Toward Establishing Criteria for Evaluating Paragraph Writing of Pre-Intermediate Learners of English

Kaoru Mita
Jissen Women's Junior College

Atsuko Shimoda
Jissen Women's Junior College

Japanese students' lack of English writing skills has been a long-standing problem. It not only originates from students' lack of grammatical knowledge and vocabulary, but also from the availability of fewer opportunities to write essays in English. Further, the current standards for evaluating the essays written by Japanese beginner-level learners in English are not established sufficiently, resulting in a heavy grading burden on the teachers. This study establishes an evaluation standard to assess the English essays written by Japanese beginner-level English learners. An "analytic scoring" is conducted for each essay, involving criteria ranging from level 1 to 4, while another group of evaluators perform a "holistic scoring", judging the quality based on their overall impression, on a scale from 1 to 4. The reliability and validity of the "analytic scoring" are established through the correlation analysis of both evaluations. Text mining is used to analyze the explanation of reasons provided by the holistic scoring evaluators to the highest-rated essays and identify the factors determining the quality of English writing.

Keywords: second language writing, rubrics, paragraph writing, topic development, analytic scoring, holistic scoring, general impression, correlation analysis, content analysis, text mining

INTRODUCTION

The majority of university students' English proficiency is at the A2 level (Basic User) according to the CEFR (Common European Framework of Reference for Languages), as indicated by the results from the universities implementing GTEC Academic.[1)] What is even more concerning is the low level of English composition skills. In 2017, the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT) conducted an English proficiency survey for third-year high school students. This survey included an opinion development writing test designed to "measure the ability to persuasively write and express one's opinions and thoughts on a given theme within a limited time." As a result, it became evident that a significant number of students were unable to write a coherent English composition, with 15.1% of students being excluded from grading due to leaving the question unanswered or scoring zero.

The deficiency in English composition among students is not simply due to a lack of grammatical knowledge or vocabulary. A significant factor is that students have overwhelmingly few opportunities to write comprehensive content in English up until they graduate from high school. One of the reasons for this is that there are still no established evaluation standards for English essays targeted at beginner-level

Japanese learners. This leads to teachers having to grade in an ad-hoc manner, resulting in a significant burden on them (Eguchi, 2012; Oi, 2015; Hirabayashi, 2016).[2)]

As a new development in high school English education, under the 2022 high school curriculum guidelines, a subject called "Logic and Expression" has been newly added. This indicates that MEXT is placing greater emphasis than ever before on sentence structure and communication skills in paragraph writing in high school English education (Yamamoto, 2020).

In this paper, an "analytic scoring" and "holistic scoring" are compared. In the analytic scoring in this paper, the evaluation criteria for each of the four levels are specified by applying Wiseman's (2012) Topic Development evaluation items. The holistic scoring in this paper is based on impressions (hereinafter referred to as "impression scoring"), in which the essay is simply rated from 1 to 4 based on whether it is good or bad. Both analytic scoring and impression scoring are conducted by multiple evaluators. The aim is to establish the reliability and validity of the "analytic scoring rubric" (evaluation criteria) by analyzing the correlation between the two scorings. In addition, the impression scorers are asked to write the reason for the essay with the highest evaluation. Their comments are analyzed by text mining[3)] to clarify the factors that determine the high quality of the English text.

PREVIOUS RESEARCH

In the English writing evaluation rubric used by Mita & Shimoda (2021), the concept of Topic Development by Wiseman (2012) was applied as the evaluation criteria for "quality of content." In this study, the authors will verify whether this setting of Topic Development has reliability and validity.

The evaluation methods for English writing can be broadly divided into two categories: holistic scoring and analytic scoring, and comparative studies have been conducted on them. Carr (2000) has clearly argued for the significance of distinguishing between these two writing evaluation methods, and in his empirical research, he has stated that "analytic scoring was more useful for learners than holistic scoring, because the general comments provided in the various subscales' band descriptors could prove helpful to students" (p.212).

The most famous writing rubric is the ESL Composition Profile by Jacobs et al. (1981). This rubric calculates the total score by five items (content: 30 points, organization: 20 points, vocabulary: 20 points, language use: 25 points, and mechanics: 5 points). Jacobs et al. argue that such an analytic assessment has the advantage of providing diagnostic feedback to the learner because it is assessed for each of the subcomponents that make up writing proficiency. The rubric describes the four criteria of content as follows:

TABLE 1
THE CONTENT OF ESL COMPOSITION PROFILE

Level	Criteria	Comments
Level 30-27	Excellent to very good	Knowledgeable, substantive, thorough development of thesis, relevant to assigned topic
Level 26-22	Good to average	Some knowledge of subject, adequate range, limited development of thesis, mostly relevant to topic, but lacks detail
Level 21-17	Fair to poor	Limited knowledge of subject, little substance, inadequate development of topic
Level 16-13	Poor	Does not show knowledge of subject, non-substantive, not pertinent, OR not enough to evaluate

Jacobs et al.（1981)

Although this descriptor clarifies the criteria for content assessment considerably, it still seems to rely on subjectivity. Thus, the content criteria need to be further refined to make them more reliable.

