English Language Teaching; Vol. 15, No. 9; 2022
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

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

Makiko Kato[1]

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

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

Received: July 27, 2022 Accepted: August 21, 2022 Online Published: August 22, 2022

doi: 10.5539/elt.v15n9p82 URL: https://doi.org/10.5539/elt.v15n9p82

Abstract

English teachers, especially those who teach summary writing to students with relatively lower proficiency in English face difficulty in teaching summary writing and while assessing their students' performances. In the classroom context, an analytic rubric is pedagogically more helpful than a holistic rubric because the teacher can confirm the strengths and weaknesses of their students' summary performance and the students can receive constructive feedback (Yamanishi et al., 2019). This study examined the practicality of the analytic rubric which consisted of four rating scales, including language use, by investigating seven in-service English teachers' honest assessment of 160 summaries of Japanese private university students who are inexperienced in writing English summaries and have lower-intermediate proficiency of English. Furthermore, this study examined the dependability of the analytic rubric using multivariate generalizability theory (Brennan, 2001). The results showed that assessing language use and judging summaries, copied, to a lesser or greater extent, from the source text was difficult because of diverse linguistic errors and the use of paraphrasing was lacking. Therefore, it is necessary to define the gravity of language errors and that of copying in more detail to develop a rubric that suits to assess the summaries written by English learners with lower-intermediate level of English.

Keywords: summary writing, analytic rubric, multivariate generalizability theory, lower-intermediate EFL learners

1. Introduction

Summary writing is considered an important academic skill that university students should possess, regardless of their varied educational backgrounds. Though the demand for teaching summary writing is increasing, English summary writing teachers, especially, those who teach students with lower proficiency in English must face difficulty in instructing because summary writers are required to possess summarizing skills, such as gisting the information and good vocabulary in addition to reading and writing abilities. In parallel with the difficultly in instructing, assessing summaries is difficult for teachers since several dimensions of rating are included in the scoring rubric of summary writing. Regarding the scoring rubric of summary writing, holistic scoring is commonly used for large scale assessments such as high-stakes test and placement test. However, analytic scoring rubric is more helpful not only for teachers to confirm the strengths and weaknesses of their students' summary performances, but also for summary writers to receive feedback (Yamanishi, Ono, & Hijikata, 2019). Therefore, though time and labor constraints are a problem, analytic rubric for summary writing is pedagogically more helpful and ideal for English learners.

Several scholars have developed an analytic scoring rubric for assessing summary performances (Li, 2014a; 2014b; Sawaki, 2003), but it is still in progress and concerns remain over some criteria are measuring the same thing. This study is a part of the future project that defines the dimensions and develops the rubric for summary writing as a classroom task. Among the existing rubrics, the one that seems to be suitable for practical use is adopted, and the reliability of the rubric will be verified by using the multivariate generalizability theory (MG theory). This study adopted the analytic rubric proposed by Li (2014a, 2014b) because the rubric is for assessment of summaries written by learners of English as a Foreign Language (EFL) focusing on the quality of summary based on the theory of text comprehension (Kintsch & van Dijk, 1978; van Dijk & Kintsch, 1983) and

it includes the dimension of language use. For summaries written by EFL learners, assessment of language use is necessary because it must affect the quality of summary performances. The author believes that it is necessary to consider practical scales which in-service teachers could easily use. This study aims to confirm that an analytic rubric is practical for in-service Japanese English teachers to assess summary writing, and examine the reliability of the rubric for measurement of Japanese EFL summary writers. Therefore, the problem of facing difficulty in deciding scores of analytic items should be cleared by collecting raters' honest assessment.

2. Literature Review

2.1 Assessing Summary Writing

As mentioned above, an analytic rubric for summary writing has been developed by several scholars (e.g., Li, 2014a; 2014b; Rivard, 2001; Sawaki, 2003). Rivard (2001, p. 174) proposed an analytic scale consisting of 10 components: four components evaluated summary products focusing on the content of the original text (i.e., main ideas, secondary ideas, fidelity to the text, and integration of ideas); five components focused on issues related to the language of the target text (i.e., organization, style, language usage, objectivity, and holistic writing score); and the last component included a score for the combination of content and language (i.e., efficiency).

Another rubric with analytic scales was proposed by Sawaki (2003, p. 285), who analyzed summary products focusing on the evaluation of main idea coverage, main idea accuracy, and integration. She defined main idea coverage as "the extent to which main ideas are covered," and main idea accuracy as "the extent to which representation of main ideas is accurate." Integration was defined as "the extent to which the information in the text is presented succinctly by using strategies such as deleting unnecessary information, combining information across idea units, reordering information in the text, and by using topic sentences," and effective integration was defined as the use of integration at the right places and without distortion of the original meaning.

