English Language Teaching; Vol. 8, No. 7; 2015
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

Evaluation by Native and Non-Native English Teacher-Raters of Japanese Students' Summaries

Yuko Hijikata-Someya[1], Masumi Ono[2] & Hiroyuki Yamanishi[3]

[1] Department of Teaching and Learning, The Ohio State University, Columbus, Ohio, USA

[2] Faculty of Law, Keio University, Kanagawa, Japan

[3] Faculty of Foreign Language Studies, Kansai University, Osaka, Japan

Correspondence: Yuko Hijikata-Someya, Department of Teaching and Learning, College of Education and Human Ecology, The Ohio State University, 196A Arps Hall, 1945 N. High Street, Columbus, OH, 43210, USA. Tel: 1-614-292-5745. E-mail: someya.2@osu.edu

Received: April 22, 2015 Accepted: May 26, 2015 Online Published: June 29, 2015

doi:10.5539/elt.v8n7p1 URL: http://dx.doi.org/10.5539/elt.v8n7p1

Abstract

Although the importance of summary writing is well documented in prior studies, few have investigated the evaluation of written summaries. Due to the complex nature of L2 summary writing, which requires one to read the original material and summarize its content in the L2, raters often emphasize different features when judging the quality of L2 summaries. Therefore, this study examines the ratings of English-language summaries written by Japanese university students in order to identify differences in EFL instructors' evaluations. Fifty-one Japanese EFL university students read a passage and then wrote an English summary without receiving any instructions concerning summary composition. The raters included three native English speakers (NESs) and three non-native English speakers (NNESs), who individually evaluated each summary using the Educational Testing Service's holistic rubric. Analysis of inter-rater reliability revealed a lower Cronbach's alpha coefficient for NNES raters ($\alpha = .39$) when compared to NES raters ($\alpha = .77$). Comments were collected from raters regarding the difficulty of evaluating summaries, and the causes of such difficulties were examined. Comments from NNES raters more concerned vocabulary use and paraphrasing, whereas the NES raters concentrated on content and language. This study also explores ways to potentially improve the holistic rubric by examining feedback from raters regarding their rating experiences.

Keywords: evaluation, holistic rubric, native English speaker, non-native English speaker, paraphrasing, reliability, summary writing

1. Introduction

Summary writing is widely recognized as an important teaching method, particularly for university students in foreign language classes, and is also an effective tool for measuring L2 proficiency. Moreover, university students are often required to perform writing tasks that involve summarization when taking academic courses conducted in English. Nevertheless, English as a foreign language (EFL) students often struggle to use source texts properly, despite this being an essential academic skill (Hirvela & Du, 2013; Shi, 2012).

To foster summary writing skills among EFL students, and also enhance the quality of L2 summary writing instruction, a research project was launched to focus on the evaluation of written summaries. This study is one from a series of studies comprising a project for the development of efficient and useful rubrics (or rating scales) for L2 summary writing in EFL academic contexts. The flow of this larger project is illustrated in Figure 1.

Study I (Hijikata, Yamanishi, & Ono, 2011) examined the reliability and validity of a holistic rubric developed by the Educational Testing Service (ETS). In the study, three Japanese raters used the ETS rubric to evaluate summaries written by 51 Japanese EFL university students. The Cronbach's alpha reliability coefficient was .51, which did not indicate sufficient reliability among the three raters. The results of Study I revealed that the rubric was difficult to use, and highlighted the need for an analytic rubric specifically targeting non-native English speakers (NNESs) who evaluate and teach L2 summary writing in higher education.

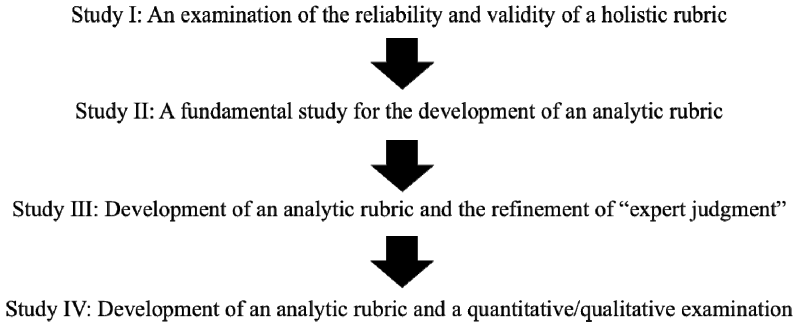

Figure 1. Flow of this research project

To determine whether an analytic or holistic rubric is better suited to evaluating and teaching L2 summary writing in an academic context, it is first necessary to identify components that NNES raters find particularly difficult to grade, and also examine whether their evaluations differ from those of native English speakers (NESs). Therefore, the purpose of the present study (Study II) is to investigate the feasibility of developing an analytic rubric that is capable of overcoming difficulties associated with the use of a holistic rubric. After identifying the causes of difficulties faced by NNESs when grading summaries, this research project intends to propose a new rubric and refine it in Study III and Study IV. The project's overall goal is to create an analytic rubric based on the results of Study II, conduct expert judgment in Study III, and quantitatively and qualitatively examine the newly developed rubric in Study IV.