Wiseman (2012) has used a holistic scoring rubric and an analytic scoring rubric to assess the writing of ESL (English as a second language) learners as a way of assessing L2 writing. The results showed that both rubrics were effective in assessing the writing of ESL learners; however, she concludes that while both rubrics explicitly categorize learners' abilities, the analytic scoring rubric is functionally superior to the holistic scoring rubric because it shows more clearly the differences between levels of ability. She also argues that while the holistic scoring rubric has the advantage of saving time and money in the short term, the analytic scoring rubric is the more preferred scoring method in the long term for the purposes of diagnosing and classifying learners' writing ability. The evaluation criterion for "quality of content" in English writing in this paper is an adaptation of the analytic scoring rubric by Wiseman, as shown in the table below:

TABLE 2
THE TOPIC DEVELOPMENT OF WISEMAN'S ANALYTIC SCORING RUBRIC

Points	Descriptors
6 points	● Provides 2+ convincing points related to topic ● Thorough development of topic
5 points	● Provides 2+ points that adequately support topic ● Substantial development of topic
4 points	● Provides 2+ points that directly relate to topic ● Adequate development of topic
3 points	● Provides 1-2 points mostly related to topic with occasional digressions ● Provides some development of topic
2 points	● May provide 1-2 points directly or indirectly related to topic ● Limited development of topic
1 point	● Fails to provide related support

Wiseman, 2012, p.91-92

Ghalib & Al-Hattami (2015) evaluated the writing of EFL (English as a Foreign Language) learners using holistic scoring without specified criteria and found significant variation in the evaluations among different assessors, leading to an inability to accurately discern between the writing abilities of various learners (p. 226). Given these results, they initiated research into developing evaluation standards that could ensure both reliability and validity, minimizing inconsistencies among different assessors' evaluations. This involved conducting a comparative analysis of holistic scoring rubrics and analytic scoring rubrics using psychometric statistical methods. They revealed that the two evaluation methods strongly correlated. Furthermore, they pointed out that the analytic scoring rubric is more effective than the holistic scoring rubric as it can clearly present learners with the criteria for measuring writing skills. They also refer to Wiseman's (2012) research as a study aimed at enhancing the precision and consistency of evaluation methods.

In Japan, Nakanishi & Akabori (2004) point out the differences in the aspects of evaluating English essays between Japanese English teachers and native English teachers. Specifically, they note that Japanese evaluators tend to focus on "content, word choice, and grammatical accuracy," while American evaluators focus more on "paragraph and composition, and connections between sentences." They emphasize the importance of native teachers' evaluations that pay attention to whether the writer's intention is sufficiently conveyed to the reader.

Eguchi (2012) proposes the use of rubrics as a way to break the status quo of writing education, which tends to rely on evaluations dependent on the instructor's experience. She identifies the reasons why rubrics have not become widespread in educational settings: the time and effort required to create and evaluate using rubrics, and the skepticism regarding their reliability due to variances in evaluation between evaluators. To resolve these issues, Eguchi devised a concrete writing rubric that applies the technique of "content analysis" and investigated its utility. In this process, three points were tested: 1) whether fair and objective evaluation of learning outcomes can be done in a short time, 2) whether grading can be done efficiently, and 3) whether it is helpful for student learning. As a result, it was confirmed that the writing rubric developed by Eguchi leads to efficient evaluation and improvements in students' understanding of content and motivation to learn. However, it is pointed out that multiple evaluators are necessary to ensure reliability since the evaluator was only the author, and that the development and use of rubrics with more detailed and specific evaluation criteria will be needed in the future.

Hirabayashi (2016) conducted a study aimed at developing a rubric for "free English composition".[4)] While output-oriented English education activities are demanded due to the internationalization of society in Japan, learning of free English composition, which involves expressing one's thoughts and opinions in English, has not necessarily been widespread. One reason for this, as Hirabayashi points out, is that the rubrics for evaluating free English composition are not sufficiently established. This can cause problems for learners in their learning of free English composition, and instructors may also be confused about how to teach it (p. 23).

Hirabayashi's rubric targets "beginner English learners" and "advanced English learners." The reason for this is that the six levels of the CEFR (ranging from A1 to C2) are not suitable for accurately assessing the writing proficiency of Japanese beginner English learners, and the evaluation descriptors for advanced English learners are often too abstract to be consistently useful. Therefore, for beginner learners, a rubric is designed to evaluate the CEFR-J A2 level[5)], which is said to encompass around 80% of Japanese learners, while for advanced learners, rubrics are designed to assess the CEFR B2 and C1 levels. They are analytic scoring rubrics using statistically validated "specific factors" as the assessment perspective.

He investigated the correlation between the holistic scoring and his analytic scoring, both for beginner-level and advanced-level compositions. The result showed that there was a strong positive correlation between the two assessments, and that his rubric could be a good substitute for the holistic scoring. However, he focused on some outliers; the rubric score was low, but the holistic score was high. In these outliers, he found that "interesting content," "excellent and interesting angle" and "excellent insight into the content" were rated higher in holistic scoring. With respect to this finding, he points out that the cause of the outliers is related to the superiority or inferiority of the content, and that his evaluation perspectives may not be sufficient to evaluate the content aspects.

Thus, research on writing assessment methods has attempted to develop rubrics that assure the usefulness of analytical assessments and their reliability and validity. However, there has been a lack of research specifically addressing the measurement of content quality in writing. This is due to the subjective nature of content quality, which makes it challenging to quantify. In light of this, the current study aims to examine the reliability and validity of criteria used to measure quality of content in the rubric.

RESEARCH QUESTIONS

(1) Does Topic Development serve as a reliable and valid criterion for assessing the quality of content in English writing evaluation?
(2) What factors determine the high quality of content in English writing?

RESEARCH METHOD

English Essays to Be Assessed

The subjects of this study were 161 first-year students taking the compulsory Integrated English[6)] course at a junior college in Tokyo. They took the GTEC Academic at the beginning of the semester (in May) and post-test (in January of the following year). The breakdown of the students who took the targeted writing test was as follows: 76 from the Japanese Communication Department and 85 from the English

Communication Department. The students' English levels, when mapped onto the CEFR levels using the GTEC Academic scores, were divided into the following four groups.