Li (2014a) investigated the role of reading and writing in summarization, which showed that writing proficiency contributed more to summary performances compared to reading skill proficiency. Li (2014a) suggested that the important elements between reading and writing should be applied in the scoring process and emphasized to allocate more weight to writing over reading in the scoring rubric. Li's (2014a, 2014b) rubric for EFL summary writers to assess the quality of a summary is based on the theory of text comprehension (Kintsch & van Dijk, 1978; van Dijk & Kintsch, 1983) and involves grammatical/lexical accuracy (i.e., language use (LU)) and content accuracy/text originality (i.e., source use (SU)) in addition to main idea coverage (MIC) and integration (INT) as the crucial skills for constructing macro structures (Sawaki, 2003). Each of the four components are scored on a 0–5-point scale (Appendix A). Regarding the first component, MIC, generally considered the central concern of a good summary (Li, 2014a; 2014b); the raters evaluate a summary focusing on the proportion of main ideas selected into the written summary. The INT component includes two phases; whether the statements in a summary are written in logical order; and whether the statements in a summary have a global interpretation. As for LU, a variety of syntax, grammar and vocabulary are the central criteria to evaluate. The final SU component refers to the accuracy and coherent use of the information in the text, and has two phases; whether the summary is written in the writer's own words; and whether the information in the summary is included correctly.

Several rubrics for assessing summary writing have been proposed by numerous researchers to assess the quality of summaries accurately, but evaluating them individually based on the produced summary is required because several skills and abilities are involved in writing a summary. Second language (L2) ability influences the quality of the summaries produced by summary writers. In particular, summary writers with low L2 proficiency are likely to make grammatical/lexical errors that hinder the interpretation more than those with higher L2 proficiency. Hence, while discussing and developing an analytic rubric that suits EFL summary writers with any level of English proficiency, assessing grammatical/lexical accuracy should factor into the quality of their summary. Using analytic rubrics, summary performances can be evaluated from multiple perspectives. However, one problem that will arise is that measurement errors increase as the number of measurement items increases. One of the methods to solve this problem is generalizability theory (G theory), which confirms the reliability of the evaluation.

2.2 Multivariate Generalizability Theory

The dependability of rating is generally examined in cases where scores are assigned by multiple raters. Dependability is defined as "the accuracy of generalizing from a person's observed scores on a measure to the average score that the test user would be equally willing to accept" (Shavelson & Webb, 1991). G theory (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991) can be used for examining the dependability of scores, and is defined as "the extension of classical testing theory" (Graham,

Hebert, Sandbank, & Harris, 2016, 73) and "a powerful framework that can assist in analyses of test data based on constructed-response items such as the performance-based assessment tasks above as well as those based on selected-response items (e.g., multiple-choice reading comprehension test items)" (Sawaki & Xi, 2019, 30).

Univariate G theory is used when the dependability of a single measurement is examined, while multivariate generalizability theory (MG theory) (Brennan, 2001) is used for analytic scales (Grabowski & Lin, 2019). The benefit of MG theory is that "it can be used to model multiple measures simultaneously as part of an investigation into the relative contribution of multiple sources of variation to the total score variability," so the indices of composite score dependability and the correlation between components in a composite (i.e., in the form of covariances and universe-score correlation) (Grabowski & Lin, 2019, 56) can be seen. According to Bachman (2004), a universe-score is a much more powerful notion than the concept of an observed score.

MG theory includes two studies, generalizability study (G study) and decision study (D study). In the G study, using the concept of variance analysis, the error variance is decomposed into different factors, and the variance component, which is the size of the capability and error variance, is statistically estimated (Brennan, 2001). The information about total score variability for each component in a composite is provided separately in the case of univariate G theory, and the interrelationships among the multiple dimensions represented in the composite can be investigated through covariances and universe-score correlations in the case of MG theory (Shavelson & Webb, 1991; Webb, Shavelson, & Maddahian, 1983). Covariance component estimates between analytic subscales presented in the D study show the directional relationship across multiple sources of variation, and the universe-score correlations indicate the magnitude of the relationship between the subscales for the object of measurement (Grabowski & Lin, 2019). Universe-score correlations are disattenuated correlations like true score correlations in classical test theory, but the effects of the sources of error are not contained. Whereas, covariance component estimates and universe-score correlations are useful for the evidence of validity because the dependencies between the variables across the multiple sources of variation can be presented. As previously mentioned, a true score with the measurement error removed is derived, and it can be expected that more accurate reliability can be measured. Moreover, D study can estimate variations of the coefficients along with different settings (e.g., different number of items, raters, and occasions) for exploring optional test administration conditions with high reliability (Bachman, 2004a; Brown & Hudson, 2002). In other words, D study is useful to determine the number of items, occasions, and raters, to include in a test to establish reliability.