1.1 Literature Review

Summarizing skills comprise the ability to extract important information, paraphrase in one's own words, and condense important ideas into a short text. Among these skills, paraphrasing is commonly quite difficult for EFL learners, although it has received little attention as a research topic. L2 summary writing has been investigated from various perspectives, such as by examining the use of strategies (Johns, 1985), the role of constructs (Asención-Delaney, 2008; Baba, 2009), composition processes (Plakans, 2008), and the relationship between strategy use and the end product in summary writing (Yang & Plakans, 2012). Some empirical studies have also been published that address paraphrasing among NNESs (e.g., Keck, 2006, 2014).

While rubrics are widely used to evaluate L2 writers' summaries, limited research has been conducted concerning the evaluation of written summaries. Here we review various holistic and analytic L2 writing rubrics in order to address key issues related to the evaluation of summaries composed by L2 writers.

The Educational Testing Service is responsible for developing one of the most popular holistic L2 writing rubrics (Hamp-Lyons, 1989), and it was originally intended for use in conjunction with the organization's *Test of Written English* (TWE). Similarly, Jacobs, Zinkgraf, Wormuth, Hartfiel, and Hughey's (1981) *ESL Composition Profile* is representative of a typical analytic L2/ESL writing rubric. Nevertheless, the primary difference between these two rubrics is that they are one-dimensional and multi-dimensional respectively.

A major strength of holistic rubrics is that their use requires minimal effort from raters (Bacha, 2001; Hamp-Lyons, 1995; Hyland, 2003; Weigle, 2002), which can be particularly beneficial in a classroom setting, where the amount of time available for evaluation is often limited. However, a major weakness of holistic rubrics lies in the fact that learners may receive less diagnostic feedback from instructors. For instance, a student who earned 3 out of 5 possible points on a writing test in April, and 5 points in September may not be able to discern which aspects he or she improved upon based on numbers alone.

Analytic rubrics are ideal for providing diagnostic feedback since they reflect multiple dimensions; that is, they divide the constructs of L2 writing skills into certain operational, defined sub-categories. For example, Jacobs et al.'s (1981) *ESL Composition Profile* comprises five dimensions: content, organization, vocabulary, language use, and mechanics. The multi-dimensional nature of analytic rubrics is their key strength, as it allows one to easily obtain diagnostic information concerning improvement in L2 writers' compositions. In the aforementioned scenario, an analytic rubric could provide an instructor with sufficient evidence to suggest that the student's writing skills improved in content and mechanics between April and September, but not in vocabulary. In this regard, analytic rubrics for L2 writing are ideal for use in classroom settings, and indeed many have been developed with specific L2 classroom settings in mind. An analytic rubric designed by Nishijima, Hayashi, Masaki, Kinshi, and Kuru (2007), for instance, addressed issues unique to higher education in Japan.

In-depth studies concerning rubrics (particularly analytic rubrics) for L2 summary writing are scant, likely because scholars consider the evaluation of summaries to be a difficult undertaking, which is primarily due to the following three factors. First, an incomplete or problematic summary may make it difficult for raters to determine whether a writer failed to understand a passage, or if the individual experienced difficulty writing about a text despite possessing a proper understanding of it. Indeed, Taylor (1986) and Winograd (1984) demonstrated that many L2 writers struggle to express ideas that they clearly understand. Second, raters may disagree concerning details that they deem essential for inclusion in a summary (Alderson, 2000). Third, researchers who develop analytic rubrics for L2 summary writing may find it difficult to address the complicated constructs of L2 summary writing using operationally defined and measurable dimensions. As such, prior investigations into summary writing often used holistic-scoring rubrics (e.g., Baba, 2009; Cumming et al., 2005; Trites & McGroarty, 2005).

Although its suitability in a classroom setting is unclear, the ETS rubric is nonetheless among the most frequently used for the purpose of evaluating L2 summary writing (e.g., Baba, 2009). While a holistic rubric should be beneficial in terms of practicality, it could potentially pose difficulties in the evaluation of summaries composed by L2 writers.

A writer's score in the ETS rubric is determined according to whether he or she fulfilled the requirements of a particular classification (see Appendix A). For example, to obtain a score of 5, a writer should satisfy each of that particular classification's four descriptors; such individuals are assumed capable of comprehending a passage's underlying message. Thus, this score should indicate that one can skillfully substitute certain phrases with equivalents written in his or her own words. In contrast, a person whose writing satisfies only three of the classification's four descriptors will receive a score of 2. Yet, grading difficulties can arise when descriptors such as "appropriate use of [one's] own language and language from the source text" along with "sentence formation and word forms [are] accurate and appropriate" fail to appear in the same summary. Therefore, if a writer merely copies the source material, his or her summary will naturally contain few grammatical errors; conversely, if a novice writer paraphrases the source material, there will likely be a significant number of grammatical errors.

