and “global,” and SU for “information (of text) accuracy” and “Paraphrase,” as shown in the present report?

3) Please indicate what items you think should be set in the rubric to be created in the future.

4) What points, if any, should we pay attention to in the descriptors (explanatory text) when we develop the analytic rubric for summary writing?

5) Please feel free to write your impressions after completing this summary evaluation.