The Effects of Japanese EFL Learners’ Proficiency on Evaluating EFL Essay Writing
The Effects of Japanese EFL Learners’ Proficiency on Evaluating EFL Essay Writing

A Thesis

Presented to

The Faculty of the Graduate Course at

Hyogo University of Teacher Education

In Partial Fulfillment

of the Requirements for the Degree of

Master of School Education

by

Shion Onishi

(Student Number: M15182F)

December 2016
Acknowledgements

I would like to give heartfelt thanks to Dr. Tomoyuki Narumi whose comments and suggestions were of inestimable value for my study, without which my research would not be successful. I would also like to thank the member of Narumi seminar, especially Mr. Katsutoshi Nakazawa, Ms. Sayaka Yoshida, and Mr. Hirotaka Nishiyama whose comments made enormous contribution to my work. I would also like to appreciate kind cooperation of the participants as raters and writers to this study. Finally, I would like to express my gratitude to my family, especially my dog Rosch for their moral support and warm encouragements.

Shion Onishi
Kato, Hyogo
December 2016
Abstract

In English education in Japan, teachers mainly teach reading and listening due to university entrance examinations that adopt the points system using an optical answer sheet. However, the government aims for students to be able to learn English in terms of the four skills equally and thus it plans to introduce external tests to university entrance examinations that test students' solid academic prowess by measuring the four skills (reading, writing, listening, and speaking). If the reorganization of university entrance examinations in Japan is realized and the pleiotropic comprehensive evaluation system is adopted, writing and speaking skills will become more important for English education. At the same time, teachers will need to teach writing and speaking to their students, and evaluate them appropriately using performance assessment.

Of the four English skills (listening, speaking, reading, writing), the process of writing, which has achieved more importance in the reorganization of English education in Japan, is a particularly confused and recursive intellectual process (Silva, 1992). Writers must produce sentences by correlating the processes of generating ideas, structuring, drafting, focusing, evaluating, and reviewing. Therefore, all native English speakers (NES's), English as a second language (ESL) learners, and English as a foreign language (EFL) learners must ultimately learn writing skills (Hamp-Lyons & Heasley, 1987). In particular, EFL learners encounter more difficulties in the writing process due to their low level English abilities. On top of that, not only EFL learners but also EFL teachers who oversee the teaching of writing may encounter difficulties because they need to acquire all of the processes of EFL writing, the techniques of teaching writing, and English itself.

To conclude with regard to teaching writing, learners are usually examined on what
they have learned using tests. However, the complex process of writing causes difficulties in evaluation. In order to evaluate writing ability with high reliability, many researchers have attempted to minimize and standardize the raters' effect. Various studies have investigated writing and its evaluation and have revealed that rater training can standardize the raters' effect, but there are numerous factors which affect raters' assessment and the factors have not been clarified. In particular, if the writer and rater are EFL learners, the writing and rating process becomes more difficult due to their English proficiency compared with native English speakers; however, few studies have been conducted with regard to raters' effects on evaluation that target EFL raters. Therefore, it is necessary to clarify the effects of Japanese EFL raters' English proficiency on their performance assessment in order to better understand the differences between NES raters and Non-native English speakers (NNES) raters.

If Japanese EFL raters' own criteria differ due to their English proficiency, the proficiency effect should be standardized in order to increase the reliability of evaluation. What is more, if Japanese EFL raters' proficiency affects their evaluation of Japanese EFL writing and the proficiency effect can be minimized or standardized, it will lead to an increase in teaching because teaching and evaluating are inseparable (Tanaka, 1998). In addition, each classroom contains students of varying English proficiency. If teachers give students the opportunity to write essays and evaluate each other, it is necessary to consider the proficiency effect of both raters and writers.

In this study, Japanese EFL raters' proficiency effects on Japanese EFL learners' evaluations of Japanese EFL essay writing are investigated in order to explicate the specific factors that affect Japanese EFL raters' assessments. 43 Japanese EFL undergraduate and graduate school students participated in this study: three writers and 40 raters. The writer participants comprised three sophomore students, all of whom were
18 years old. Their proficiency levels were A2, B1, and B2 according to CEFR, respectively. The rater participants were divided into two groups according to their placement test scores: 26 participants were lower level (A2) and 14 were higher level (B1, B2).

The writers wrote essays and the raters rated them using a rating scale (content, organization, vocabulary, grammar). The rating scale was composed by arranging the five scales (content, organization, vocabulary, language use, and mechanics) by Jacobs et al. (1981) and the English free writing criteria by Toyama Minami high school in order to make it appropriate to the case of English education in Japan. The original criteria of Jacobs et al. (1981) comprised five rating scales with different weightings, but the weightings of all rating scales of criteria were equalized in this study by using a six-point Likert scale. When removing these weightings, however, the rating scale “mechanics” was excepted because the weighting of this scale was regarded as too light compared with the others. After the writers evaluated the writers’ essays, the two-way ANOVA assessed differences between the raters’ proficiency and writers’ proficiency on each evaluating scale (content, organization, vocabulary, grammar, total) based on the scores of evaluations.

This investigation has shown that English proficiency had little effect on the evaluation of content in this study and the raters seem to have been able to understand the content of all three essays and assess them without linguistic problems.

As regards “organization,” raters assessed the essay written by the A2 writer lower than the essays written by the B1 and B2 writers. In addition, lower proficiency raters gave a higher score to the essay written by the B1 writer than higher proficiency raters. Regarding “vocabulary,” the participating raters agreed that the vocabulary of the A2 writer’s essay was the lowest scoring. However, the assessment did not depend on raters’
English proficiency and English proficiency seems to have had no effect on the evaluation of vocabulary in this study.

Regarding "grammar," raters assessed the essay written by the A2 writer lower than the essays written by the B1 writer and B2 writer. In addition, lower proficiency raters tended to give higher scores to the essays than higher proficiency raters.

To sum up, the raters' proficiency effect is one of the factors affecting raters' judgment and this elucidation could lead to an increased effectiveness of rater training and the reliability of the evaluation of EFL essay writing in Japan.
# Table of Contents

Acknowledgements ........................................................................................................... i
Abstract ............................................................................................................................ ii
Table of Contents ............................................................................................................... vi
List of Tables ...................................................................................................................... viii
List of Figures ..................................................................................................................... ix

Chapter 1 Introduction ....................................................................................................... 1

Chapter 2 Literature review ............................................................................................... 5
  2.1 Writing ......................................................................................................................... 5
  2.2 Measurement, evaluation, and tests ............................................................................. 6
  2.3 Raters' effect on evaluation ......................................................................................... 9
  2.4 Role of writing tests in Japan ..................................................................................... 13
  2.5 Research questions .................................................................................................... 14

Chapter 3 Method ............................................................................................................... 16
  3.1 Participants .................................................................................................................. 16
  3.2 Rating scale ............................................................................................................... 16
  3.2.1 Content .................................................................................................................... 17
  3.2.2 Organization .......................................................................................................... 18
  3.2.3 Vocabulary ............................................................................................................. 19
  3.2.4 Grammar ............................................................................................................... 21
  3.3 Procedure ................................................................................................................... 22

Chapter 4 Results and discussion ...................................................................................... 24
  4.1 Total .......................................................................................................................... 24
4.2 Content.................................................................................................................... 25
4.3 Organization............................................................................................................. 26
4.4 Vocabulary............................................................................................................... 27
4.5 Grammar................................................................................................................... 28
Chapter 5 General discussion....................................................................................... 29
Chapter 6 Conclusion and further studies................................................................... 32

References..................................................................................................................... 35

Appendix A Writing Format........................................................................................... 40
Appendix B Direction for Writers.................................................................................. 41
Appendix C Writing Samples......................................................................................... 42
Appendix D Direction for Raters.................................................................................... 45
Appendix E Rubric........................................................................................................... 46
List of Tables

Table 1 The means of scores................................................................. 24
Table 2 Results of Two-way ANOVA of raters' proficiency (independent) x writers' proficiency (repeated) (Total)......................................................... 25
Table 3 Results of Two-way ANOVA of raters' proficiency (independent) x writers' proficiency (repeated) (Content)......................................................... 26
Table 4 Results of Two-way ANOVA of raters' proficiency (independent) x writers' proficiency (repeated) (Organization)......................................................... 27
Table 5 Results of Two-way ANOVA of raters' proficiency (independent) x writers' proficiency (repeated) (Vocabulary)......................................................... 27
Table 6 Results of Two-way ANOVA of raters' proficiency (independent) x writers' proficiency (repeated) (Grammar)......................................................... 28
List of Figures

Figure 1. The rating scale of "content"................................................................. 18
Figure 2. The rating scale of "organization"......................................................... 19
Figure 3. The rating scale of "vocabulary"............................................................. 21
Figure 4. The rating scale of "grammar"............................................................... 22
Chapter 1
Introduction

In the past two decades, the Ministry of Education, Culture, Sports, Science and Technology (MEXT) has begun to use three key phrases: "redefinition of solid academic prowess," "introduction of criterion-referenced assessment system," and "further advancement of liberal, flexible and comfortable school life (Yutori)."