(i) CEFR B2, B1 level: 27 students (16.8%)
(ii) CEFR A2 level upper: 43 students (26.7%)
(iii) CEFR A2 level lower: 62 students (38.5%)
(iv) CEFR A1 level: 29 students (18.0%)

The writing test was administered online in both pre- and post-test. Specifically, the students took the test using the learning management system (LMS) *manaba* ver. 2.95 (Asahi-net). Because students were unable to attend school due to the Corona pandemic at the time of the study, students took both the pre-test in May and the post-test in January of the following year by accessing *manaba* from their home computers while taking the online course (via zoom).

The time required for the writing test, including brainstorming, was 15 minutes. The test screen provided not only a space for writing input, but also a space for the test-taker to record their personal brainstorming content. The topic of the writing test was "my favorite place." The instructions for the essay are as follows.

Writing test topic: My favorite place

"Write down a place you would like to visit and three reasons why you would like to go there. It can be either overseas or in Japan. Begin with the following expression."

The place I want to visit most is (   ). There are three reasons.

Evaluation Criteria for Students' English Writing

The results of the students' writing tests were evaluated using a rubric of the following four assessment items. This study focuses on (iv) "content" among the following four items.

Assessment items of the writing test analysis rubric.

(i) Fluency: word count
(ii) Structure of the essay: presence or absence of discourse markers set in class
(iii) Grammar accuracy: presence/absence of specified grammatical errors set in class
(iv) Content: quality of details

Topic Development, included in research question (1), is originally one of the items in Wiseman's (2012, p.91) analytic scoring rubric. Wiseman established Topic Development, which reflects the "quality of English writing," and created an analytic scoring rubric to evaluate it with six levels of criteria. In this survey, the authors applied this item and established four-level evaluation criteria as follows. The English sentences numbered (1) to (4) below are written by students who received each respective evaluation (grammatical errors are left uncorrected). "Detailed sentences" in the four-level evaluation refer to sentences that provide detailed explanations or specific examples, thereby adding persuasiveness to the argument.

Evaluation Criteria for Topic Development (4 Levels)

(1) No detailed sentences (Level 1)

Sample: The place I want to visit most is Okinawa. There are three reasons. First, I like hot place. Second, beach is beautiful and rich in nature. And finally, I've never eaten Okinawa food, and I want to try them.

(2) Sentences with a few detailed sentences but limited content (Level 2)

Sample: The place I want to visit most is Hawaii. There are three reasons. I want to eat delicious food in Hawaii. For example, I like garlic shrimp. I want to surf because my father is doing it. I want to go to the beach in the evening. Because the setting sun is beautiful.

(3) Detailed sentences with multiple sentences that develop the topic and deepen the content (Level 3)

Sample: The place I want to visit most is Korea. there are three reasons. First, I want to go to different cities and go shopping. because I've been watching Korean dramas every day lately. therefore, the reason I want to go to Korea is probably influenced by the Korean drama. Second, I want to buy Korean cosmetics. they are so good for my skin. so, I often use Korean cosmetics. Third, there are many delicious foods in Korea. I especially like spicy food. so, I want to eat a lot of spicy food when I go to Korea.

(4) Particularly good among “Level 3” (Level 4)

Sample: The place I want to visit most is Okinawa. There are three reasons. First, I have been to Okinawa only once. In addition, the weather was very bad when I went to Okinawa. So, I want to see the beautiful sea of Okinawa on a sunny day. Next, I want to eat delicious food in Okinawa. When I went to Okinawa, I ate delicious food such as Okinawa soba and Seta Andagi. And I like them very much. So, if I go to Okinawa again, I want to eat them. Finally, I like Okinawa time. Because it is very slowly, we can relax in Okinawa time. That’s why I want to go to Okinawa.

Evaluators and Evaluation of the Students’ English Essays

A total of 322 English essays from 161 students’ writing tests at the beginning (in May) and the end (in January following the year) were assessed by the following six assessors.

Breakdown of evaluators (the names of (1) to (4) are pseudonyms)

(1) ATSUKO (Japanese)
(2) EIKO (Japanese)
(3) BRIT (native English speaker)
(4) JANEY (native English speaker)
(5) MITA and SHIMODA (the authors)

The four external evaluators (1) to (4) were commissioned through the English manuscript correction service company[7]. These four external evaluators were asked to give an “impression evaluation” ranging from level 1 to 4, as well as provide a comment on why they assigned a level 4 (Excellent to very good) rating. They were shown some students’ English manuscripts evaluated by the aforementioned “Topic Development” criteria as a sample.

Requests to the External Scoring Evaluators

These are essays written in 15 minutes by junior college students who do not specialize in English. Please rate your impression of the essays without regard to spelling or grammatical errors and give us a brief comment on the best essay, saying why you rated it the best.

- Level 1. Very poor
- Level 2. Fair to poor
- Level 3. Good to average
- Level 4. Excellent to very good

Each of the authors (Mita and Shimoda) gave an “analytic scoring” of 1 to 4 based on the evaluation criteria for Topic Development to the 322 English essays, and then discussed and modified their evaluations if they differed from each other.

Analysis Method

For research question (1), a correlation analysis was conducted between each of the four external evaluators' impression scoring and the two authors’ adjusted analytic scoring based on Topic Development.

For research question (2), text mining[7] was conducted on each of the four external evaluators’ reason comments for assigning a level 4 rating in the impression scoring. KH Coder 3 was used for the text mining.

RESULTS

Correlation Analysis of Analytic Scoring and Impression Scoring

The impression scorings of 322 student essays in English by four external evaluators were compared with the analytic scorings by the two authors. Table 3 shows the number of essays at each of the four levels assessed by evaluators.