Since summary writing requires a number of different skills, and because rubrics for evaluating summaries are in short supply, a rater's first language, educational background, and teaching experience could contribute to determining the dimensions that are given precedence. Prior studies have investigated the potential influence of rater background on the evaluation of writing performance, both by using a holistic rubric, and from the standpoint of speaking and writing assessment. Studies comparing the scores of speaking or those of writing tests rated by NESs and NNESs revealed notable qualitative scoring differences between both rater groups although neither group significantly differed in their actual scores.

Earlier research concerning the evaluation of speaking and writing using holistic rubrics has shown that judgment is influenced by rater background. Kim (2009) and Winke, Gass, and Myford (2012) demonstrated the effect of rater background on determining which dimensions are prioritized in the evaluation of speaking. Kim's study compared the speaking performance ratings given to NNES students by two separate groups of 12 NES and NNES teachers. The results revealed that NES raters were more critical than their NNES counterparts in their evaluation of pronunciation, grammar use, and the accuracy of transferred information.

Zhang and Elder (2011) examined differences in rater judgment among 20 and 19 NNES and NES raters respectively, who were tasked to evaluate the oral proficiency of 10 English speakers' speech samples. By using the multifaceted Rasch model of measurement, the researchers determined that there was no significant difference between the two groups' scores. However, qualitative analysis revealed differences between NESs and NNESs in their justifications for allotting certain scores, with the latter group placing greater emphasis on linguistic resources.

Shi (2001) conducted a study comprising two groups of 23 NES and NNES raters, who examined 10 expository essays. The bases for their evaluations were classified into five categories related to general issues, content, organization, language, and length. The results of a multivariate analysis of variance did not show any significant differences in scores, although a chi-square test revealed that the reasons for allotting each score differed between groups. Specifically, NESs tended to leave positive comments concerning content and language, whereas NNESs often left negative comments regarding organization and length.

1.2 Research Questions

A limited number of studies have investigated the evaluation of summaries, although its importance has been acknowledged. Consequently, few rubrics have been developed for evaluating L2 summary tasks. Furthermore, the usefulness of holistic rubrics in a classroom context remains unclear. Likewise, it has not been established

whether such holistic rubrics are of equal use to NES and NNES raters for the evaluation of L2 summaries.

Thus, to ascertain whether holistic rubrics are effective in assessing L2 summaries, and if the development of an analytic rubric is necessary, it is imperative to discern whether it is possible for NES and NNES raters to score summaries similarly. Moreover, determining whether raters experience any difficulties while using a common holistic rubric is also key. With these aforementioned issues in mind, the present study will compare English summary writing scores allotted to Japanese EFL students by NES and NNES teacher-raters. In particular, we examine which dimension(s) the two groups tend to focus on, and also the dimension(s) wherein the two groups often differ. Therefore, this study will address the following research questions:

RQ1: Does the holistic ETS rubric have similar inter-rater reliability, and do NES and NNES raters experience difficulty when scoring L2 summaries?

RQ2: Which dimensions do NES and NNES raters emphasize when scoring L2 summaries: content, organization, vocabulary, language, mechanics, paraphrasing, or length?

2. Method

2.1 Participants

Fifty-one first-year EFL students from two Japanese universities participated in this study, whose English proficiencies ranged from intermediate to lower-intermediate. Students from universities A and B possessed average TOEIC-IP and TOEFL-ITP scores of 532.1 (*SD* = 117.4) and 420.7 (*SD* = 31.6), respectively. Participants from University A were management majors, whereas students from University B specialized in various fields related to the English-language. Members of both groups had studied English for more than six years.

The NES and NNES rater groups included three NES teachers of English, and three NNES Japanese teachers of English respectively. Table 1 shows a summary of each rater's educational background and teaching experience.

Table 1. Educational background and teaching experience of each rater

Variable	NES 1	NES 2	NES 3	NNES 1	NNES 2	NNES 3
Position	Ph.D. student; part-time lecturer	Ph.D. student	Master's student	Associate professor	Associate professor	Lecturer
Years of university teaching experience	4	1	0	8	8	1
Years of teaching experience elsewhere	15	14	5	1.5	0	3
Past experience scoring summaries	Yes	Yes	No	No	No	Yes
Past experience using an ETS rubric	No	No	No	Yes	No	No

The NES raters were British graduate students majoring in applied linguistics, whose experience teaching English ranged between 5 and 19 years; the NNES raters possessed between 4 and 9 years of English teaching experience. Of the six instructors, two NESs and one NNES had prior experience scoring summaries. None of the raters previously used the ETS rubric for the evaluation of summaries, although one NNES rater had used it in conjunction with the TWE.