3. Research Questions

This study addresses the following five research questions, using MG theory. RQ 1: What are the problems in facing the difficulty of giving scores on each item of Li's (2014a; 2014b) rubric? RQ 2: What is the relative contribution of multiple sources of variation (i.e., test takers and raters) to the total score variability in the summary writing test for each of the four subscales in the analytic rubric? RQ 3: How dependable are the scores for each criterion and at the composite score level? RQ 4: To what extent are the components of English summary performance (i.e., MIC, INT, LU, and SU) related in the test? (In other words, how is each criterion in the analytic rubric correlated?) RQ 5: In the case of one rater, how reliable can they be? What is the minimum number of raters for which reliability can be confirmed?

4. Methodology

4.1 Participants

A total of three raters took part in evaluating 160 summaries written by Japanese private university students majoring in engineering, using Li's (2014a; 2014b) rubric. These three raters were Japanese English teachers at universities in Japan.

The English level of summary writers was measured using Nation's vocabulary size test (Nation, 2007; Vocabulary size. com, 2010) because the test correlates with English language proficiency, is free and takes only about 30 minutes to complete. The average vocabulary size of participants was 4887.63 (SD = 1167.02). According to Chujo and Oghigian (2009), less than 5,500-word vocabulary corresponds to around the level of Grade 2 holders of EIKEN, the Test in Practical English Proficiency.

4.2 English Text

Considering that short length was appropriate since participants in this study were prohibited from looking at the text while writing a summary, the text to be summarized was adopted from a 171-word expository text in their textbook (Appendix B). Readability of the text material was measured using Microsoft Word, Vocab-profile (Heatley, Nation & Coxhead, 2002), and Online D-Tools 2.21 – Lognostics (Meara & Miralpeix, 2018). The

results for the text were as follows: Flesch Reading Ease was 54.5; Count of paragraphs was 3; Count of sentences was 11; Type/Token Ratio was 0.57; D-measure was 83.00; and Guiraud's Index was 7.56.

4.3 Rubric

This study adopted Li's (2014a; 2014b) analytic rubric for EFL summary writers, described in more detail previously (Appendix A). The average number of words in all the summaries written by the participants was within tolerance. So, the word limitation perspective is included in the original Li's rubric, but was not observed in this study.

4.4 Design and Data Collection Procedure

A total of 160 English summaries were written by Japanese university students. Before they wrote the summaries, all of them received a lecture on the definition of summary writing including the following points: They should (1) produce a summary, a third as long as the original text; (2) put a thesis statement in the first sentence; (3) include the main ideas; (4) summarize for an audience who does not know the content of the text; and (5) not copy from the source text but paraphrase it in their own words.

Essentially, the participants in this study should have been made aware in advance of the purposes and contents of this research, and should have been asked if they would participate in the research in advance. However, the summary tasks were added to the curriculum of their regular lessons. Regarding the consent form, permission was granted by the English department of the university where this study was conducted with the condition of getting approval from the participants.

Three raters joined the two-hour rater training, and practiced segmenting the sentences into idea units using Kroll's (1977) idea unit definition. They also discussed and defined the five main ideas which should be in the summary. After they practiced giving scores using the rubric, they were asked to evaluate all the summaries accordingly. Then they were also asked to answer the open-ended questionnaire asking about the points where they struggled to evaluate the summaries using the rubric same.