Yutori education was initiated by the Course of Study in 1998, when 30% of the syllabus content was cut and a complete five-day school week was enforced as a means to counteract the culture of cramming education and the advancement of liberal, flexible, and comfortable school life. However, Yutori invited criticism as society regards children's academic ability to have declined, according to the ranking of the Programme for International Student Assessment (PISA) by the Organization for Economic Cooperation and Development (OECD). Refutations to this criticism of Yutori, however, have said that the decline in ranking is due to the increased number of countries attending PISA. Accordingly, Yutori was abolished without achieving its original purpose when the number of school hours was increased by 10% in the 2008 Course of Study.

Since the end of the Yutori era, under globalization, the necessity of the four skills in English has been increasing and educational reorganization is being accelerated. The government aims to develop global human resources and nurture of the four English skills is now included in the Course of Study. In 2011, "foreign language activities" were initiated in elementary school classrooms, while students now begin learning English through English in high school classrooms. However, entrance examinations test students' reading, listening, and limited writing skills; therefore, English teachers tend to teach only these skills in their English classes. To solve this problem, the university entrance
examination is going to be reorganized. Specifically, the Central Council for Education plans to introduce "the pleiotropic comprehensive evaluation system" by using an essay writing test and/or interview test instead of the active "points system", which uses an optical answer sheet, for the university entrance examination by 2020. Simultaneously, the government plans to introduce an external test, such as the EIKEN Tests that are "Japan's most widely recognized English language assessment" (Eiken Foundation of Japan, n.d.), TOEFL iBT, TOEIC, TOEIC (S/W), IELTS, GTEC CRT, and TEAP, for the university entrance examination in order to evaluate students' four English skills (reading, writing, listening, and speaking).

Regarding English education in Japan, performance assessment for "solid academic prowess," such as cultivating introspection, the desire to learn and think, independent decision-making and action, as well as the talent and ability for problem-solving, has become famous in elementary and secondary education since being referred to in a cumulative guidance record in the revised Course of Study in 2008. In addition, on the assumption that multifaceted assessment criteria are necessary to evaluate solid academic prowess precisely, the performance assessment and rubric are mentioned in the report as: "The reorganization for unification of the education in high schools and universities and university entrance examinations in order to actualize the connection between high schools and universities, which is fitting for the new era" (Central Council for Education, 2014; translated by the author). The Central Council for Education has also mentioned a plan to introduce assessments with multifaceted criteria, such as performance assessment and writing tests, to university entrance examinations in the future.

To sum up, in English education in Japan, teachers mainly teach reading and listening due to university entrance examinations that adopt the points system using an optical answer sheet. However, the government aims for students to be able to learn
English in terms of the four skills equally and thus it plans to introduce external tests to university entrance examinations that test students’ solid academic prowess by measuring the four skills (reading, writing, listening, and speaking). In other words, it is conceivable that speaking and writing abilities will become more important in English classrooms due to this reorganization.

Writing ability, which has achieved more importance in the reorganization of English education in Japan, involves a complex process in which learners may encounter some difficulties. Simultaneously, the complex process of writing causes difficulties in evaluation. In order to evaluate writing ability with high reliability, many researchers have attempted to minimize and standardize the raters’ effect. However, “the rating process is complex and there are numerous factors which affect raters’ judgment” (Weigle, 1994); moreover, these factors have not been clarified. Therefore, it is necessary to clarify the factors to increase the reliability of writing evaluation and possible factors include: gender, age, cultural background, and language background.

In particular, if the writer and rater are English as a foreign language (EFL) learners, the writing and rating process becomes more difficult due to their English proficiency compared with native English speakers (NES’s); however, few studies have been conducted with regard to raters’ effects on evaluation that target EFL raters.

In the future, teachers may have to simultaneously teach and rate their students’ writing. However, as mentioned, it is difficult for raters to evaluate writing; moreover, if the raters are EFL learners, it becomes more difficult due to their language background. In this study, Japanese EFL raters’ proficiency effects on Japanese EFL learners’ evaluations of EFL essay writing are investigated in order to explicate the specific factors that affect EFL raters’ assessments. If these factors are elucidated, it may become possible for Japanese English teachers to increase their teaching of writing. Moreover, when
students learn writing in English classes, they will be able to evaluate each other’s essay writing.
Chapter 2

Literature Review

2.1 Writing

Of the four English skills (listening, speaking, reading, writing), the process of writing is a particularly confused and recursive intellectual process (Silva, 1992). Writers must produce sentences by correlating the processes of generating ideas, structuring, drafting, focusing, evaluating, and reviewing. Therefore, all NES’s, English as a second language (ESL) learners, and EFL learners must ultimately learn writing skills (Hamp-Lyons & Heasley, 1987).

Raimes (1983) explained six ways of teaching writing: the controlled-to-free, freewriting, paragraph-pattern, grammar-syntax-organization, communicative, and process. As teachers use these techniques, they must be conscious of the balance between the quantity and quality of writing depending on the traits of each approach. For instance, in the freewriting approach, the teacher makes the students focus on the quantity and fluency of writing and the writers try to write a lot. This approach is necessary for beginner learners of writing because it can reduce the Affective filter (Krashen, 1982), which comprises the psychic hindrance to learning. Moreover, teaching writing would not succeed if writers were to spend a great deal of time writing texts accurately. On the other hand, the quality of writing and grammatical accuracy cannot be ignored because grammatical competence is included amongst the lower level acknowledgements of communicative competence (Canale, 1983). Therefore, in teaching writing, teachers should first allow students to write a lot, and then gradually let them focus on the accuracy of the texts (Mochizuki, 2014).

Above all, the process approach has been addressed in some recent studies. In this
approach, teachers help students to write compositions using all writing processes such as generating ideas, structuring, drafting, focusing, evaluating, and reviewing. In particular, ESL/EFL writing produced using the process approach consists of prewriting, drafting, evaluating, and revising (Zhang, 1995) and entails various difficulties that cannot be recognized in the process approach of NES writing. Specifically, the processes of composing, transcribing, planning, and reviewing are more difficult for EFL learners than NES's; therefore, the texts produced by EFL learners are short and their paragraphs lack consistency. Moreover, they rarely use metaphors and cannot express subtle nuance due to their lack of vocabulary (Chelala, 1981; Cook, 1988; Dejesus, 1983, 1984; Gaskill, 1986; Hall, 1987, 1990; Indrasuta, 1987, 1988; Jones & Tetroe, 1987; Lin, 1989; Norment, 1986; Schiller, 1989; Skibniewski, 1988; Skibniewski & Skibniewska, 1986). As mentioned above, the writing process is complex and time-consuming, and, in particular, EFL learners encounter more difficulties in the writing process due to their low level English abilities. On top of that, not only EFL learners but also EFL teachers who oversee the teaching of writing may encounter difficulties because they need to acquire all of the processes of EFL writing, the techniques of teaching writing, and English itself.

To conclude with regard to teaching writing, learners are usually examined on what they have learned using tests. As raters evaluate writing tests, an appropriate rating scale should be selected and the criteria established depending on the purpose of the evaluation (Bacha, 2001). However, evaluation tends to be complicated and error-prone because writing raters use their own rating scales for evaluation (Schaefer, 2008).