2.2 Materials

Students were given a 199-word passage from Oshima and Hogue (2007) to read. The passage had a clear comparative text structure, which compared the left and right sides of the human brain (see Appendix B); its readability, as measured by the Flesch-Kincaid Grade Level scale, was 9.3. Considering the English proficiencies of the study's participants (see Section 2.1), and the processing-difficulty level of the aforementioned passage, the researchers did not anticipate that readers would experience any difficulty comprehending the material.

2.3 Procedure

Data collection spanned two weeks. During the first week, participants composed a 50-60-word summary in English without receiving any explicit instruction concerning how a summary should be written. A time limit was not specified, although the summaries were not to exceed 25-30% of the original text's length (Sherrard,

1989; Taylor, 1984).

The following week, participants were shown model summaries, accompanied by a description of how summaries should generally be written. Two NESs pursuing a Ph.D. in linguistics were tasked with composing the model summaries (see Appendix C), and informed that their summaries should be roughly 25-30% of the original's in length. While familiarizing students with the conventions of summary writing, the instructors focused on three concepts: deletion, generalization, and construction (Kintsch & van Dijk, 1978). Students were provided examples of each of these concepts, which were based on prior research conducted by Muramoto (1998).

After the second week, participants were asked to summarize the first passage a second time. During the process, participants were permitted to refer back to their initial summary. These summaries were then submitted as Microsoft Word documents, which included both their original and most recent summaries.

2.4 Scoring and Data Analyses

Once the written data were collected, raters used the ETS rubric to score the 102 summaries; to prevent order effects, the summaries were randomized before being distributed to each rater. The raters then a) scored the summaries, b) noted any difficulties encountered during the evaluation process, and c) provided information pertaining to their educational and teaching backgrounds—in addition to general comments concerning the task itself. The first data set (i.e., item a) was composed of the scores allotted by each rater, which were based upon their interpretation of the rubric; as stated in Section 1.1, the ETS rubric allows for a maximum obtainable score of 5. These scores were primarily used to measure inter-rater reliability.

The second data set (i.e. item b) contains information concerning the difficulty experienced by raters while scoring. To categorize raters' comments, five components were borrowed from Jacobs et al.'s (1981) analytic rubric. These components were selected due to the rubric's extensive use in the evaluation of ESL compositions for research and teaching purposes in a wide range of contexts. Moreover, the rubric partially addresses areas that are particularly relevant to the summary writing process.

Despite the suitability of Jacobs et al.'s rubric in relation to this study's research goals, it was nonetheless necessary to add two components in order to increase its applicability to summary writing—namely components related to paraphrasing and length. Thus, the finalized analytic framework included the following seven components, which were also identified in the ETS holistic rubric:

1) Content: How well the writer understood the passage's content

2) Organization: How well the summary was organized as a paragraph

3) Vocabulary: Word choice and the appropriate use expressions

4) Language: Sentential and discourse grammar

5) Mechanics: Writing rules (e.g., punctuation)

6) Paraphrasing: To what extent the writer explained a passage's content using his or her own expressions

7) Length: Whether the writer's summary stayed within the specified word limit

The third data set (i.e., item c) included information related to the raters' backgrounds, in addition to general feedback concerning their use of the ETS rubric. These details were obtained through a questionnaire comprising four open-ended and six multiple-choice questions (see Appendix D).

3. Results

3.1 Evaluation Results

Table 2 shows the results for scoring and difficulty of evaluation. An alpha level was set at 5% ($p < .05$).

Table 2. Scoring and difficulty of evaluation ($N = 102$)

Rater	Scores	Difficulty (Frequency)		
	M (*SD*)	Easy	Moderate	Difficult
Total for NES raters	3.31 (0.74)	123	172	11
NES rater 1	4.07 (0.99)	87	15	0
NES rater 2	3.09 (0.94)	0	102	0
NES rater 3	2.76 (0.76)	36	55	11
Total for NNES raters	3.28 (0.65)	70	155	80
NNES rater 1	3.45 (0.84)	9	63	29
NNES rater 2	2.71 (1.02)	20	42	40
NNES rater 3	3.70 (1.01)	41	50	11

An independent *t*-test analysis of data from the NES and NNES raters did not reveal a statistically significant difference between the two groups' mean scores, t (202) = 0.24, p = .815, r = .02. However, analysis of inter-rater reliability revealed a high (α = .77) and low (α = .39) Cronbach's alpha coefficient for the NES and NNES rater groups respectively. Regarding the difficulty of evaluation, a statistically significant difference was detected, χ^2 (2) = 67.76, $p < .001$, Cramer's V = .33. Post-hoc residual analysis indicated that NES raters tended to deem the task of evaluation "easy"; comparatively, NNES raters most often considered it to be "difficult." The adjusted standardized residuals for the NES and NNES groups were statistically significant (4.59 and 7.86, respectively). Thus, these results confirm that the NNES group experienced greater difficulty using the ETS holistic rubric.