TABLE 3
NUMBER OF ENGLISH ESSAYS AT EACH OF THE FOUR LEVELS ASSESSED BY EVALUATORS

	Level 1	Level 2	Level 3	Level 4	Sum
ATSUKO	105	91	92	34	322
EIKO	94	143	61	24	322
BRIT	56	101	123	42	322
JANEY	5	30	156	131	322
Authors	103	101	86	32	322

Table 4 shows the Pearson's product-moment correlation coefficients, means and standard deviations for the impression scorings of each of the four evaluators and the authors' analytic scorings.

TABLE 4
CORRELATION COEFFICIENTS, MEANS AND STANDARD DEVIATIONS FOR THE IMPRESSION SCORINGS OF THE FOUR EVALUATORS AND THE AUTHORS' ANALYTIC SCORINGS (N=322)

	ATSUKO	EIKO	BRIT	JANEY	Authors	*M*	*SD*
ATSUKO	—					2.171	1.0041
EIKO	$.782^{**}$	—				2.047	.8830
BRIT	$.820^{**}$	$.772^{**}$	—			2.469	.9276
JANEY	$.738^{**}$	$.673^{**}$	$.731^{**}$	—		3.283	.6955
Authors	$.830^{**}$	$.793^{**}$	$.810^{**}$	$.678^{**}$	—	2.146	.9829

$^{**}p < .01$

All correlation coefficients are significant at the 1% level, with the correlation coefficient between EIKO and JANEY and between teachers and JANEY indicating a 'relatively strong correlation' (<0.7). All other correlation coefficients indicate a "strong correlation" (>0.7).

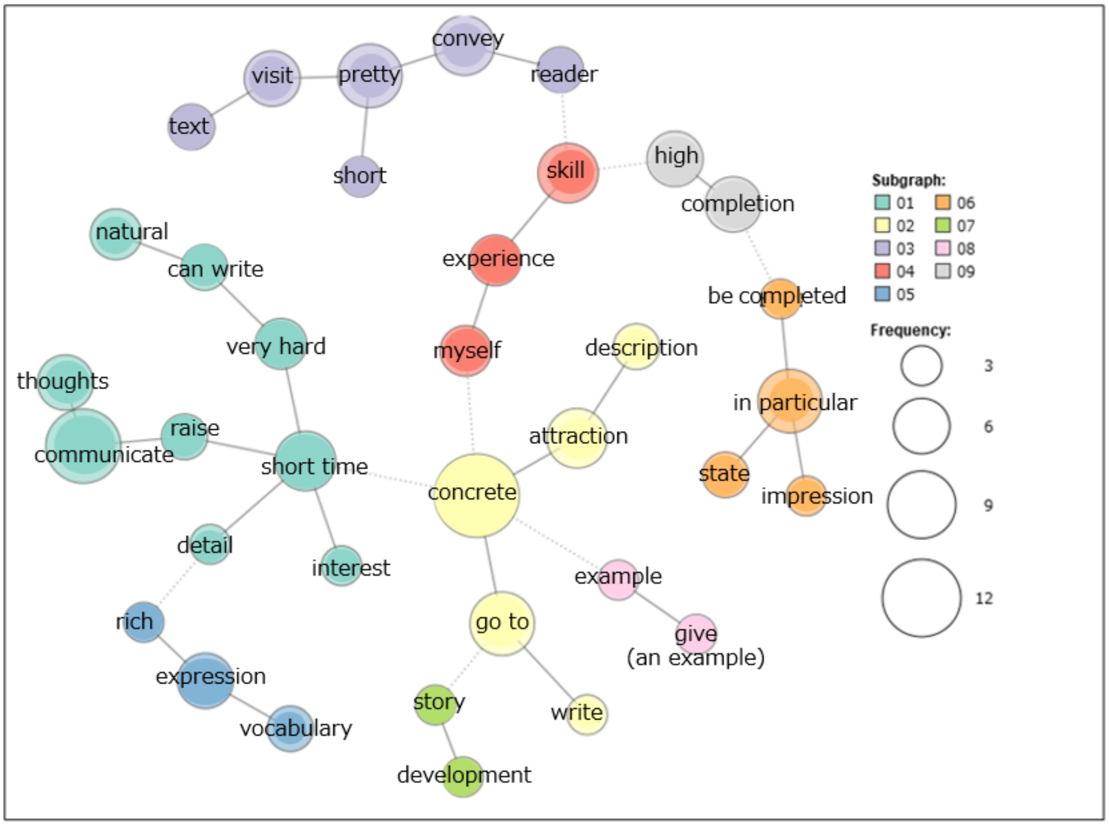
Text Mining of Comments by Four Evaluators Who Rated Based on Their Impressions

For the Level 4 essays evaluated by the four external evaluators, the explanation of why each rater judged the essay to be the best was analyzed by text mining.

ATSUKO

A co-occurrence network between characteristic words and phrases of Atsuko is as follows (Fig.1):