2.2 Measurement, evaluation, and test

In this thesis, the words "measurement," "evaluation," and "test" are often used and should be defined clearly because they are sometimes confused. Measurement means
"systematic deeds for explanation of size or strength of a trait of the subject by numbers" (Shizuka, 2002) and it can be applied in various contexts, including the educational, architectural, and medical fields. On the other hand, evaluation entails providing a value judgment to elicited numbers by measuring, and is distinguished from the word "measuring." With reference to measurement, tests are amongst the most popular measuring techniques in an educational context and include written and practical tests. Testing is designed to elicit samples of participants' behavior in order to quantify their abilities, aptitudes, and motivations (Bachman, 1990), and the three conditions for developing a "good test" are reliability, validity, and practicality.

Test reliability is the degree to which a test reflects the real ability of participants. If the same participant takes the same test many times and the scores are always the same, the test has high reliability. Moreover, if different raters evaluate the same test and the scores are the same, the test also has high reliability. In this case, the reliability is called inter-rater reliability. If the same rater evaluates the same test a number of times and the scores are always the same, the test has high inner-rater reliability.

Test validity is the degree to which a test appropriately measures the ability it seeks to measure. There are three types of validity: content validity, construct validity, and face validity.

Test practicality is the degree to which a test is easy to implement. If a test is impossible to implement, it is meaningless despite having high reliability and validity. Tests are also classified into various patterns according to their different purposes or types of question (Brown, 1996; Ito, 2011; Mochizuki, 2013). For example, an indirect test is one whose participants are asked for lower-level knowledge about a subject. In contrast to indirect tests, in direct tests, participants are asked to perform "real" tasks (Schooneen, Vergeer, & Eiting, 1997) and multiple assessment criteria are applied to evaluate that "real"
performance, which has multiple aspects. There are also norm-referenced tests and criterion-referenced tests. Norm-referenced tests are relative tests that measure total language ability and find a participant’s level in comparison to all participants. They include proficiency tests and placement tests. On the one hand, in criterion-reference tests, participants’ scores are compared not with each other but with their learning objectives, and these tests include achievement tests, diagnostic tests, and aptitude tests. Moreover, tests are classified into objective tests and subjective tests by their method of evaluation. Multiple-choice tests and true-false tests are types of objective test because the score is always the same regardless of who assesses them. On the other hand, speaking tests are subjective tests because their score is variable depending on the rater’s subjectivity or feeling.

As mentioned above, performance assessment, which is classified into direct tests and subjective tests, tends to be complicated and error-prone due to its various traits.

**Direct tests**

Direct tests can measure skills or abilities directly. Their validity tends to be higher than that of indirect tests and objective tests and they are “the most valid way to gather information about the general level of the students’ writing proficiency” (Schoonen et al., 1997).

However, according to Schoonen et al. (1997), the scores for different essays often have low consistency and raters are not always consistent in their assessments, nor do they agree with other raters of writing assessments. In other words, it is difficult to evaluate with high inner-rater reliability and inter-rater reliability.
Subjective tests

Regarding the evaluation of subjective tests such as a writing test, there are two types of evaluation: holistic evaluation and analytic evaluation. Holistic evaluation uses just one scale to give a comprehensive score to participants (Brown, 1996). According to Cooper (1977), in holistic evaluation, the raters evaluate using a holistic scoring guide after practicing the procedure with other raters and this involves the process of placing, scoring, and grading the written pieces. This is also the most valid way of placing students according to their writing ability because the raters can evaluate students’ writing “quickly and impressionistically” (Cooper, 1977). However, Cooper (1977) also mentioned that “a group of raters will assign widely varying grades to the same essay” and the effect of this is incontrovertible. In other words, the inter-rater reliability tends to be low.

In contrast to holistic evaluation, analytic evaluation uses various scales to evaluate various aspects of participants’ skills. While it takes much more time to evaluate, the reliability and consistency of evaluation tend to be higher than for holistic evaluation. According to Popham (2002), performance assessment has three conditions: (1) it should have multiple rating scales; (2) the criteria should be decided in advance; and (3) it should include the raters’ judgments. Regarding the second condition, the table that shows the criteria is called a rubric.

2.3 Raters’ effect on evaluation

Performance tests have high validity, but their reliability tends to be lower. Hence, attempts are made to reduce the variability of raters’ behaviors. Lumley and McNamara (1995) noted that “subjectivity ... may be reduced by the adoption of scoring rules when essay examinations are to be graded,” and that rater training may reduce the variability
and randomness of raters’ scores.

According to Lumley and McNamara (1995), rater training usually involves three sessions: (1) raters are introduced to the assessment criteria; (2) raters rate a series of performances and discuss it with others; and (3) raters are given examples of a range of abilities and characteristic issues arising in the assessment. Accordingly, rater training can reduce the variability of raters’ severity and extreme differences in harshness or leniency (McIntyre, 1993). In addition, it can reduce random errors in raters’ assessments and make raters more self-consistent, which is the most crucial quality in a rater (Wiseman, 1949), but does not dramatically alter severity (Lumley & McNamara, 1995; Weigle, 1998). Moreover, trained raters are more reliable than untrained raters (Kraemer, 1992; Weigle, 1994, 1998), and, according to Stalnaker (1934), “rater reliability could be improved from a range of .30 to .75 before training to a range of .73 to .98 after training.”

With regard to recent research, Matsuo (2009) compared self-assessment and peer-assessment with teachers’ assessment, on a target group of 91 university students and four teachers. The students received guidance on essay writing in seven lessons, including on aspects such as format, mechanics, construction, and content, and then wrote a 300-word essay on a given topic in the eighth lesson. In the ninth lesson, the students practiced evaluation using the criteria of the Multifaceted Rasch Measurement (MFRM) and evaluated six essays, including their own. As a result of data analysis, (a) the evaluation was more severe than expected in the self-assessment, (b) the evaluation was lenient in the peer-assessment, (c) the peer-assessment had fewer bias effects compared with other assessments, and (d) the evaluation was more severe on the criteria of grammar and more lenient on the criteria of spelling. In addition, the peer-assessment had internal consistency and its pattern of evaluation did not depend on the raters’ writing skills. Moreover, the author disclosed that the bias hardly affected the evaluation in peer-
assessment and the quality of the peer-assessment could be improved by using MFRM. In the conclusion of this study, the assessment of the students showed an internal consistency that was as reliable as the teachers' assessment.

Schaefer (2008) investigated rater tendency amongst a target of 40 assistant language teachers (ALTs) working at junior high or high schools in Tokyo. The ALTs were untrained as raters. In this study, rater-category and rater-writer relations were studied using MFRM; it was revealed that the raters evaluated individually and it was difficult for untrained raters to assure the reliability and validity of writing assessments. Note that the results did not depend on the texts as the raters evaluated the same essays. The author recommended using the MFRM as a guideline for rater training, and explained that doing so could improve the reliability and validity of the evaluation.

Johnson (2009) studied the effect of raters' backgrounds on evaluation, using 19 writing test raters of Michigan English Language Assessment Battery (MELAB) on the premise that raters' language backgrounds should be considered in writing evaluation because there are various raters of each language background in the worldwide writing test. The raters comprised 15 native English speakers and 4 non-native English speakers whose English proficiency was native-like, and they had been given the rater training by MELAB. As a result, the author showed that language background does not affect evaluation and raters can evaluate adequately, regardless of their language background.

According to Ling (2010), a group of NES raters who had taken teacher training evaluated EFL essays more reliably than a group of non-native English speakers without the training. In addition, Schoonen et al. (1997) examined the effect of raters' experience of evaluation by comparing lay raters and expert raters. The raters were trained by one of the researchers and a sample of essays was evaluated by them, and the scores discussed. Next, evaluation practice continued until the raters felt confident about the criteria and
rating. As a result, the expert raters’ evaluation was found to be more reliable than the lay raters’ evaluation.

Finally, Weigle (1994) compared two groups of raters: new raters and old raters. The new raters were raters who had never rated the ESL placement examination (ESLPE) given by the University of California, Los Angeles, while the old raters were experienced ESLPE raters. Accordingly, the rater training was effective for new raters because “it helped the raters to understand and apply the intended rating criteria” and “it modified the raters’ expectations in terms of the characteristics of the writers and the demands of the writing tasks.”