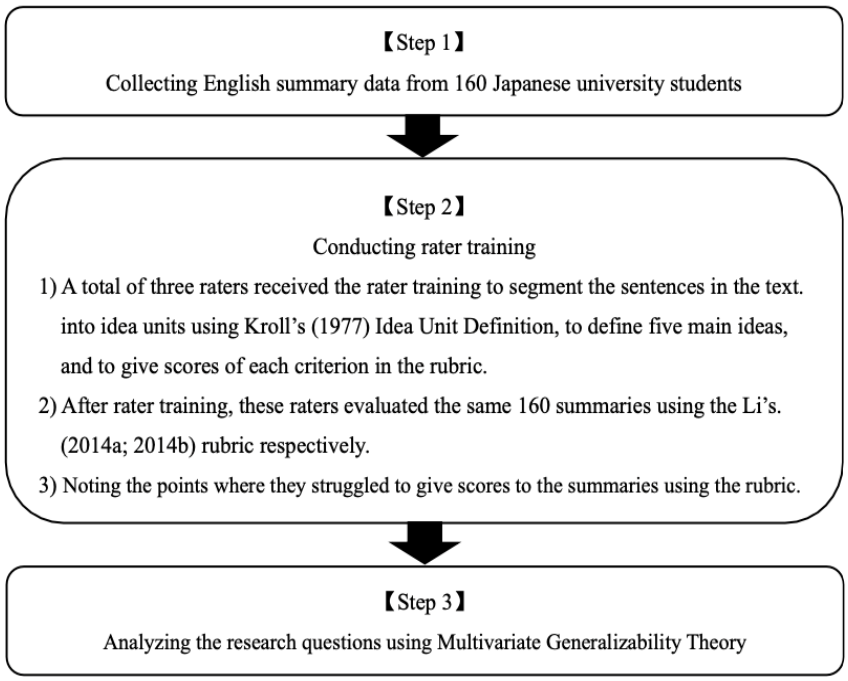


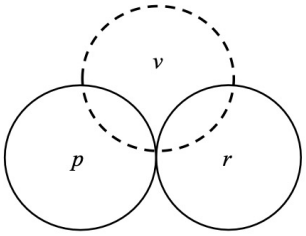
Figure 1. Brief Procedures of Data Collection and Data Analyses (revised from Kato (2021a))

4.5 Data Analyses Procedure

All the data were analyzed using MG theory by mGENOVA, Version 2.1 (Brennan, 2001). The MG theory was adopted in this study to: 1) examine the dependability of the scoring of each item, the variance components are estimated, and the severity of the evaluation is examined for each evaluation item (for RQ2); 2) examine the degree to which the sources of variation associated with each item covaried with those of other items (RQ 2); 3) examine how reliable the scores of individual items (i.e., MIC, INT, LU, and SU) and compound scores are (RQ 3); 4) examine the extent to which the components of L2 summarizing ability (i.e., MIC, INT, LU, and SU) are related in the test (RQ 4); and 5), examine estimate variations of the coefficients along with different settings

(i.e., different number of items and raters) (RQ 5). In MG theory, the concept of dependability is applicable to both norm-referenced and criterion-referenced score interpretations for making relative and absolute decisions, respectively. In the case of this study, the criterion-referenced score interpretation (i.e., absolute decision) is adopted since the summary is compared with the criteria and the quality of the summary is assessed using the stage of each item in the rubric (Hirai, 2018, 76).

In this study, first of all, the objects of measurement, person ability (p) and a facet of measurement, raters (r), were defined, and the design of a single-facet with a single random facet ($p^\bullet \times r^\bullet$) was adopted. By modeling r, the effects of rater severity on a set of ratings each participant earns can be also observed. A superscript filled circle $^\bullet$ refers to a facet crossed with multiple dependent variables (i.e., the level of the fixed facet) (Brennan 2001). In the $p^\bullet \times r^\bullet$ design, each level of r is associated with every level of v (i.e., students who wrote a summary evaluate each of a set of items with respect to MIC, INT, LU, and SU). The random facet r is crossed with the fixed facet v, and p is crossed with both facets. As shown in Figure 2, the multivariate design is $p^\bullet \times r^\bullet$, its univariate counterpart is $p \times r \times v$, there exist a variance components $p \times r$ design for each level of v, and a covariance components $p \times r$ design for each pair of levels of v (Brennan, 2001). Variance components of the following source of variability are also examined using mGENOVA, Version 2.1 (Brennan, 2001); person (p), raters I, person by raters ($p^\bullet \times r^\bullet$), undifferentiated error with the interactions of all objects (pr, e). Subsequently, in D study, as the second step of MG theory, variance components and covariance components are used to examine the dependability using index of dependability (i.e., phi (Φ)). Moreover, universe-score correlations between the four items of Li's (2014a; 2014b) rubric are also examined.



p = person (n =160)

r = raters (n = 3)

v = (n = 4 (i.e., evaluation items; MIC, INT, LU, and SU))

	v_1			v_2			v_3			v_4		
p	r_1	r_2	r_3	r_1	r_2	r_3	r_1	r_2	r_3	r_1	r_2	r_3
1	X	X	X	X	X	X	X	X	X	X	X	X
:	:	:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:	:	:
n_p	X	X	X	X	X	X	X	X	X	X	X	X

Figure 2. Representations of $p^\bullet \times r^\bullet$ Design (revised from Brennan 2001, 277)

5. Results and Discussion

Descriptive statistics for the scores of each criterion scored by three raters are as follows: 2.75 for MIC (SD = 0.87); 2.63 for INT (SD = 0.81); 2.79 for LU (SD = 0.57); and 3.27 for SU (SD = 0.91). Three trained raters gave scores to 160 summaries and the inter-rater reliability measured by the Cronbach alpha were: α = .89 for MIC, α = .86 for INT, α = .67 for LU, and α = .81 for SU.

To examine the first research question, the honest assessment on using the rubric from three raters were examined. Table 1 presents the qualitative results.

Examining the assessment of the raters in Table 1, it seems that it was not difficult to evaluate MIC and INT since they decided the points to be selected as the main ideas after a long discussion in rater training. Meanwhile, LU and SU scoring was difficult even though the raters received training. The summaries included various types of language errors, and their evaluation was difficult by judging the error ambiguity. In addition, some raters might have given lenient LU and SU scores as the summarizer seemed to have made efforts to partially replace words to avoid directly copying from the text, as instructed.