3.2 Qualitative Analyses

3.2.1 Comments from Raters Concerning the Difficulty of Scoring Summaries

In examining why certain summaries proved to be more difficult for raters to score, 16 were identified whose grading difficulty on average exceeded 2.0 (moderate). Therefore, we specifically focused on these summaries by qualitatively analyzing each rater's comments concerning them. As described in Section 2.3, the study's qualitative framework is an adaptation of Jacobs et al.'s (1981), wherein two components have been added.

Rater comments concerning the difficulties that they encountered while grading the summaries were both positive and negative in nature. Hence, two researchers experienced in the grading of summaries were asked to independently code rater comments according to the seven components described in Section 2.3, and also differentiating between positive and negative elements within them. An inter-coder reliability check was conducted by dividing the number of components that both researchers agreed upon by the number of components identified in the 16 summaries; consequently, a relatively high reliability coefficient (82.4%) was obtained. A discussion between both researchers subsequently occurred, wherein discrepancies in their findings were resolved.

Table 3 shows the distribution of positive and negative components in raters' comments.

Table 3. Components of raters' comments

Component	NES raters ($n = 3$)		NNES raters ($n = 3$)	
	Positive	Negative	Positive	Negative
Content	4	7	10	8
Organization	0	1	1	3
Vocabulary	0	1	0	8
Language	0	7	2	11
Mechanics	0	0	0	0
Paraphrasing	3	0	1	5
Length	0	1	0	1

As shown in Table 3, both groups frequently attributed difficulties in evaluating summaries to content and language-related issues. However, comments from NNES raters more often concerned vocabulary use and paraphrasing when compared to NES raters.

The following two examples highlight specific causes of difficulties encountered by raters with regard to their use of the holistic rubric. In the first example, the NNES raters expressed somewhat similar opinions concerning Summary A-19 and the writer's failure to paraphrase. NNES Rater 1 noted that, "many parts are just copied," while NNES Rater 2 deemed it "extremely difficult [to score] due to [the writer's] very limited use of [his/her] own language."

The left and right sides of brain process information in different ways. Left brains think in words (logical, rational, linear, and verbal) and analyze carefully, but right brains think in pictures (visual, intuitive, and sensual) and create. One side is stronger, but both brains are well-balanced and work normally together. Summary A-19 (Score: 3.33; Difficulty: 2.17)

Despite the absence of paraphrasing in Summary A-19, NNES Rater 1 nonetheless believed that it "included all necessary information." This indicates that, at least from a content perspective, that NNES Rater 1 had a somewhat favorable impression of the summary, since it managed to convey the main idea expressed in the original passage.

Only two raters (one an NES and the other an NNES) noted any difficulties in grading Summary A-3. However, these two raters focused on different aspects in the summary, and therefore associated different causes with their difficulties.

The left and right sides of your brain process information in different ways. The left side is more logical. On the other hands, the right side uses the five senses more. So a left-brained and a right-brained person think in different ways. Though, usually people's both sides of their brain work together. Summary A-3 (Score: 3.83; Difficulty: 2.17)

Whereas the NES rater's comments focused on paraphrasing and content, the NNES rater called attention to a grammatical error (i.e., "On the other *hands*"). NES Rater 1 recognized that the student used his/her own words to write the summary, but also noted that he/she nevertheless failed to adequately explain the differences between the brain's left and right sides; moreover, the writer did not mention that humans generally make use of a combination of both sides. Therefore, NES Rater 1 acknowledged that the writer successfully paraphrased the text, but also highlighted the writer's incomplete understanding of the passage's main idea.

3.2.2 Rater Reflections

Reflections from each of the six raters were obtained through open-ended questions, and subsequently examined in order to gather opinions concerning the general difficulty of using the rubric, as well as to identify possible ways of improving upon it. Feedback from the raters tended to address the rubric's formal aspects and the content of its descriptors. Overall, the raters' difficulties seemed to be attributable to the rubric's holistic nature. For instance, NES Rater 3 related that he/she often found the summaries to be "on the border between marks," wherein they included the characteristics of two different categories. Furthermore, NNES Rater 1 ascribed the rubric's difficulty of use to the absence of clear and concise descriptors, which made the task of evaluating each summary tedious. NNES Rater 3 also found the descriptors to be ambiguous, particularly in summaries wherein language errors were present despite being otherwise adequate from a content perspective.

Regarding the rubric's content, the raters' comments primarily focused on how to properly evaluate each student's ability to paraphrase, since L2 writers with low English proficiencies often copy sentences and phrases directly from the original text (Keck, 2006). For example, NNES Rater 2 indicated that he/she struggled to distinguish between paraphrased information and information that had been copied directly from the source text—which in many cases constituted the bulk of students' summaries. Hence, careful consideration should be given to these aforementioned observations when developing a rubric for summary writing, especially in light of the need for students to avoid plagiarism in academic writing.