All in all, a number of studies have shown the availability of rater training. Moreover, they have ascertained that such training is necessary for evaluation and the reliability of evaluation seems to depend on rater training regardless of language background or teacher experience. However, rater training cannot reduce the significant and substantial differences that persist between raters (McIntyre, 1993) because raters often have unique standards, which it is difficult for them to alter (Lunz & Stahl, 1990). For instance, Saito (2008) investigated the effect of rater training on the evaluation of oral presentations by comparing two groups of raters. The participants comprised 74 Japanese university freshmen who were majoring in economics and were divided into two groups: treatment and control. In study 1, both groups first received instruction on skill aspects and evaluated each other in the oral presentation, but only raters in the treatment group were trained for 40 minutes as raters before the evaluation. In study 2, the 40-minute rater training was replaced with a long training: the total training time amounted to approximately 200 minutes. However, there were no differences between the scores of the two groups in either of the studies. In addition, Lumley and McNamara (1995) argued that the results of training may not endure for long after a training session. Moreover,
research on rater training is inconclusive with regard to what background variables could relate to rater training (Saito, 2008). In Johnson’s (2009) study, the reason for the lack of language background effects was probably because the participating raters’ English proficiency was relatively high.

In sum, rater training undoubtedly has a positive effect on performance assessment; however, it is still not the perfect way to reduce raters’ negative effects. “The rating process is complex and there are numerous factors that affect raters’ judgment” (Weigle, 1994); moreover, these factors have not been clarified. One of the factors is raters’ language background, which differs greatly between NES raters and EFL raters. This factor has rarely been examined because, in the studies above, almost all participants were NES’s or had high English skill levels. However, it has become more necessary to investigate the differences between NES raters and EFL raters because it is obvious that both groups are required to evaluate writing abilities in the era of globalization.

2.4 Role of writing tests in Japan

If the reorganization of university entrance examinations in Japan is realized and the pleiotropic comprehensive evaluation system is adopted, writing and speaking skills will become more important for English education. At the same time, teachers will need to teach writing and speaking to their students, and evaluate them appropriately using performance assessment.

However, teachers are too busy teaching to manage testing and a great deal of data for investigating, for example, examination questions has been treated as secret.

It has been said that teaching and testing are inseparable; however, teaching has been mainstream in English educational contexts and
testing has hardly ever been considered (Tanaka, 1998).

Moreover, there is a major difference between NES raters and EFL raters, i.e., language background. Most raters are Japanese and their mother tongue is Japanese. As mentioned above, both EFL learners and EFL raters are severely burdened by EFL writing and its evaluation. Therefore, it is necessary to clarify the effects of Japanese EFL raters’ English proficiency on their performance assessment in order to better understand the differences between NES raters and EFL raters. If Japanese EFL raters’ own criteria differ due to their English proficiency, the proficiency effect should be standardized in order to increase the reliability of evaluation. What is more, if Japanese EFL raters’ proficiency affects their evaluation of EFL writing and the proficiency effect can be minimized or standardized, it will lead to an increase in teaching because teaching and evaluating are inseparable.

In addition, each classroom contains students of varying English proficiency. If teachers give students the opportunity to write essays and evaluate each other, it is necessary to consider the proficiency effect of both raters and writers.

2.5 Research Questions

In conclusion, the importance of teaching writing has increased in classrooms in Japan due to the reorganization of English education in Japan by the reformation of the university entrance examination. However, the writing process is complex and both learners and teachers encounter difficulties during teaching, learning, and evaluating. Moreover, when evaluating EFL essays, most raters are EFL learners and have specific factors that affect their raters’ judgment that are not found in evaluations by NES’s. One of these factors is their language background, the effect of which has not yet been
explicated in Japan.

In order to explicate the effect of a Japanese language background, this study investigated raters' proficiency effect for the evaluation of EFL writing. The research questions for this study were as follows:

1. Does EFL raters' proficiency affect their evaluation of EFL essay writing?
2. When EFL raters evaluate EFL essays, do they need to consider the writers' proficiency?
Chapter 3
Method

3.1 Participants

43 Japanese EFL undergraduate and graduate school students participated in this study: three writers and 40 raters. They took the Quick Placement Test of the Oxford University Press, and were divided by English proficiency level according to the Common European Framework of Reference for Languages (CEFR). All participants had no teaching experience apart from teacher practice in university. There were two reasons why Japanese EFL university students without teaching experience were assembled: (1) to investigate the effect of raters’ proficiency, and (2) to eliminate the bias effect of teaching experience.

The writer participants comprised three sophomore students (two females and one male), all of whom were 18 years old. Their proficiency levels were A2, B1, and B2 levels, respectively.

The rater participants comprised 40 undergraduate and graduate school students in Japan (13 males and 27 females), ranging in age from 18 to 24 (mean = 21.55). The participants were divided into two groups according to their placement test scores: 26 participants were lower level and 14 were higher level. 26 participants were A2 level, 12 were B1, and 2 were B2.

3.2 Rating scale

The rating scale was composed by arranging the five scales (content, organization, vocabulary, language use, and mechanics) by Jacobs et al. (1981) and the English free writing criteria by Toyama Minami high school in order to make it appropriate to the case
of English education in Japan. Toyama Minami high school was designated the Super English Language High School (SELHi) from 2003 to 2007; it provides an international course and attaches importance to English education. The original criteria of Jacobs et al. (1981) comprised five rating scales with different weightings, but the weightings of all rating scales of criteria were equalized in this study by using a six-point Likert scale. When removing these weightings, however, the rating scale “mechanics” was excepted because the weighting of this scale was regarded as too light compared with the others. In addition, the author bound sentences of criteria to each score and removed the scores “2” and “5” in order to partly control the raters’ evaluations with a minimum effect of the criteria.

3.2.1 Content

Figure 1 shows the content rubric that the raters used to evaluate in this study. As mentioned above, scores “2” and “5” have no criteria in order that raters’ own criteria may be used and the effect of the author’s criteria is minimized. “Content” dealt with the information given in the essay in three respects: (1) persuasiveness, (2) appropriateness, and (3) incontestability.

To evaluate the persuasiveness of the essay writing, the abundance and appropriateness of the supporting sentences used for each opinion were assessed. If there were sufficient supporting sentences to back up the writer’s opinion or argument, the essay was regarded as well-grounded with sufficient persuasiveness.

More importantly, the appropriateness of the essay writing was evaluated by assessing the consistency of the writer’s composition. If the given topic and essay context corresponded with a high inner-writer consistency, the essay was regarded as appropriate.

In particular, the evaluation gave priority to the incontestability of the writer. First,
the raters assessed the essay on their ability to understand the writer’s argument. If the essay fit the given topic and had an appropriate level of persuasiveness, it could not receive a high score without a clear argument.

To sum up, essays that achieved a high score on the “content” scale needed to fulfill the following three conditions: (1) the writer’s argument was obvious; (2) the given topic and essay context corresponded; and (3) each opinion was supported by appropriate supporting sentences.

<table>
<thead>
<tr>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

Figure 1. The rating scale of “content”

3.2.2 Organization

Figure 2 shows the criteria for evaluation of “organization”. Organization dealt with the smoothness of the construction of the overall essay in four respects: (1) flow of the texts, (2) logic of the composition, (3) construction using an introduction-body-conclusion form, and (4) linking words.

In evaluating the smoothness of an essay, the flow of the text and logicality of the composition are usually regarded as the foremost criteria. However, these aspects tend to be abstract for raters, which causes difficulty in evaluation. In addition, the flow of the
text and logic of the composition could be regarded as criteria not only of "organization" but also of "content." This overlap caused complexity of evaluation and these should be differentiated clearly.

Therefore, the criteria gave priority to the use of linking words. The appropriateness of linking words was regarded as the most important criterion for the following two reasons: (1) linking words have a higher concreteness than other aspects of evaluation of "organization," and (2) The "organization" scale should be clearly differentiated from the "content" scale.