Table 1. Comments on Using the Rubric from the Raters

Items	*Comments*
MIC	• There were no problems in scoring MIC.
INT	• I felt that the INT score will almost be the same as MIC.
	• It was difficult to determine whether conjunctions and transition words were used effectively to arrange the propositions logically.
LU	• I was confused as to whether the LU and INT scores should be deducted when conjunctions or transition words that connect sentences were used incorrectly.
	• I was troubled with the degree of the LU score when the sentence/phrase was partially copied.
	• It was difficult to score LU because of too many linguistic errors in the summaries. Although there were no major grammatical errors, I was confused if I should make a strict evaluation when there were many spelling mistakes.
SU	• Judging the degree of copying was difficult.
	• When there was an easy-to-understand language error (e.g., tense error, misspelling), it was difficult to evaluate the accuracy of the content.

To examine the second and third research questions, all data scored by the raters were analyzed using the MG theory. Table 2 shows the results of estimated D study variance components and their proportions in each case. The ratios of fluctuation factors of all items MIC, INT, LU, and SU were the highest in persons (p) (i.e., 87.17 %, 84.85 %, 64.71 %, and 72.04 %). In all four items, persons (p) explained the fluctuation of the score, not rateI(r), which was a desirable result because the results represented the difference in participants' performances, not the severity of evaluation among the raters. The index of dependability (Φ) of composite score was $\Phi = .82$, and that of each item was as follows: $\Phi = .89$ for MIC, $\Phi = .86$ for INT, $\Phi = .66$ for LU, and $\Phi = .72$ for SU. Raters seemed to understand how to rate MIC and INT, but they seemed to still struggle to give LU and SU scores even though they received rater training.

Table 2. Estimated D study Variance Components and Their Proportions ($p^\bullet$ x $r^\bullet$ Design)

	Estimated variance component			
Source of variation	MIC	INT	LU	SU
Person (p)	0.68	0.56	0.22	0.67
	(87.17 %)	(84.85 %)	(64.71 %)	(72.04 %)
Rater (R)	0.00	0.00	0.01	0.10
	(0%)	(0 %)	(2.94 %)	(10.75 %)
pR,e	0.10	0.10	0.11	0.16
	(12.82 %)	(15.15 %)	(32.35 %)	(17.20 %)
Total	0.78	0.66	0.34	0.93
Φ	0.89	0.86	0.66	0.72
Composite Φ	0.82			

Subsequently, Table 3 shows the results of examining the degree to which the source of variation associated with each item covaried with those of other items. The covariance component estimates clarify the contribution of true differences in test-taker ability, rater severity, and error to the observed score covariance (Grabowski & Lin, 2019). The interpretation of the test-taker ability having the largest proportion of covariance for all items indicates that those who scored high for any one item also obtained high scores for the other items. The effect besides the person effect, such as the rater effect and measurement error were generally minor or close to zero (Grabowski & Lin, 2019), and it can be interpreted that there was little contribution of these item covariances to the total observed score covariance. Accordingly, the proportions of variance and covariance indicated by the person effect in all items were the largest among other sources of variation, so these findings support the dependability of the analytic items more precisely.

Table 3. Variance and Covariance Component Estimates for the Four Items

Source of variation	Estimated variance-covariance component				
	items	(1) MIC	(2) INT	(3) LU	(4) SU
Person (*p*)	(1)	0.68	0.93	0.58	-0.03
	(2)	0.58	0.56	0.61	0.01
	(3)	0.22	0.21	0.22	-0.58
	(4)	-0.02	0.01	-0.22	0.67
Raters (*R*)	(1)	0.00			
	(2)	0.00	0.00		
	(3)	0.00	0.01	0.01	
	(4)	0.01	0.01	0.02	0.10
pR,e	(1)	0.09			
	(2)	0.01	0.09		
	(3)	0.01	0.01	0.11	
	(4)	0.02	-0.00	0.02	0.16

Note. Variance component estimates are underlined on the diagonal; covariance component estimates are on the off-diagonals in plain text.

Given the correlation between the components of L2 summarizing skill (i.e., MIC, INT, LU, and SU), Table 4 shows the universe-score correlation with one another. The universe score refers to the true score (Hirai, 2017) in classical test theory, which is the average of the scores that can be obtained by a test-taker, considered to be the same an infinite number of times (Hirai, 2018). Among all the rating scales, the universe-score correlations that range from -1 to 1 clarified that test-takers who obtained a high score for one item also obtained a high, or at least relatively high, score for another. As shown in Table 4, a strong positive correlation exists between the universe-score of MIC and that of INT (r = 0.94). Furthermore, there were also medium positive correlations between MIC and LU (r = 0.57) and between INT and LU (r = 0.34). However, no correlations exist between MIC and SU (r = - 0.03) and between INT and SU (r = 0.02). Regarding the correlation between LU and SU, there was a medium negative correlation between LU and SU (r = - 0.60).