Both rater groups proposed ways in which the use of rubrics could be improved upon. First, the raters indicated the need for an analytic rubric for educational purposes, and asserted that the development of such a rubric could reap greater benefit than the refinement of a preexisting holistic rubric. Second, NNES Rater 1 believed that more comprehensive and concise descriptors were needed. He/she further suggested that a how-to styled rubric could help in achieving this, since it would consequently make the rubric more self-explanatory and transparent, both for teachers and students. Moreover, both rater groups desired an improved approach to evaluating the paraphrasing component of summary writing, and believed that this would assist in determining the appropriate balance between textual borrowing and paraphrasing in students' summaries.

4. Discussion

Is a holistic or analytic rubric best suited for the scoring of L2 summaries? Moreover, is the development of an

EFL-specific analytic rubric truly needed? To answer these questions, it was first necessary to determine whether NES and Japanese NNES raters would score summaries written by university-level Japanese EFL students differently. Using the ETS holistic rubric, raters scored students' summaries, and later documented any difficulties that were encountered while doing so. These scores and responses were subsequently examined quantitatively in terms of inter-rater reliability, and qualitatively in order to identify reasons why the raters may have experienced difficulties in scoring summaries. The following paragraphs discuss the study's findings in relation to the research questions presented in Section 1.2.

As for RQ1, neither group differed significantly, a finding consistent with prior studies that examined teachers' evaluations of oral and written EFL proficiency (Shi, 2001; Zhang & Elder, 2011). Therefore, it may seem that raters' language backgrounds had no effect on their judgment, and that the use of a holistic rubric did not lead to significant scoring differences. However, contrary to the results attained through a comparison of both groups' means, low inter-rater reliability was obtained for the NNES group's data when compared to the NES group's. Furthermore, the chi-square test results revealed that the NNES group experienced greater difficulty scoring summaries. These findings indicate that using a holistic rubric may result in scoring difficulties and inconsistencies, particularly for NNES raters. Hence, despite a lack of significant differences between the mean scores of summaries graded by either group, the present study's results are in agreement with earlier research indicating qualitative differences between NES and NNES raters (e.g., Kim, 2009; Shi, 2001; Zhang & Elder, 2011).

With respect to RQ2, qualitative analysis revealed that the NNES raters gave more comments regarding the use of vocabulary and paraphrasing than did the NES raters. In contrast, the NES raters emphasized the aspects of content and language use, which is consistent with the findings of Shi (2001). The emphasis placed on paraphrasing by the NNES group may be indicative of a strong belief that paraphrasing in L2 summary writing is important. Alternatively, it could be because the group was less confident in judging general language use, since only one of its members had prior experience grading L2 summaries; consequently, he/she may have paid greater attention to the ratio of paraphrased to copied text.

Having examined the incidences in which raters encountered difficulties using the ETS rubric, and after reviewing the various perceptions of raters regarding these difficulties, it is worthwhile to discuss additional problems that could arise when using the rubric to evaluate EFL students' summaries. First, the holistic rubric's scores cannot always be used to distinguish between a summary that contains substantial use of paraphrasing (albeit with numerous grammatical errors), and a summary that was primarily copied from the original source. Although substantial revision should be recommended as an effective paraphrasing strategy, students may hesitate to make such revisions if language accuracy is factored into the grading process. Therefore, it would be helpful to show students how different types of paraphrasing might be reflected in their final scores.

Second, due to its one-dimensional nature, the holistic rubric cannot provide students or teachers with a sufficient amount of constructive or informative feedback concerning changes in student performance (Bacha, 2001; Carr, 2000; Cumming, 1997; Hamp-Lyons, 1995). For example, if a student receives a score of 3 for his or her first and second summaries, he or she may find it difficult to discern whether the writings were nearly identical, or somewhat different but similar in overall quality. Thus, this characteristic renders the rubric ineffective as a tool for the teaching and learning of summary writing.

Third, the root cause of disparities among raters concerning a summary's evaluation cannot always be easily identified. Summary writing is a complex and dynamic process involving, "the comprehension, evaluation, condensation, and frequent transformation of ideas that have been presented" (Hidi & Anderson, 1986, pp. 473-474). Therefore, when scoring summary writing one must consider several factors, such as the written English proficiency of the author, whether the passage was understood, and if it was paraphrased (i.e., not copied from the original text).

The limitations of this study are as follows. First, the two groups of students did not take the same proficiency test, since their respective universities incorporated either the TOEIC-IP or TOEFL-ITP into their programs. Although this inconsistency does not change the study's main findings, future research should ideally include participants who have taken an identical test.

Second, this study did not factor prior summary writing experience into its analysis. Given the crucial role of paraphrasing in summary writing, prior experience composing summaries could have affected participants' final scores; consequently, this feature will be accounted for in future studies.

The third limitation of this study concerns the lack of diversity among its raters, specifically in terms of their teaching experience and educational backgrounds. Admittedly, because the number of raters in this study was

relatively small, it was not feasible to control for whether they had ever used the ETS rubric or evaluated L2 summary writing—although this issue should be addressed in future research.