To sum up, essays that received a high score on the "organization" scale needed to fulfill the following four conditions: (1) linking words are used appropriately; (2) the essay has an introduction-body-conclusion form; (3) the overall composition is logical; and (4) the sentences flow naturally.

<table>
<thead>
<tr>
<th>Organization</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>The linking words are used appropriately and the essay has the introduction-body-conclusion form. The whole composition is logical and the sentences flow naturally.</td>
</tr>
<tr>
<td>5</td>
<td>Some linking words are used and the essay has clear construction, but the sentences do not flow perfectly.</td>
</tr>
<tr>
<td>4</td>
<td>The linking words are used limitedly. The essay does not have clear construction and the sentences do not flow naturally.</td>
</tr>
<tr>
<td>3</td>
<td>Few linking words are used. The essay does not have construction.</td>
</tr>
</tbody>
</table>

Figure 2. The rating scale of "organization"

3.2.3 Vocabulary

Figure 3 shows the criteria for evaluation of "vocabulary". Vocabulary dealt with lexical richness and the comprehensibility of the essay in three respects: (1) diversity of
vocabulary, (2) appropriateness of vocabulary, and (3) consideration for readers.

According to Read (2000), there are three ways to measure the lexical richness of a composition: The Type-token ration (TTR), which shows the ratio of type and token; Lexical Density (LD), which shows the ratio of content words to function words; and Lexical Variation (LV), which measures the TTR of content words only. In addition, the Lexical Frequency Profile (LFP) can analyze the level of written words (Sugimori, 2009). However, in many cases, raters evaluate essay vocabulary based on their own judgment due to time considerations (Sugimori, 2009); thus, the measures were not used in this study and the raters assessed the diversity of vocabulary and appropriateness of vocabulary based on their own judgment.

On the other hand, the criteria did not give priority to the appropriateness of vocabulary, but rather to the diversity of vocabulary. As mentioned above, learners should shift their focus from quantity to quality when learning writing. Essays whose writers use a range of vocabulary intuitively with some errors should be evaluated more highly than essays whose writers use a narrow range of vocabulary in order to avoid errors. However, needless to say, appropriateness of vocabulary is as necessary a criterion as diversity of vocabulary.

Additionally, difficult words or proper nouns may cause difficulty in understanding an essay. Essays that are conscious of readers by paraphrasing or explaining difficult words and proper nouns should achieve a higher score.

To sum up, essays that can achieve a high score on the "vocabulary" scale need to fulfill the following three conditions: (1) a variety of words and expressions is used; (2) their use is appropriate; and (3) the essay is conscious of readers in terms of its lexical aspects.
Vocabulary

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Various words and expression are used and the use is appropriate. The essay conscious the readers on lexical aspect.</td>
</tr>
<tr>
<td>5</td>
<td>Various words and expression are used, but the use is partly not appropriate.</td>
</tr>
<tr>
<td>4</td>
<td>The words and expression are used limitedly and there are some mistakes.</td>
</tr>
<tr>
<td>3</td>
<td>There are few information about words and expressions and the essay is incomprehensible due to too many mistakes.</td>
</tr>
</tbody>
</table>

Figure 3. The rating scale of “vocabulary”

3.2.4 Grammar

Figure 4 shows the criteria for the evaluation of “grammar”. The “grammar” scale dealt with the comprehensibility of the essay in two respects: (1) diversity of expression and (2) grammatical accuracy of sentences.

These criteria did not give priority to sentence accuracy because if this was regarded as the most important criterion, writers would be liable to use only simple grammatical constructions. To avoid this risk, the criteria prioritized the diversity of sentences and expressions. Of course, the grammatical accuracy and appropriateness of the sentences were necessary criteria for “grammar”.

To sum up, essays that can achieve high scores on the “grammar” scale need to fulfill the following three conditions: (1) various sentence structures are used appropriately; and (2) there are no grammatical (in terms of consistency of subject and verb, tense, numbers, pronouns, articles, and prepositions) errors.
Grammar

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Various sentence structures are used appropriately and there are no grammatical mistakes. (consistency of subject and verb, tense, numbers, pronoun, article, and preposition)</td>
</tr>
<tr>
<td>5</td>
<td>Various sentence structures are used appropriately, but there are some grammatical mistakes.</td>
</tr>
<tr>
<td>4</td>
<td>The sentence structures are used limitedly and there are grammatical mistakes.</td>
</tr>
<tr>
<td>3</td>
<td>The essay is incomprehensible due to too many grammatical mistakes.</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 4.** The rating scale of “grammar”

### 3.3 Procedure

In this study, three writers wrote essays and the raters rated them using the rating scale. All participants were measured in terms of their English proficiency and the proficiency effect of the evaluation was investigated.

**Writers**

The writers were selected from amongst the students who had already taken the Quick Placement Test. At the beginning, in order to partly control the writers’ essay writing and secure the validity of the rating scales, the writers were asked to check the rating scale before writing. They then had 30 minutes in which to write a 150-word essay on the following topic: “A foreign visitor has only one day to spend in your country. Where should this visitor go on that day? Why?” quoted from the TOEFL essay questions. The reason for selecting this topic was that it seemed easy for (1) writers to produce an essay with concrete grounds, and (2) the writers’ own opinions needed to be reflected in the essay because the topic seemed familiar to writers.

After the writers had written their essays, they were typed up by the author to avoid
the handwriting affect in the evaluation (Chase, 1968). The paragraphing used was the same as in the original.

**Raters**

The raters rated the three essays in counterbalanced order. Before the evaluation, the raters were asked to check the rating scale once, and received general instruction and explanation of the criteria, which included the following two points: (1) The raters had to assess the essays one by one and were prohibited from going back to evaluate the previous essays again; (2) the sentences in the criteria were constructed in order of priority of evaluation.

The raters were asked to evaluate essays using the rating scale within 30 minutes without rater training. All raters took plenty of time to assess the writings. After evaluating, raters took the Quick Placement Test for 30 minutes.
Chapter 4

Results and discussion

The two-way analysis of variance (two-way ANOVA) assessed differences between the raters' proficiency (independent-measures: lower proficiency raters and higher proficiency raters) and writers' proficiency (repeated-measures: A2, B1, B2) on each evaluating scale (content, organization, vocabulary, grammar, total) based on the scores of evaluations.

Table 1 The means of scores

<table>
<thead>
<tr>
<th>Writers</th>
<th>Content</th>
<th>Organization</th>
<th>Vocabulary</th>
<th>Grammar</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>All raters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>4.08</td>
<td>3.03</td>
<td>3.08</td>
<td>3.40</td>
<td>13.58</td>
</tr>
<tr>
<td>B1</td>
<td>4.85</td>
<td>4.90</td>
<td>4.55</td>
<td>4.45</td>
<td>18.75</td>
</tr>
<tr>
<td>B2</td>
<td>4.55</td>
<td>4.85</td>
<td>4.53</td>
<td>4.35</td>
<td>18.28</td>
</tr>
<tr>
<td>Lower Proficiency Raters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>4.15</td>
<td>3.04</td>
<td>3.15</td>
<td>3.54</td>
<td>14.00</td>
</tr>
<tr>
<td>B1</td>
<td>5.12</td>
<td>5.27</td>
<td>4.58</td>
<td>4.58</td>
<td>19.54</td>
</tr>
<tr>
<td>B2</td>
<td>4.50</td>
<td>4.73</td>
<td>4.69</td>
<td>4.50</td>
<td>18.42</td>
</tr>
<tr>
<td>Higher Proficiency Raters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>3.93</td>
<td>3.00</td>
<td>2.93</td>
<td>3.14</td>
<td>12.79</td>
</tr>
<tr>
<td>B1</td>
<td>4.36</td>
<td>4.21</td>
<td>4.50</td>
<td>4.21</td>
<td>17.29</td>
</tr>
<tr>
<td>B2</td>
<td>4.64</td>
<td>5.07</td>
<td>4.21</td>
<td>4.07</td>
<td>18.00</td>
</tr>
</tbody>
</table>

4.1 Total

Table 1 compares the means of the scores given to the three essays by the raters. The mean scores of the two groups indicate that the three essays ranged from a low of 2.93 to a high of 5.27 on the 6-point scale. Table 2 also shows the result of the two-way analysis of variance (two-way ANOVA) between the raters' proficiency (independent-
measures: higher proficiency group and lower proficiency group) and writers’ proficiency (repeated-measures: A2, B1, B2) on the mean of total scores. According to this result, both groups of raters agreed that the essay written by the A2 writer was the lowest (writers’ proficiency: \( F (1.82, 69.17) = 1.28, p < .01, \eta^2 = .56 \)). However, as regards the essays written by the B1 and B2 writers, there were no differences in the scores, although the essay written by the B1 writer received the highest score from the lower proficiency raters while the essay written by the B2 writer received the highest score from the higher proficiency raters. This is because the writing process comprises not only linguistic competences but also various other components such as correlating the processes of generating ideas, structuring, drafting, focusing, evaluating, or reviewing.