The results of the universe-score correlations show that MIC/INT and LU/SU have a strong relationship, respectively. The reason behind negative correlation between LU and SU was that the raters presumably did not deduct points for the sentences/phrases close to copied. Some raters reported that it was difficult to evaluate the LU score since judging the degree of copying was difficult. Considering the above, developing a rubric that can accommodate any levels of student's summary evaluation will require a focus on paraphrases and a more detailed definition of language use descriptors.

Table 4. Universe-Score Correlations

Second rating session			
MIC	INT	LU	SU
1.0			
0.94	1.0		
0.57	0.34	1.0	
-0.03	0.02	-0.60	1.0

Table 5 illustrates the change in phi coefficients depending on the number of raters. One rater was enough to achieve .70 in MIC and INT, yet, approximately four raters were needed to evaluate LU and three raters were needed to evaluate SU. Therefore, as reported by the raters (see Table 1), they tried to score MIC after they discussed and defined the points to be selected as the main ideas in the rater training in advance, so it was easy for them to give the MIC score. As for INT, reliability can be obtained with scoring only by one person; however, it seems the raters found it difficult to determine whether conjunction and transition words were used effectively to arrange propositions logically. In some cases, a single rater may not be enough to obtain reliability. With regard to LU, it was found that more detailed discussions among the raters are needed on how to evaluate the

degree of various types of errors. In addition, it is considered that the low reliability of SU was due to the degree of copying and the difficulty of the evaluation criterion.

Table 5. Decision Study of Raters for Each Criterion ($p^{\bullet}$ x $r^{\bullet}$ design)

	1 Rater	2 Raters	3 Raters	4 Raters	5 Raters	6 Raters
Main idea coverage	0.72	0.84	0.89	0.91	0.93	0.94
Integration	0.67	0.80	0.86	0.89	0.91	0.92
Language use	0.39	0.56	0.66	0.72	0.76	0.79
Source use	0.46	0.63	0.72	0.77	0.81	0.84
Composite Phi	0.61	0.76	0.82	0.86	0.89	0.90

6. Conclusion

In this study, an existent analytic rubric consisting of four important dimensions for summary writing was used to examine its practicality for in-service EFL teachers and its reliability. This study adopted Li's (2014a; 2014b) rubric and three in-service Japanese English teachers rated 160 summaries written by Japanese private university students using the rubric.

Results of examining the evaluation reliability using MG theory were consistent with the content of this report. In this study that adopted a short source text, only one rater was necessary for scoring MIC and INT but no less than three or four raters were needed to assess LU and SU.

LU and SU, which were of concern through the evaluation in this study, were also strongly negatively correlated in universe-score correlations. As mentioned by Brown and Day (1983), learners with lower English proficiency tend to copy when they write a summary. Though participants in this study tried to use synonyms and combine parts of the text to avoid copying, paraphrasing might have been difficult for them since the original text was brief.

In a recent study of the development of summary writing for Japanese EFL students, Yamanishi, Ono and Hijikata (2019) adopted the perspective of paraphrasing quantitatively and qualitatively, and developed an analytic rubric consisting of content, paraphrase (quantity), paraphrase (quality), and language use in addition to the overall quality as a holistic rubric. Their study reported a high positive correlation between the overall quality score and the four derived scores, and a positive correlation between the overall quality score and the commonly used ETS's holistic rubric for summary writing. Another recent study on the rubric for EFL summary writers by Sawaki (2020), examined the functioning of two types of rating scales, content point score and holistic summary content rating scale called Integration, in addition to language quality rating scale. Unlike Yamanishi, Ono and Hijikata (2019), Sawaki (2020) included a perspective of paraphrasing in the Integration scale due to the practicality and administrative efficiency of the scoring. In their study, the author analyzed/discussed the concept of copying instead of the concept of paraphrasing and reported that it has a large correlation with the score of language use. However, paraphrasing is recognized as a part of the summarization process (Headgcock & Ferris, 2009) and as an important perspective to be included in the overall evaluation. Conversely, Sawaki (2020) targeted students with relatively high English proficiency, but considered the development of the feedback-type rubric which suits participants who cannot avoid copying as in this study. It is effective to inform students of their paraphrasing score in addition to the overall evaluation including the perspective of paraphrasing. Including the perspectives of language use and paraphrasing is important to evaluate summary writing, so observing the relationship between language use and paraphrasing more in detail needs further consideration.