Despite these limitations, the study's findings yielded important educational implications. A holistic rubric cannot convey the true significance of the final scores allotted by raters, and therefore does not seem to be effective in a classroom context. Summary writing is particularly complicated in this regard, since students are expected to grasp the gist of a passage and then express it in their own words. Accordingly, an analytic rubric is more desirable from a pedagogical perspective for the evaluation of summary writing, as it can be used to distinguish between the various dimensions involved in the summarization process (e.g., properly identifying the gist of the source material, writing in an accurate and organized way, and paraphrasing). To increase ease of use, the inclusion of a how-to guide for the rubric could prove effective. Furthermore, the rubric should incorporate a dimension to address paraphrasing specifically.

Finally, Japanese NNES raters encountered greater difficulty in scoring, an observation supported by a low reliability coefficient, frequency analysis of the three difficulty levels, and the group's reflections on their rating experience. An analytic rubric would likely be useful in overcoming this problem.

This study was the first, to the researchers' knowledge, to compare scores allotted by NES and NNES raters in L2 summary writing. In an ensuing study, the authors will propose a suitable analytic rubric for the evaluation of summaries written by L2 writers in a classroom context.

Acknowledgements

The project was funded by grants from the Japan Society for the Promotion of Science, KAKENHI (# 23520725 and # 26580121). An earlier version of this paper was presented at the 38th Annual Meeting of the Japan Society of English Language Education.

References

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

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

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

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

Carr, N. T. (2000). A comparison of the effects of analytic and holistic rating scale types in the context of composition tests. *Issues in Applied Linguistics*, *11*, 207-241.

Cumming, A. (1997). The testing of writing in a second language. In C. Clapham, & D. Corson (Eds.), *Encyclopedia of language and education* (Vol. 7, pp. 51-64). Dordrecht, Netherlands: Kluwer Academic Publishers.

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

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

Hamp-Lyons, L. (1989). *Newbury house TOEFL preparation kit: Preparing for the test of written English*. New York City: Newbury House.

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

Hidi, S., & Anderson, V. (1986). Producing written summaries: Task demands, cognitive operations, and implications for instruction. *Review of Educational Research*, *56*, 473-493. http://dx.doi.org/10.3102/00346543056004473

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

Hirvela, A., & Du, Q. (2013). Why am I paraphrasing? Undergraduate ESL writers' engagement with

source-based academic writing and reading. *Journal of English for Academic Purposes*, *12*, 87-98. http://dx.doi.org/10.1016/j.jeap.2012.11.005

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

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

Johns, A. M. (1985). Summary protocols of "under-prepared" and "adept" university students: Replications and distortions of the original. *Language Learning*, *35*, 497-517. http://dx.doi.org/10.1111/j.1467-1770.1985.tb00358.x

Keck, C. (2006). The use of paraphrase in summary writing: A comparison of L1 and L2 writers. *Journal of Second Language Writing*, *15*, 261-278. http://dx.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. http://dx.doi.org/10.1016/j.jslw.2014.05.005

Kim, Y. H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing*, *26*, 187-217. http://dx.doi.org/10.1177/0265532208101010

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

Muramoto, T. (1998). *Bunsyou rikai ni tsuite no ninchi shinrigaku teki kenkyu: Kioku to youyaku ni kansuru jikken to rikai katei no moderuka* [*Cognitive psychological research on reading comprehension: Modeling of comprehension process and experiments on memorization and summarization*]. Tokyo: Kazama Shobo.

Nishijima, H., Hayashi, K., Masaki, M., Kinshi, K., & Kuru, Y. (2007). Developing a writing rubric for classroom use in Japanese higher education. *JACET Journal*, *45*, 109-115.

Oshima, A., & Hogue, A. (2007). *Introduction to academic writing*. New York City: Pearson Longman.

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

Sherrard, C. (1989). Teaching students to summarize: Applying textlinguistics. *System*, *17*, 1-11. http://dx.doi.org/10.1016/0346-251X(89)90055-9

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing*, *18*, 303-325. http://dx.doi.org/10.1177/026553220101800303

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

Taylor, K. K. (1984). Teaching summarization skills. *Journal of Reading*, *27*, 389-393.

Taylor, K. K. (1986). Summary writing by young children. *Reading Research Quarterly*, *21*, 193-208.

Trites, L., & McGroarty, M. (2005). Reading to learn and reading to integrate: New tasks for reading comprehension tests? *Language Testing*, *22*, 174-210. http://dx.doi.org/10.1191/0265532205lt299oa

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

Winke, P., Gass, S., & Myford, C. (2012). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*, 231-252. http://dx.doi.org/10.1177/0265532212456968

Winograd, P. N. (1984). Strategic difficulties in summarizing texts. *Reading Research Quarterly*, *19*, 404-425.