On the other hand, each scale (content, organization, vocabulary, grammar, total) was also examined by two-way ANOVA between raters’ proficiency (independent-measures: higher proficiency group and lower proficiency group) and writers’ proficiency (repeated-measures: A2, B1, B2).

**Table 2** Results of Two-way ANOVA of raters’ proficiency (independent) x writers’ proficiency (repeated) (Total)

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>( F )</th>
<th>( p )</th>
<th>( \eta^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>raters' proficiency error</td>
<td>45.903</td>
<td>1</td>
<td>45.903</td>
<td>2.726</td>
<td>0.107</td>
<td>0.067</td>
</tr>
<tr>
<td></td>
<td>639.963</td>
<td>38</td>
<td>16.841</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>writers' proficiency</td>
<td>587.891</td>
<td>1.82</td>
<td>322.995</td>
<td>48.985</td>
<td>0.00</td>
<td>0.563</td>
</tr>
<tr>
<td>writers' proficiency x raters' proficiency error (writers proficiency)</td>
<td>15.325</td>
<td>1.82</td>
<td>8.420</td>
<td>1.277</td>
<td>0.283</td>
<td>0.033</td>
</tr>
<tr>
<td>error (writers proficiency)</td>
<td>456.059</td>
<td>69.165</td>
<td>6.594</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 4.2 Content

With regard to the rating scale of content, no significant differences were found in scores between the two groups of raters (Table 3). The raters evaluated the essays similarly, regardless of the writers’ proficiency or raters’ proficiency. In other words,
English proficiency had little effect on evaluation for "content" in this study and the raters seem to have been able to understand the content of all three essays and assessed them without linguistic problems. Writers do not always consider the content of essays in English but sometimes do so in their own language. For this reason, it is unlikely that English proficiency would affect the evaluation of content.

In addition, lower proficiency raters gave the highest score to the B1 writer, while higher proficiency raters gave the highest score to the B2 writer.

**Table 3** Results of Two-way ANOVA of raters’ proficiency (independent) x writers’ proficiency (repeated) (Content)

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
<th>( \eta^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Between Subjects</td>
</tr>
<tr>
<td>raters’ proficiency</td>
<td>2.144</td>
<td>1</td>
<td>2.144</td>
<td>1.263</td>
<td>0.268</td>
<td>0.032</td>
</tr>
<tr>
<td>error</td>
<td>64.515</td>
<td>38</td>
<td>1.698</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Within Subjects</td>
</tr>
<tr>
<td>writers’ proficiency</td>
<td>9.602</td>
<td>1.873</td>
<td>5.127</td>
<td>4.082</td>
<td>0.023</td>
<td>0.097</td>
</tr>
<tr>
<td>writers’ proficiency x raters’ proficiency</td>
<td>3.736</td>
<td>1.873</td>
<td>1.995</td>
<td>1.588</td>
<td>0.213</td>
<td>0.040</td>
</tr>
<tr>
<td>error (writers proficiency)</td>
<td>89.381</td>
<td>71.167</td>
<td>1.256</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.3 Organization

As regards the rating scale of "organization" (see Table 4), the main effect between the raters’ English proficiency and writers’ English proficiency (A2-B1, A2-B2) was significant (writers’ proficiency: \( F (1.86, 68.96) = 43.95, p < .01, \eta^2 = .54 \)). In other words, raters assessed the essay written by the A2 writer lower than the essays written by the B1 and B2 writers. In addition, the raters’ proficiency x writers’ proficiency interaction was significant (writers’ proficiency x raters’ proficiency: \( F (1.82, 68.96) = 5.26, p < .01, \eta^2 = .12 \)) and the simple main effect of each factor was tested statistically in order to understand the interaction. As a result, lower proficiency raters gave a higher score to the essay written by the B1 writer than the higher proficiency raters did (\( p = .007 \)). In short, English proficiency affected the evaluation of "organization" and this result
should be investigated when evaluating the organization of EFL essay writing.

### Table 4 Results of Two-way ANOVA of raters’ proficiency (independent) x writers’ proficiency (repeated) (Organization)

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
<th>η²</th>
</tr>
</thead>
<tbody>
<tr>
<td>raters' proficiency</td>
<td>1.719</td>
<td>1</td>
<td>1.719</td>
<td>1.168</td>
<td>0.287</td>
<td>0.030</td>
</tr>
<tr>
<td>error</td>
<td>55.940</td>
<td>38</td>
<td>1.472</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>writers' proficiency</td>
<td>79.278</td>
<td>1.815</td>
<td>43.689</td>
<td>43.954</td>
<td>0.000</td>
<td>0.536</td>
</tr>
<tr>
<td>writers' proficiency × raters' proficiency</td>
<td>9.478</td>
<td>1.815</td>
<td>5.223</td>
<td>5.255</td>
<td>0.009</td>
<td>0.121</td>
</tr>
<tr>
<td>error (writers proficiency)</td>
<td>68.538</td>
<td>68.955</td>
<td>0.994</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### 4.4 Vocabulary

On the scale of “vocabulary” (see Table 5), the main effect between the raters’ English proficiency and writers’ English proficiency (A2-B1, A2-B2) was significant (writers’ proficiency: $F(1.74, 66.04) = 44.94, p < .01, η² = .54$), but there was no significant interaction ($p = 0.99$). The participating raters agreed that the vocabulary level of the A2 writer’s essay was the lowest. However, the assessment did not depend on the raters’ English proficiency and English proficiency seems to have had no effect on evaluation of vocabulary in this study.

### Table 5 Results of Two-way ANOVA of raters’ proficiency (independent)x writers’ proficiency (repeated) (Vocabulary)

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
<th>η²</th>
</tr>
</thead>
<tbody>
<tr>
<td>raters' proficiency</td>
<td>1.847</td>
<td>1</td>
<td>1.847</td>
<td>1.389</td>
<td>0.246</td>
<td>0.035</td>
</tr>
<tr>
<td>error</td>
<td>50.520</td>
<td>38</td>
<td>1.329</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>writers' proficiency</td>
<td>51.482</td>
<td>1.738</td>
<td>29.622</td>
<td>44.937</td>
<td>0.000</td>
<td>0.542</td>
</tr>
<tr>
<td>writers' proficiency × raters' proficiency</td>
<td>0.749</td>
<td>1.738</td>
<td>0.431</td>
<td>0.653</td>
<td>0.503</td>
<td>0.017</td>
</tr>
<tr>
<td>error (writers proficiency)</td>
<td>43.555</td>
<td>66.044</td>
<td>0.659</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.5 Grammar