FIGURE 1
ATSUKO



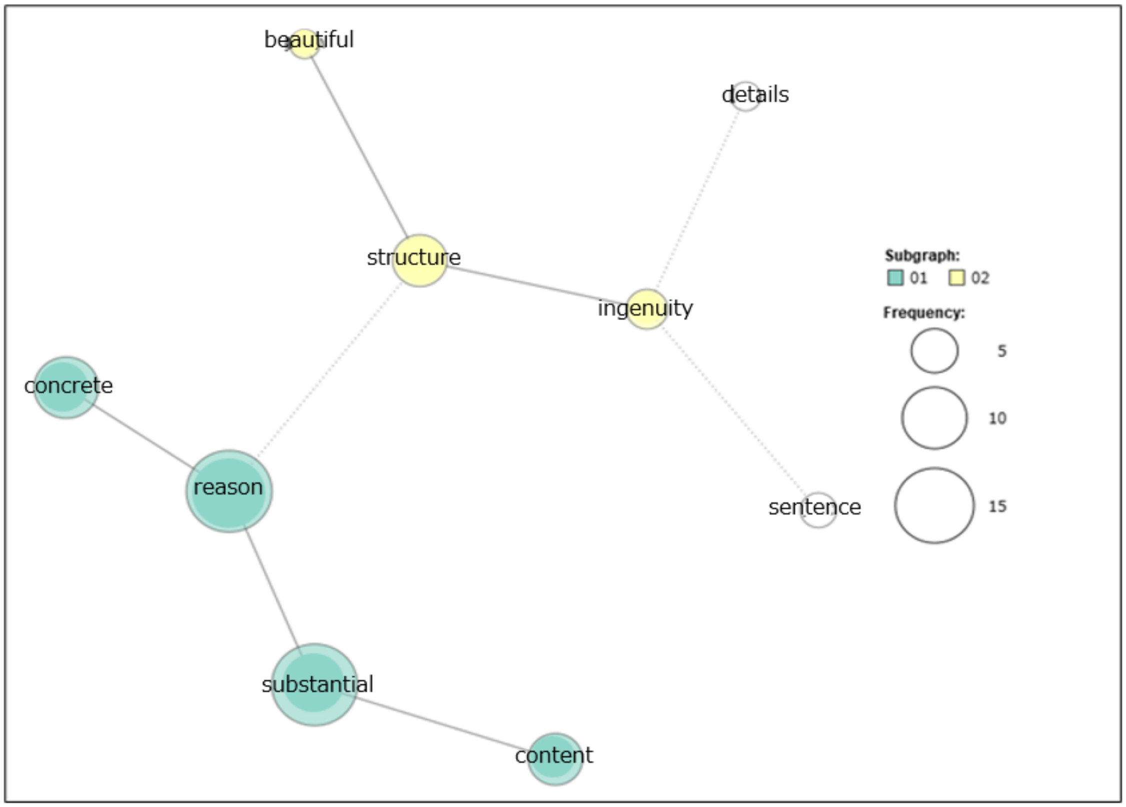
The words that appear frequently in Figure 1 are ‘short time,’ ‘concrete,’ ‘pretty,’ ‘skill’ and ‘expression.’ The word “short time” strongly co-occurs with the following word groups: ‘interest,’ ‘detail’, ‘raise,’ ‘communicate,’ ‘thoughts’ and ‘very hard.’ The word ‘details’ also co-occurs with the word groups ‘expression,’ ‘rich’ and ‘vocabulary.’ ‘Concrete’ has strong co-occurrence with the word groups ‘go to,’ ‘attraction’ and ‘description.’ and also with ‘example,’ ‘give,’ ‘myself,’ ‘experience’ and ‘skill.’ From this, it can be inferred that ATSUKO places the criteria for evaluating excellent essays on the ability to write within a time limit, richness of vocabulary, and the ability to give specific examples, such as personal experiences. The following are excerpts from the comments:

> *I think you have written very well in a very short time. I liked the way you summarised in detail what you learned about France, which you have never visited, and what interested you. Your English writing is also very good.*
>
> *I could feel your strong passion for Disney World from the whole text. I like the way you have concretely and vividly summarised your past experiences and what you want to do in the future.*

EIKO

A co-occurrence network between characteristic words and phrases of Eiko is as follows (Fig. 2):

FIGURE 2
EIKO



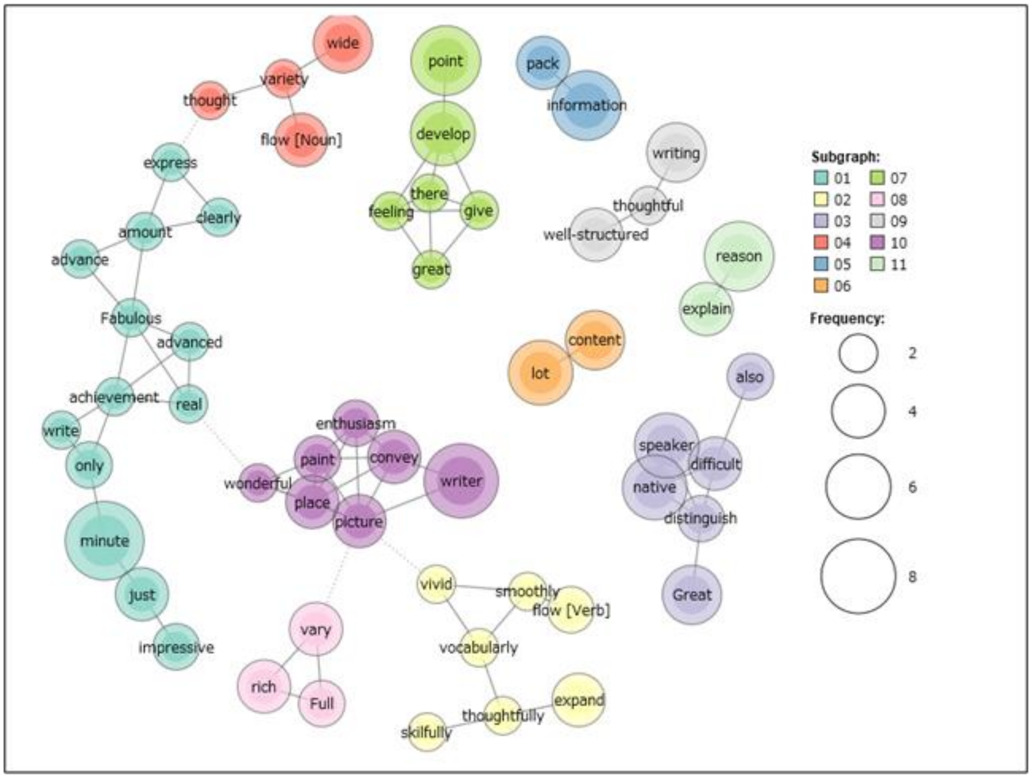
The words that appeared most frequently in Figure 2 were 'reason' and 'structure.' The word 'reason' has strong co-occurrence with 'concrete,' 'substantial' and 'content.' 'Reason' also co-occurs with 'structure,' and 'structure' strongly co-occurs with 'beautiful' and 'ingenuity,' indicating that ERIKO bases her evaluation of a good essay on the concreteness and fullness of the reasons, and on the ingenuity and beauty of the structure. The following are excerpts from her comments:

> *The reasons are concrete and substantial.*
> *Well-structured and detailed reasons.*

BRIT

A co-occurrence network between characteristic words and phrases of Brit is as follows (Fig 3):

FIGURE 3
BRIT



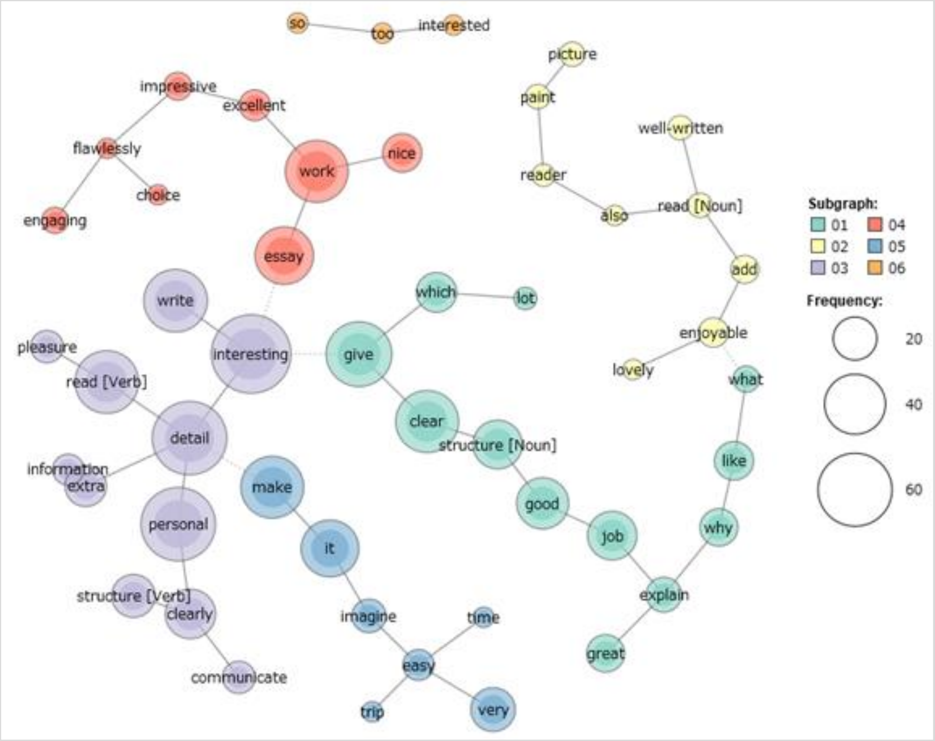
The most frequent words in Figure 3 are ‘minute,’ ‘vocabulary,’ ‘native,’ ‘information,’ ’content,’ ‘develop,’ ‘well-structured’ and ‘picture.’ ‘Minute’ has strong co-occurrence with ‘just,’ ‘only’ and ‘achievement.’ ‘Vocabulary’ strongly co-occurs with ‘vivid,’ ‘smoothly,’ and ‘thoughtfully.’ ‘Native’ strongly co-occurs with ‘distinguish’ and ‘difficult; ‘information’ strongly co-occurs with ‘pack’; ‘content’ with ‘lot’; ‘develop’ with ‘point’; ‘well-structured’ with ‘thoughtful; ‘picture’ with ‘convey,’ ‘paint,’ ‘enthusiasm,’ and ‘place.’ BRIT’s criteria for assessing a good essay are: ability to write within a time limit, richness of vocabulary use, native-like, richness of information, volume of content, development of the topic, good organization, and the ability to communicate with enthusiasm so that the reader can visualize the scene. The following are excerpts from the comments:

> *The writer painted a vivid picture of the city and the points flowed smoothly into each other. Also, the vocabulary was rich and varied.*
>
> *Very thoughtful writing and a lot of content for just 15 minutes. Well structured, too.*
>
> *Paints a really vivid picture of the place and conveys a lot of enthusiasm. Difficult to tell apart from a native speaker.*
>
> *Full of rich, descriptive and varied vocabulary. Each point is developed in detail.*

JANEY

A co-occurrence network between characteristic words and phrases of Janey is as follows (Fig.4):

FIGURE 4
JANEY



The most frequent words in Figure 4 are ‘structure,’ ‘detail,’ ‘work’ and ‘read.’ ‘Structure’ strongly co-occurs with ‘clear’ and ‘good.’ ‘Detail’ strongly co-occurs with ‘personal,’ ‘information’ and ‘interesting’; ‘work’ with ‘excellent,’ ‘impressive,’ ‘flawlessly’ and ‘engaging’; ‘read’ with ‘well-written,’ ‘enjoyable,’ ‘lovely,’ ‘paint’ and ‘picture.’ JANEY’s criteria for a good essay are good organization, personal, detailed and interesting content, an engaging and memorable essay, and an enjoyable read that allows the reader to visualize the scene. The following is excerpts from the comments:

> *This is an impressive essay about why you would like to visit France. it is flawlessly written, and you have clearly structured your work. The interesting details with personal reasons to back up your choice make it engaging. Excellent work!*
>
> *Your essay was interesting to read because of the clear picture you have painted of what you would like to do in England. Great work!*
>
> *It is very easy to imagine your trip to the US after reading your well-structured and detailed reasons for wanting to visit. Well done!*

DISCUSSION

As for the research question (1) (Does Topic Development serve as a reliable and valid criterion for assessing the quality of content in English writing evaluation?), a correlation analysis was conducted between the “impression scoring” by four external evaluators and the “analytic scoring” by the authors. A relatively strong correlation was shown at the 1% level for any combination of evaluators. Therefore, it is considered that reliability and validity have been confirmed for the four-level evaluation criteria of the analytic scoring applying Wiseman’s (2012) Topic Development to the “quality of content.”

As for the research question (2) (What factors determine the high quality of content in English writing?), the comments of four external evaluators (two Japanese and two native English speakers) were analyzed by text mining. As a result, some common concepts about good writings were extracted.

It is noteworthy that words related to Topic Development appeared in the comments of all four external evaluators: ‘concrete’ and ‘detail’ for the two Japanese, and ‘develop’ and ‘detail’ for the two native speakers. This leads to the conclusion that concepts related to ‘concrete’ and ‘detail’ are major factors in the essay being judged as excellent.

Next, it was observed that the word “structure” frequently appeared in the comments of three out of four external evaluators. Even in the comment of the one evaluator where the word “structure” did not appear, there is a description saying “a very well-organized sentence,” which is presumed to suggest “structure.” Therefore, it can be said that the elements that determine the “quality of content” in Research Question (2) are the description being “concrete” and the essay being written with an orderly “structure.”

Furthermore, from the results of text mining, it was possible to capture concepts characterizing excellent content in English. Descriptions co-occurring with “vocabulary” such as “rich,” “vivid,” “smoothly,” and “thoughtfully” allowed for an understanding of the importance of “vocabulary” in a good essay. In addition, the co-occurrence of “information” & “pack”, and “content” & “lot” indicates that a “large amount of information” is highly valued.

In the comments from native evaluators, “paint” and “place” appear as words co-occurring with “picture.” It is interesting that both of the native evaluators shared comments attributing high evaluations to descriptions that vividly enable one to imagine the scene. These co-occurring words can be interpreted as sub-concepts of “concreteness,” which is one of the essential elements of an excellent essay.

Lastly, it is interesting that two out of the four external evaluators highly appreciate that the essay was completed in a limited time, as indicated by phrases like “short time” and “just 15 minutes.” Although this is not a factor directly related to the quality of the content, it suggests that the ability to write an impressive essay within a limited time could also become one of the criteria for evaluating writing skills.

Previous studies have shown that the “overall evaluation” of English essays varies depending on the evaluator. In this study, the comments of the four external evaluators also indicated that they each evaluated the essays from their own unique perspectives. For example, JANEY wrote the comment with the intention of addressing the students directly, as can be seen from the use of “you” as the subject and the expressions of encouragement (e.g., *Your essay was interesting to read because of the clear picture you have painted. Great work!*) Janey was the one who awarded the highest evaluation, Level 4, the most among the four evaluators. On the other hand, Eiko was the one giving stringent evaluations with the bare minimum of words (e.g., *reasons are specific and substantial*). While Eiko had the lowest average evaluation and Janey had the highest, the result of the correlation analysis of their evaluations showed a relatively strong correlation (.673**). Although it has been pointed out in previous studies that there is variation in the evaluation by the evaluators, the results of the present study confirm that there is a correlation in the impression scoring of Levels 1 to 4, even between evaluators with different evaluation stances.

As described in Nakanishi & Akabori (2004), the present study also revealed that Japanese and native English evaluators have different perspectives on evaluation. Japanese evaluators tended to evaluate the concreteness of content and the structure of the essay, whereas native evaluators tended to evaluate the richness of the vocabulary and the content that the reader can read while imagining the scene.

These concepts extracted from the text-mining results can be useful not only for teachers to evaluate students’ English writing, but also for teaching students to write in English. For instance, it can be helpful to introduce some tools such as a checklist or a rubric with descriptors reflecting each of the concepts identified in this study (e.g., “Does it contain specific examples?”, “Does it include your own experiences?”, “Does it use many types of words?”, “Can the reader picture the scene?”, “Are the introduction, main thesis, and conclusion written?”). By using a checklist that includes each of the concepts clarified in this study as individual items and having students confirm them, it might become easier to make students aware of the conditions for obtaining good evaluations and understanding the traditionally ambiguous criterion of “quality of content”.

SUMMARY

Multiple evaluators conducted an “analytic scoring” applying Wiseman’s (2012) Topic Development to 322 student English essays, as well as an “impression scoring” simply based on the quality of the essays. The reliability and validity of the analytic scoring rubric were confirmed through a correlation analysis between the two evaluations. Furthermore, the reasons for the highest evaluation given by each of the four external evaluators to English essays were analyzed using text mining, and elements determining the “quality of content” were extracted. While utilizing these results in the teaching of English essay writing to students, we would like to continue our research on the evaluation criteria for the “quality of content” of English essay writing.

ENDNOTES

[1.] GTEC® (Global Test of English Communication) is a score-based English two- or four-skills test to measure English language proficiency conducted by Benesse Corporation. The two-skills test was used for this study.