Finally, the raters in this study reported that the evaluation scale consisting of multiple phases should have been discussed more in detail in rater training. With respect to the report, sufficient quality of rater trainings is needed in the current study, several scholars (e.g., Heidari et al., 2022; Lumley, 2002; Shohamy et al., 1992; Weigle, 1994; 1998) also presented similar reports in their well-designed studies. Thus, as further study, quality of rater training and detailed observation of decision making of scoring for assessing summary writing, especially, for nonproficient learner of English is required.

Acknowledgement

This article is partly based on my doctoral dissertation submitted to Sophia University, Tokyo, in 2021. I would like to express my deepest thanks to Professor Yoshinori Watanabe, my research supervisor at the Graduate School of Languages and Linguistics, Sophia University, for his strong guidance and valuable advice throughout

this process. Also, I would like to acknowledge the committee members of my dissertation, Professor Yasuyo Sawaki at Graduate School of Education, Waseda University, and Professor Takanori Sato at the Center for Language Education and Research, Sophia University. They gave me valuable advice on statistics and helpful comments. Finally, this work was supported by JSPS KAKENHI Grant Number JP 22K13174.

References

Baba, K. (2009). Aspects of lexical proficiency in writing summaries in a foreign language. *Journal of Second Language Writing, 18*(3), 191-208. https://doi.org/10.1016/j.jslw.2009.05.003

Bachman, L. F. (2004). *Statistical analyses for language assessment*. New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9780511667350

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

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced Language Testing*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524803

Chiu, C. H., Wu, C. Y., & Cheng, H. W. (2013). Integrating reviewing strategies into shared electronic note-taking: Questioning, summarizing and note reading. *Computers & Education, 67*, 229-238. https://doi.org/10.1016/j.compedu.2013.04.015

Cohen, A. D. (1993). The role of instructions in testing summarizing ability. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 132-159).

Cohen, A. D. (1994). English for academic purposes in Brazil: the use of summary tasks. In C. Hill & K. Parry (Eds.), *From testing to assessment: English as an international language* (pp. 174-204). London, UK: Longman.

Cohen, L., Manion, L., & Morrison, K. (2011). *Research Methods in Education* (7th ed). Oxford, UK: Routledge Publishers (part of the Taylor & Francis group).

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York, NY: Wiley.

Grabowski, K., & Lin, R. (2019). Multivariate generalizability theory in language assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment: Vol. 1. Fundamental techniques* (pp. 30-53). New York, NY: Routledge. https://doi.org/10.4324/9781315187815-4

Graham, S., Hebert, M., Sandbank, M. P., & Harris, K. R. (2016). Assessing the writing achievement of young struggling writers: Application of generalizability theory. *Learning Disability Quarterly, 39*(2), 72-82. https://doi.org/10.1177/0731948714555019

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). *Range* [Computer software]. Retrieved from https://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx

Hedgcock, J. S., & Ferris, D. R. (2009). *Teaching readers of English: Students, texts, and contexts.* New York, NY: Routledge. https://doi.org/10.4324/9780203880265

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

Hirai, A. (2017). *Kyoiku, Shinri Kenkyu no Tameno Data Bunseki Nyumon*. Tokyo: Tokyo-tosho.

Hirai, A. (2018). *Kyoiku, Shinri Kenkyu no Tameno Data Bunseki: Kenkyu no haba wo hirogeru toukei syuhou*. Tokyo: Tokyo-tosho.

Holmes, J. L., & Ramos, R. G. (1993). Study summaries as an evaluation instrument: Questions of validity. *English for Specific Purposes, 12*(1), 83-94. https://doi.org/10.1016/0889-4906(93)90029-N

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

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

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

Kintsch, W., & Van Dijk, T. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363-394. https://doi.org/10.1037/0033-295X.85.5.363

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

Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing, 19*(2), 193-220. https://doi.org/10.1191/0265532202lt227oa

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). Los Angeles, CA: University of Southern California.

Li, J. (2014a). The role of reading and writing in summarization as an integrated task. *Language Testing in Asia, 4*. https://doi.org/10.1186/2229-0443-4-3

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

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

Manchón, R. M., Roca de Larios, J., & Murphy, L. (2007). Second and foreign language writing strategies: Focus on conceptualizations and impact of the first language. In A. D. Cohen & E. Macaro (Eds.), *Language Learner Strategies: 30 Years of Research and Practice* (pp. 229-250). New Yord, NY: Oxford University Press.

Norris, R. W. (2007). Dealing with plagiarism at a Japanese university: A foreign teacher's perspective. *The East Asian Learner, 3*(1), 1-20.