On the scale of "grammar" (Table 6), the main effect of the writers’ English proficiency was significant (writers’ proficiency: \( F(1.94, 38.00) = 15.74, p < .01, \eta^2 = .29 \)). In other words, raters assessed the essay written by the A2 writer lower than the essays written by the B1 writer and B2 writer. In addition, the main effect of raters’ proficiency was marginally significant (raters’ proficiency: \( F(1, 38) = 3.17, p = .083, \eta^2 = .08 \)) and the simple main effect of each factor was tested statistically in order to understand the interaction. Accordingly, the difference between the means of the scores of each raters’ group (lower proficiency raters: 4.21 and higher proficiency raters: 3.81) were marginally significant. In short, lower proficiency raters tended to give higher scores to the essays than higher proficiency raters. It seems that it was difficult for lower proficiency raters to find the grammatical mistakes in the essays or to score the grammar harshly due to their lack of grammar skills compared with higher proficiency raters.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
<th>( \eta^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>raters’ proficiency</td>
<td>4.273</td>
<td>1</td>
<td>4.273</td>
<td>3.171</td>
<td>0.083</td>
<td>0.077</td>
</tr>
<tr>
<td>error</td>
<td>51.194</td>
<td>38</td>
<td>1.347</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>writers’ proficiency</td>
<td>24.486</td>
<td>1.936</td>
<td>12.650</td>
<td>15.741</td>
<td>0.000</td>
<td>0.293</td>
</tr>
<tr>
<td>writers’ proficiency ( \times ) raters’ proficiency</td>
<td>0.020</td>
<td>1.936</td>
<td>0.010</td>
<td>0.013</td>
<td>0.986</td>
<td>0.000</td>
</tr>
<tr>
<td>error (writers proficiency)</td>
<td>59.114</td>
<td>38.000</td>
<td>1.556</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6 Results of Two-way ANOVA of raters’ proficiency (independent) \( \times \) writers’ proficiency (repeated) (Grammar)
Chapter 5

General discussion

This investigation has shown that English proficiency had little effect on the evaluation of content in this study and the raters seem to have been able to understand the content of all three essays and assess them without linguistic problems. As regards "organization", raters assessed the essay written by the A2 writer lower than the essays written by the B1 and B2 writers. In addition, lower proficiency raters gave a higher score to the essay written by the B1 writer than higher proficiency raters. Regarding "vocabulary", the participating raters agreed that the vocabulary of the A2 writer's essay was the lowest scoring. However, the assessment did not depend on raters' English proficiency and English proficiency seems to have had no effect on the evaluation of vocabulary in this study. Regarding "grammar", raters assessed the essay written by the A2 writer lower than the essays written by the B1 writer and B2 writer. In addition, lower proficiency raters tended to give higher scores to the essays than higher proficiency raters.

In this chapter, I discuss about each research question based on the results.

RQ 1: Does EFL raters' proficiency affect their evaluation of EFL essay writing?

Japanese EFL raters' proficiency affected the evaluation of EFL essay writing and it was attested that the proficiency effect is one of the factors that affect raters' judgment. In other words, the raters' proficiency effect undermines the reliability of evaluation and this effect should therefore be standardized.

Specifically, raters' proficiency affects evaluation when they evaluate organization and grammar, with lower proficiency raters tending to give higher scores using these criteria than higher proficiency raters. Therefore, it is possible that it was difficult for
lower proficiency raters to find the mistakes in the organization and grammar of the essays or to score them harshly due to their lack of English skills compared with higher proficiency raters. In addition, the point of the "organization" and "grammar" scales was not to evaluate the appropriateness of the essay or the meaning of the text, but rather the construction of the composition. To sum up, lower proficiency raters tended to evaluate the construction of the composition higher due to their English proficiency.

In other words, rater proficiency did not affect the evaluation of "content" or "vocabulary." The raters seem to have been able to understand most of the vocabulary in all three essays and assessed it without linguistic problems. Moreover, when an EFL writer writes an essay, the writers usually considered the essay content in their own language, which means it is unlikely that English proficiency affected their evaluation of content and vocabulary.

However, if the essay was too difficult for the raters to understand the content, it seems to have become more difficult for raters whose proficiency was much lower than the writer's proficiency to rate the essays with high reliability. Therefore, it is necessary to investigate this field more, particularly with raters of a wide range of proficiency levels.

**RQ 2: When EFL raters evaluate EFL essays, do they need to consider the writers' proficiency?**

This effect was partly related to writers' proficiency. Specifically, lower proficiency raters gave a higher score to the organization of the essay written by the B1 writer than higher proficiency raters and lower proficiency raters tended to give a higher score to the grammar of the essays than higher proficiency raters. From these results, it is likely that lower proficiency raters tended to give higher scores than higher proficiency raters.

If teachers and students could standardize themselves as raters by comprehending
their own English proficiency and its tendency for each level, the reliability of evaluation would be increased and this may lead to an increase in the effectiveness of teaching.

On the other hand, no relation was found between raters’ proficiency and writers’ proficiency except that mentioned above. In other words, raters were able to evaluate the essays without having to worry about their English proficiency and the writers’ proficiency when they evaluate the content and vocabulary. Moreover, according to Matsuo (2009), students’ peer assessment showed an internal consistency that was as reliable as teachers’ assessment. Considering these results, students’ peer assessments could be applied in the classroom with a relatively high reliability.

As mentioned, teaching and evaluation are inseparable. If the reliability of teachers’ evaluations and students’ peer assessments increases, teachers will come to be able to use teaching methods related to evaluation without hesitation and this will lead to an improvement in the quality of English classes.
Chapter 6
Conclusion and further studies

Various studies have investigated writing and its evaluation and have revealed that rater training can standardize the raters’ effect. On the one hand, “the rating process is complex and there are numerous factors which affect raters’ judgement” (Weigle, 1994), factors that have not been clarified. One of the factors is raters’ language proficiency, which has not been investigated because previous studies have mainly been conducted for NES’s. However, the opportunity for EFL/ESL raters to evaluate EFL/ESL essays is increasing in the era of globalization, and thus it has become increasingly necessary to investigate the proficiency effect on evaluation.

In Japan, it is planned to reorganize the university entrance examination by 2020, and English teachers therefore need to make provisions for the new entrance examination. It may become necessary for teachers to teach students writing and speaking and evaluate these skills. However, the factors affecting evaluation have not been clarified in Japan and it remains difficult for English teachers to teach writing and evaluate their students’ essays with high reliability. As mentioned above, teaching and evaluating are inseparable and if the quality of evaluation increases, the quality of teaching could increase simultaneously. Thus, teachers would come to be able to make provisions for the new entrance examination and students would be able to learn the four skills equally.

In this study, EFL raters’ proficiency affected the evaluation of the organization and grammar of EFL essay writing. The Japanese EFL raters’ proficiency effect should be standardized to increase the reliability of writing evaluation, and it is possible to increase the efficiency of rater training. It is necessary for raters to understand their own English proficiency and their tendencies when evaluating EFL essay writing.
In addition, according to previous research, teaching and evaluating are inseparable and students’ peer assessment is as reliable as teachers’ assessment. Moreover, there is little effect between raters’ proficiency and writers’ proficiency to evaluation of content and vocabulary. For these reasons, it is possible that peer assessment can be applied to English classrooms to teach English writing, and methods of teaching writing can be further developed by understanding the proficiency effect in evaluation.

In conclusion, the raters’ proficiency effect is one of the factors affecting raters’ judgment and this elucidation could lead to an increased effectiveness of rater training and the reliability of the evaluation of EFL essay writing in Japan. One should consider the possibility that students would obtain more opportunities to express their opinions or arguments in English in English classrooms in future.

Finally, there were a number of limitations in the experimental stage of this study and these should be investigated in further studies.

Writers

The three participating writers’ proficiencies were A2, B1, and B2 and they were specifically divided. However, the differences of the writers’ proficiency do not verify the differences of levels of the essays due to the complicated writing process, which depends not only on writers’ language proficiencies but also the other skills of writing. Therefore, it is necessary to increase the number of writers and the size of the experiment in order to define the result more clearly.

Raters

Forty university and graduate students in Japan participated in this study as raters and their English proficiencies ranged from low A2 to high B2: however, there were only
two participants of B2 level. In this study, the number of raters and raters' proficiency levels were limited and they were divided into just two groups: lower proficiency raters and higher proficiency raters. The participants, particularly the raters whose proficiency was B2 level or higher, should have participated more in this study and the effect of raters whose proficiency is much higher should be investigated in further studies to increase the validity of the results.

**Rating scale**

An original rating scale was used in this study and its reliability and validity were not revealed, so the result of evaluation was not necessarily appropriate even though the higher proficiency raters used it for evaluation. However, this is not a major problem as the aim of this study was not to determine the reliability of evaluation, but the differences between raters of different English proficiency levels: all raters used the same rating scale and the differences were defined by raters' proficiency.
References


Shizuka, T (2002). *Eigo test no sakusei tatsujiin manual* [The expert manual for
developing the English tests]. JP: Taisyukan.