Yang, H. C., & Plakans, L. (2012). Second language writers' strategy use and performance on an integrated reading-listening-writing task. *TESOL Quarterly*, *46*, 80-103. http://dx.doi.org/10.1002/tesq.6

Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing*, *28*, 31-50. http://dx.doi.org/10.1177/0265532209360671

Appendix A

Rubric (Reading/Writing Task Scoring Guidelines) Developed by ETS (2002, p. 47)

5) A response at this level has all of the following qualities:

- principal ideas presented accurately with ample and accurately connected key supporting points/elaboration as required to fulfill the task effectively
- organization effective in response to the task
- sentence formation and word forms accurate and appropriate; response may have occasional minor grammatical or lexical errors
- appropriate use of own language and language from source text

4) A response at this level has all of the following qualities:

- principal ideas presented accurately as required by the task, though one or two key supporting points/details/elaboration may be omitted, misrepresented, or somewhat unclear, inexplicit, or inexplicitly connected
- organization generally effective in response to the task
- sentence formation and word choice generally accurate and appropriate; response may have noticeable minor errors and some imprecision and/or unidiomatic language use and/or imprecise connections among ideas; however, these do not obscure meaning
- generally appropriate use of own language and language from the source text

3) A response at this level is marked by inconsistency:

- principal ideas inconsistently presented: some are discussed accurately with key supporting points/elaboration; other support/elaboration may be absent, incorrect or unclear/obscured by weaknesses in language; or
- inconsistent facility in sentence formation and word choice present (meaning may be unclear and may be occasionally obscured); **or**
- efforts at paraphrasing may result in a number of sentence and word form errors, but meaning is not usually obscured, or there are efforts at paraphrasing, but they do not move sufficiently away from exact wordings and/or structures in the source text; **or**
- inconsistent facility in expressing connections between and among ideas (connections exist but are not effective)

2) A response at this level is marked by flaws in presentation of information or language:

- significantly incomplete, inaccurate, or unclear presentation of principal ideas and key supporting points; **or**
- consistent lack of facility in sentence formation, word choice, word forms and/or connection between and among ideas; **or**
- efforts at paraphrase usually unsuccessful or very limited attempts at paraphrase

1) A response at this level exhibits one or more major flaws:

- little or no comprehensible presentation of principal ideas and key supporting points required by the task
- failure to connect points to the required task
- pervasive language errors that make it difficult for the reader to derive meaning
- text too brief or too borrowed to allow for judgment of writing proficiency

Appendix B

A Passage Used in This Study

Right Brain/Left Brain

The left and right sides of your brain process information in different ways. The left side is logical, rational, linear, and verbal. The right side, on the other hand, processes information intuitively, emotionally, creatively, and visually. Left brains think in words, whereas right brains think in pictures. People who depend more on the

left side of their brain are list makers and analysts. They are detailed, careful, and organized. In contrast, right-brained people are visual, intuitive, and sensual. When a left-brained person has to make an important decision, he or she makes a mental list of all the factors involved and arrives at a decision only after careful analysis. When a right-brained person has to make the same decision, on the other hand, he or she is more likely to base it on intuition and feelings. For example, a left-brained automobile shopper will consider a car's cost, fuel efficiency, and resale value, whereas a right-brained shopper bases a decision on how shiny the chrome is, how soft the seats are, and how smoothly the car drives. Of course, no one is 100 percent left-brained or 100 percent right-brained. Although one side may be stronger, both sides normally work together.

(Extracted from Oshima and Hogue, 2007, p. 109)

Appendix C

Model Summaries Written by Native English Speakers

Model 1:

The right side of the brain processes information intuitively, whereas the left side processes information more logically. Some people tend to use the left side of their brain more and some people use the right, which can lead to different factors being taken into account in decision-making, but despite differences in dominance both sides of the brain normally work together. (60 words)

Model 2:

The left and right sides of the brain process information in different ways: the left side is logical and rational, dealing in words; while the right side is intuitive and creative, dealing in pictures. Each person has a balance of these characteristics, although one side may be stronger than the other. (51 words)

Appendix D

Questions Answered by Raters

Q1) Current academic position (e.g., Ph.D. student, lecturer): ( )

Q2) How many years have you taught at university or elsewhere?

At university ( ) years / Elsewhere ( ) years

Q3) Have you marked summaries before? Please choose one. (Yes/No)

Q4) Have you used the ETS's (Educational Testing Service) rubrics in marking written compositions? Please choose one. (Yes/No)

Q5) How long did it take you to complete marking the whole summaries this time? ( ) hours

Q6) Please evaluate the overall difficulty in marking the summaries by using the ETS's rubrics. Please choose one. ( )

←==========→

(1) very easy (2) fairly easy (3) fairly difficult (4) very difficult

Q7) Following Q6, why do you think so?

Q8) In what cases, did you find it difficult to mark the summaries?

Q9) Do you have any opinions or requests in order to improve the rubrics?

Q10) Other comments