Riley, G. L., & Lee, J. F. (1996). A comparison of recall and summary protocols as measures of second language reading comprehension. *Language Testing, 13*(2), 173-189. https://doi.org/10.1177/026553229601300203

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

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

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. https://doi.org/10.1007/978-981-15-8952-2_10

Sawaki, Y., & Xi, X. (2019). Univariate generalizability theory in language assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment: Vol. 1. Fundamental techniques* (pp. 30-53). New York, NY: Routledge. https://doi.org/10.4324/9781315187815-3

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: SAGE Publications.

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

Taylor, K. K. (1986). Summary writing by young children. *Reading Research Quarterly, 21*(2), 193-208. https://doi.org/10.2307/747845

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York, NY: Academic Press.

Watanabe, Y. (2001). *Read-to-Write tasks for the assessment of second language academic writing skills: Investigating text features and rater reactions.* (Unpublished doctoral dissertation). University of Hawaii, HI.

Webb, N. M., Shavelson, R. J., & Maddahian, E. (1983). Multivariate generalizability theory. In L. J. Fyans (Ed.), *New directions in testing and measurement: Generalizability theor*y (pp. 67-82). San Francisco, CA: Jossey-Bass.

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

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

Westby, C., Culatta, B., Lawrence, B., & Hall-Kenyon, K. (2010). Summarizing expository texts. *Topics in Language Disorders, 30*(4), 275-287. https://doi.org/10.1097/TLD.0b013e3181ff5a88

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.* https://doi.org/10.1186/s40468-019-0087-6

Appendix A

Scoring Rubric

(1) Main Idea Coverage

5. EXCELLENT: A response has complete coverage of main ideas.

4. VERY GOOD: A response has coverage of most main ideas.

3. GOOD: A response has moderate coverage of main ideas.

2. MODERATE: A response has some coverage of main ideas.

1. POOR: A response has coverage of very few ideas.

0. NO: A response has no coverage of main ideas.

(2) Integration

5. EXCELLENT: A response rearranges the order of the statements logically, displays excellent examples of integration and connectives, and demonstrates global interpretation of the source text.

4. VERY GOOD: A response rearranges the order of the statements logically, displays good examples of. integration and connectives, and demonstrates global interpretation of the source text.

3. GOOD: A response rearranges the order of the statements logically, displays moderate examples of. integration and connectives, and demonstrates global interpretation of the source text.

2. MODERATE: A response basically follows the order of source text with few cases of re-ordering and. integration, and is not global in the interpretation of the source text.

1. POOR: A response follows the original order of the statements in the source text, shows rare instance of. proper integration and connectives, and is not global in their interpretation of the source text.

0. NO: A response has no instances of integration or connectives at all.

(3) Language Use

5. EXCELLENT: A response displays consistent facility in the use of language, demonstrating syntactic. variety, appropriate word choice.

4. VERY GOOD: A response displays facility in the use of language, demonstrating syntactic variety and. range of vocabulary, though it will probably have occasional noticeable minor errors in structure, or word form that do not interfere with meaning.

3. GOOD: A response demonstrates inconsistent facility in sentence formation and word choice that may result in lack of clarity and occasionally obscure meaning.

2. MODERATE: A response has a noticeably inappropriate choice of words or word forms, an accumulation. of errors in sentence structure and/or usage.

1. POOR: A response has serious and frequent errors in sentence structure or usage, the text shows a lack of. control of vocabulary and/or grammar.

0. NO: A response is totally incomprehensible due to language errors, or because the response is left blank.

(4) Source Use

5. EXCELLENT: A response is predominantly in the summarizers' own words and sentence structures, in. addition to the accurate use of the information from the source text.

4. VERY GOOD: A response is mostly in the summarizers' own words and sentence structures, in addition to. the accurate use of the information from the source text.

3. GOOD: A response is basically in the summarizers' own words and sentence structures, in addition to. appropriate use of information from the source text.

2. MODERATE: A response has some use of the summarizers' own words and sentence structures, in addition to the adequate use of the information from the source text.

1. POOR: A response is predominately verbatim copying the source text.

0. NO: A response demonstrates completely verbatim copying from the source text.

Appendix B

Source Text to be summarized

Campus Life in America

As far as campus life is concerned, going to college in America is very different from Japan. For a start, you'll only hear spoken English, and learning activities seem more informal. Many students even call their teachers by their first names. As a result, their relationship seems much closer than between teachers and students in Japan. Moreover, in class, they sit around the teacher and classes are generally smaller than in Japan.

On the other hand, studying in Japan is generally more formal. Students address their teachers as "sensei," and never use first names. By and large, classes in Japan are bigger than in America with more students attending them. On campus, you'll find both fun-loving students and hard-working, too.

In conclusion, because of differing lifestyles in America and Japan, it is no surprise that life on campus in the two countries is pretty different, too. If you are Japanese and can study in America, or vice versa, you will have many opportunities to learn about varying social attitudes and customs.

(172 words)