The Central Council for Education. (2014). *Atarashii jidai ni fusawashii kodai setsuzoku no jitsugen ni muketa koto kyoiku, daigaku kyoiku, daigaku nyugakusya senbatsu no ittaikei kaikaku ni tsuite [The reorganization for unification of the education in high schools and universities and university entrance examinations in order to actualize the connection between high schools and universities, which is fitting for the new era]*. http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo12/shiryo/_icsFiles/afieldfile/2014/11/11/1353318_02_1.pdf (accessed 2016-12-18).


A foreign visitor has only one day to spend in your country. Where should this visitor go on that day? Why?

Name ___________________
Appendix B Direction for Writers

今日はお忙しい中、お集まりいただきありがとうございます。私は兵庫教育大学大学院
において、自由英作文の評価に、評価者の英語熟達度が与える影響についての研究を行って
います。大変発表と申します。

この実験は、実験の参加者の方々に、こちらで用意した自由英作文を評価していただき、
その評価の結果と評価した方の英語熟達度（英語能力）の関係を分析することで、英語教育
学、特に評価の分野において貢献しようとするものです。

今回は、その実験で使用するための自由英作文を書いて頂きたく、お集まり頂きました。

今回書いて頂いた自由英作文は、名前など個人情報が特定される情報を隠し、30名から
40名ほどの実験参加者の方々に評価されます。各評価項目は6点満点で、評価の観点は以
下の4点です。以下の評価観点を念頭に置いて自由英作文を書いて頂きますので、よく読
んで理解してください。

①内容
主張が明白である。
解答は問題の趣旨と一致しており、それぞれの考えに適切な根拠が示されている。

②構成
つながりが適切に使用され、導入・展開・結論の構成になっている。
また文章は論理的であり、流れがスムーズである。

③語彙
多様な語や表現が適切に選択、使用されている。
また、読者が理解しやすいような工夫がみられる。

④文法
多様な文・構文や表現が適切に使用されており、
主語・動詞の一致、時制、数、代名詞、冠詞、前置詞などに誤りがない。

今回の実験の内容を理解し、参加を同意してくださる場合は、測定同意書へ署名
をお願いします。
Appendix C Writing Samples

Essay 1 (A2 writer)

自由英作文問題

下記のテーマについて、150字程度の英語で自分の考えを述べなさい。その際、筆記体は
使用しないようにしてください。制限時間は30分です。

A foreign visitor has only one day to spend in your country. Where should this
visitor go on that day? Why?

I think this visitor should go to Kami town in Hyogo on that day. Because the sea is very
beautiful. And foods, for example, Kasumi clubs and fishies and Tajima beafs are very
nice. And Kinosaki onsen is very good. And kounotori in Toyoka is very cute. And there
is Amarube tekkyo and Misaki todai in Amarube. Both Amarube tekkyo and Misaki
todai are high. Kami town is very good place. So this visitor should go to Kami town on that
day.

内容____ 構成____ 話順____ 文法____ 合計____点
I think the visitor should go to Kyoto on that day. I have three reasons.

First reason is that Kyoto has many Japanese culture, for example, kimono, green tea and maiko-san. Foreign visitors are very pleased because they can get a good experience.

Second reason is that there are many foreigners in Kyoto. They don't worry and have a good time.

Last reason is that we can see many sightseeings on only one day. We recommend that they go to Kyoto if they want to visit many places.

Kyoto has many historical buildings and many delicious foods. I want foreigners to know Japanese history and culture.

So I recommend foreigners to visit Kyoto.
Essay 3 (B2 writer)

自由英作文問題

下記のテーマについて、150字程度の英語で自分の考えを述べなさい。その際、筆記体は使用しないようにしてください。制限時間は30分です。

A foreign visitor has only one day to spend in your country. Where should this visitor go on that day? Why?

I think it depends on his age or the season, but I will recommend him go to Arashiyama in Kyoto. There are some reasons for it.

First I want him to take trains while he moves. This is because he can find how exact Japanese trains run or come on time. So he can imagine the hardworking temperment of Japanese people by his experience.

Second Arashiyama is popular sightseeing spot, so people are willing to welcome foreigners. Some Japanese people speak English or Chinese. They will help him understand Japanese culture. Also he can experience traditional Japanese culture. For example, kimono, tea ceremony, eating dumplings, visiting temples and so on.

Therefore, I think that Arashiyama is the best place for him staying Japan for a shot time because he can enjoy a lot for only a day.

内容____構成____語彙____文法____合計____点.
Appendix D Direction for Raters

本日お忙しい中お集まりいただきありがとうございます。本実験は、実験の参加者の方々にこちらで用意した自由英作文を評価して頂きます。その評価の結果と評価した方の英語
熟達度の関係を分析することで、英語教育学、特に評価の分野において貢献しようとするものです。
実験の流れは以下の通りです。

①自由英作文の評価 (30分)
②英語熟達度の測定 (30分)

①自由英作文の評価に関して
・3つの自由英作文を合計30分以内で評価して頂きます。
・2つの自由英作文は、前後せずに提示された順番に評価してください。
・評価する際には評価項目に従って、各自自由英作文を独立して評価してください。
(3つの自由英作文を比べながら評価することはありませんお願いします。
・4つの評価項目があり、各6点、合計24点満点で評価して頂きます。
・各評価項目の各点数に、その点数をつける根拠となる文章を提示しています。その文章の中には評価の観点が2つ以上存在しますが、優先して頂きたい評価観点の順に文章を構成しています。
例「きれいな字であるが、文法的な誤りが多い」ものが3点に相当するとした場合、文法的には完璧でも字がきたない場合は2点以下ということになります。

②英語熟達度の測定に関して
・今回の実験では、Oxford Universityの"Quick placement test"を使用します。
・制限時間は30分です。
・問題用紙に書き込みはしないようにお願いします。

今回の実験の内容を理解し、参加を同意していただく場合は、測定同意書へ署名をお願いします。
Appendix E Rubrics

<table>
<thead>
<tr>
<th>内容</th>
<th>6</th>
<th>主張が明白である。解答は問題の趣旨と一致しており、それぞれの考えに適切な根拠が示されている。</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>主張が理解できる。解答は問題の趣旨に沿うものであるが、根拠が明白でない部分がある。</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>主張が明白でない。解答に示されている情報は限られており、それを支える根拠に説得力がない。</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>主張が理解できない。解答は問題の趣旨と一致しておらず、根拠となる考えも示されていない。</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>構成</th>
<th>6</th>
<th>つなぎ言葉が適切に使用され、導入・展開・結論の構成になっている。また文章は論理的であり、流れがスムーズである。</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>つなぎ言葉が使用され、構成がはっきりしているが、文章の流れがスムーズでない部分がある。</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>使用されているつなぎ言葉は限られている。また、文章の流れが明白でない部分があり、流れがスムーズでない。</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>つなぎ言葉がほとんど使用されていない。また、文章の構成が明白でない。</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>語彙</th>
<th>6</th>
<th>多様な語や表現が適切に選択、使用されている。また、読者が理解しやすいような工夫がみられる。</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>多様な語や表現が使用されているが、その選択が適切でない場合がある。</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>示されている語や表現は限られており、誤りもいくつか認められる。</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>語や表現に関する知識がほとんどなく、多くの誤りが含まれるために意味が不明瞭である。</td>
</tr>
<tr>
<td></td>
<td>内容</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>多様な文・構文や表現が適切に使用されており、文法的（主語・動詞の一致、時制、数、代名詞、冠詞、前置詞など）に誤りがない。</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>適切な文・構文や表現が使用されているが、文法的（主語・動詞の一致、時制、数、代名詞、冠詞、前置詞など）な誤りがいくつか認められる。</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>文・構文や表現は限られており、文法的（主語・動詞の一致、時制、数、代名詞、冠詞、前置詞など）な誤りもたびたび認められる。</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>文法的（主語・動詞の一致、時制、数、代名詞、冠詞、前置詞など）な誤りが多く含まれるために、意味が不明瞭である。</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>