[2.] Many researchers have pointed out that there are no established criteria for the assessment of English essay writing by Japanese learners and that this places a heavy burden on teachers. For example, Eguchi (2012) states that “It is difficult to say that rubrics have become a common method of assessment. One of the reasons for this is that it takes time and effort to create rubrics and to assess using rubrics.” (pp.73-74), noting that the evaluation of writing is time-consuming and usually results in fluctuations in evaluation, even within the evaluator.” (p.82)

[3.] Text mining is a method of content analysis that cuts text data word by word, analyses it in a quantitative way and visualizes the results.

[4.] Hirabayashi defined the term “free English composition” as “an expository or argumentative essay on a given theme in which the writer constructs his or her own ideas and opinions, placing emphasis on content rather than errors in language form, and constructs an English text within a certain time frame” (p.61)

[5.] CEFR-J is a customized version of the CEFR for Japanese learners of English.

[6.] Integrated English is a compulsory English course for all junior college students, and both Integrated English a (first semester of year 1) and Integrated English b (second semester of year 1) are offered in the first year. Both the first and second semesters consist of two lessons per week, one of which is taught by a Japanese teacher and the other by a native English teacher. In the survey year, the online classes started late due to the spread of corona infection, so the first survey was conducted in May.

[7.] This company has English correction teachers all over the world and has a track record of 11 years.

REFERENCES

Carr, N.T. (2000). A comparison of the effects of analytic and holistic rating scale types in the contest of composition tests. *UCLA Issues in Applied Linguistics*, *11*(2), 207–241.

Eguchi, M. (2012). Creating specific rubrics for writing English sentences using the content analysis. *Shimane Journal of Policy Studies*, *24*, 73–84.

Ghalib, T.K., & Al-Hattami, A.A. (2015). Holistic versus analytic evaluation of EFL writing: A case study. *English Language Teaching*, *8*(7), 225–236.

Hirabayashi, K. (2016). Development of a Rubric for the Free Writing of CEFR-J A2-Level Learners of English, *International Journal of Curriculum Development and Practice*, *39*(2), 61–70.

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

Mita, K., & Shimoda, A. (2020). An analysis of Japanese Students’ Improvement on EFL Writing---- Changes of errors by L1 transfer and the ability of expression in essays by students who improved fluency. *Jissen English Communication*, *50*, 6–33.

Mita, K., & Shimoda, A. (2021a). Analysis of students’ improvement in English writing. Part 2: Changes in writing proficiency with focused instruction in grammar, organization, and logic. *Jissen English Communication*, *51*, 14–46.

Mita, K., & Shimoda, A. (2021b). An analysis of students' writing improvement observed in different proficiency levels―Changes in writing through focused instructions for overcoming weaknesses. *Jissen Women's Junior College Review*, *42*, 63–83.

Nakanishi, C., & Akahori, K. (2005). Differences in evaluation of Japanese college students' writing between Japanese English teachers and native English teachers. *Japan Journal of Educational Technology*, *28*, 229–232.

Oi, K. (2015). What instruction and assessment are needed for writing in the age of four-skills testing? *The English Teachers' Magazine*, *12*, 10–12.

Wiseman, C.S. (2012). A comparison of the performance of analytic vs. holistic scoring rubrics to assess L2 writing. *International Journal of Language Testing*, *2*(1), 59–92.

Yamamoto, A. (2020). What's new and what's inherited in 'Logic and Representation'. *Chart Network*, *93*, 18–20. Retrieved June 27, 2023, from https://www.chart.co.jp/subject/eigo/cnw/93/93-5.pdf

APPENDIX

The profiles of four external evaluators are as follows:

Atsuko

(i) Nationality: Japanese

(ii) Years of living and experience abroad: Atsuko and her husband lived in New Zealand for seven years after their international marriage. They then moved to Australia together, where they have lived for 14 years. Her total time living abroad is 21 years.

(iii) Experience as an English language corrector: Atsuko has been working as an English language corrector for six years now, and she has experience of living overseas and teaching Japanese at a junior high school in Japan, as well as teaching Japanese to local people. She has received a lot of feedback from her customers that they found the corrections easy to understand.

Eiko

(i) Nationality: Japanese

(ii) Years of living and experience abroad: Eiko has lived in the suburbs of New York, USA, for about 32 years, since 1990. She has worked as a translator for television programs, as a Japanese teacher for students and adults, and as a teacher at a preschool in her area. Prior to coming to the USA, she worked as a high school and junior high school English teacher in Japan.

(iii) Experience as an English language corrector: Since 2011, Eiko has been working as a corrector at her company for 12 years. To date, she has completed over 12,000 English corrections for individuals and companies of all kinds, as well as for students and housewives.

Brit

(i) Nationality: British

(ii) Years of living and experience abroad: Brit has been to Norway and it is her favorite country to visit. In her spare time, she loves learning foreign languages. She has also trained as a tutor of Welsh for adults with the University of Bangor and completed an online course in Spanish to English translation with the University of Cardiff.

(iii) Experience as an English language corrector: Brit has 10 years' experience as a primary school teacher and has also worked as a counsellor with MIND, a UK mental health charity. She has been teaching English on Skype since 2015 and done essay proofreading for students at the local university.

Janey

(i) Nationality: British

(ii) Years of living and experience abroad: Janey has lived abroad for around 20 years, 10 of which were spent teaching English as a second language in South-East Asia. Following this, she became a content writer/manager and freelance editor. She enjoys living in Asia because of the wonderful weather, unique culture and friendly people.

(iii) Experience as an English language corrector: After graduating in English literature in the UK, Janey took a Cambridge TEFLA course to teach English as a second language. She has taught people of most ages and various class sizes and ability levels in Thailand, Hong Kong and the UK. She now enjoys assisting Japanese people with their written English.