A comparison of Aptis trained raters and Japanese teachers' holistic scores and judgments of Japanese students' Aptis writing performance

by Chiho Takeda

British Council's Master's Dissertation Awards 2023
Commendation

A comparison of Aptis trained raters and Japanese teachers' holistic scores and judgments of Japanese students' Aptis writing performance

Chiho Takeda

Dissertation submitted in partial fulfillment of the requirements for the degree of

MA in TESOL

Supervised by: Emma Bruce and Parvaneh Tavakoli

Proofread by: Lynda O'Brien

School of Literature and Languages

University of Reading

Submission date

12 September 2022

Word count

15,682

Table of Contents

List of tables v

List of figures vi

List of appendices vii

List of abbreviations viii

1. Introduction 1

2. Literature Review 3

2.1 Introduction 3

2.2 Validity 3

2.3 Weir’s (2005) socio-cognitive framework 4

2.3.1.1 Construct validity 6

2.4 Reliability of writing assessment 7

2.4.1 Rater reliability 7

2.5 Rater variance for writing assessment 8

2.5.1 Rater bias 8

2.5.2 Rater background 9

2.5.3 Rating environment and conditions 10

2.6 Rating scales 11

2.6.1 Holistic scale 11

2.6.2 Analytic scale 12

2.7 Evaluation of writing 13

2.7.1 NES and NNES evaluation of error 14

2.7.2 Evaluation of authentic written work 15

2.8 Japanese context 17

2.9 Summary 18

3. Methodology 20

3.1 Introduction 20

3.2 Design Approach 20

3.3 The data collection instruments 22

3.3.1 Evaluation sheet 23

3.4 The participants 30

3.4.1 Interview participants 32

3.5 Ethical considerations 33

4. Results 35

4.1 Introduction 35

4.2 Research Question 1 36

4.2.1 Reliability 36

4.2.2 Comparison of scores 36

4.3 Research Question 2 39

4.3.1 Reasons for holistic rating 39

4.3.2 Rater perceptions of grading analytic features 43

5. Discussion 49

5.1 Introduction 49

5.2 Research Question 1 49

5.3 Research Question 2 51

5.4 Implications and limitations of the study 58

5.4.1 Implications 58

5.4.2 Limitations and suggestions for further research 60

6. Conclusion 62

7. References 64

8. Appendices....71

Appendix 1: Type of research: NES and NNES error gravity....71

Appendix 2: Type of research: NES and NNES evaluation of authentic written work....72

Appendix 3: Rater training sheet – English version....73

Appendix 4: Rater training sheet – Japanese version....78

Appendix 5: Evaluation sheet – English version....81

Appendix 6: Evaluation sheet – Japanese version....84

Appendix 7: Questionnaire – English version....87

Appendix 8: Questionnaire – Japanese version....90

Appendix 9: Semi-structured Interview “Question Examples”....93

Appendix 10: Coding categories for the semi-structured interview....94

Appendix 11: Coding example for the semi-structured interview....95

Appendix 12: Profile of participants....96

Appendix 13: Consent form – English version....98

Appendix 14: Consent form – Japanese version....99

Appendix 15: Information sheet – English version....100

Appendix 16: Information sheet for participants – Japanese version....101

Appendix 17: Raters’ reason for holistic scoring....102

Appendix 18: Feature ranking of the difficulty of holistic score determination....104

Appendix 19: Essay example....105

List of tables

Table 1: Operationalization of this study..22

Table 2: Organization of questions 1 to 10 from the questionnaire..27

Table 3: Details about the interviewees from the AR group 33

Table 4: Details about the interviewees from the JT group..33

Table 5: Breakdown of data used to answer research questions..35

Table 6: Comparisons of ARs and JTs' holistic scores on the 20 essays 38

Table 7: Reasons for holistic scores (ranking and percentage)..41

Table 8: The most difficult criterion...44

Table 9: The least difficult criterion 44

List of figures

Figure 1:A socio-cognitive framework for conceptualizing writing test performance.............5

Figure 2: Degree of participants 31

Figure 3: Rater training experience...31

Figure 4: Years of teaching English...32

Figure 5: Reasons for holistic scores (ARs) 40

Figure 6: Reasons for holistic scores (JTs) 40

List of appendices

Appendix 1: Type of research: NES and NNES error gravity..71

Appendix 2: Type of research: NES and NNES evaluation of authentic written work...........72

Appendix 3: Rater training sheet – English version 73

Appendix 4: Rater training sheet – Japanese version 78

Appendix 5: Evaluation sheet – English version 81

Appendix 6: Evaluation sheet – Japanese version 84

Appendix 7: Questionnaire – English version 87

Appendix 8: Questionnaire – Japanese version 90

Appendix 9: Semi-structured Interview “Question Examples” 93

Appendix 10: Coding categories for the semi-structured interview 94

Appendix 11: Coding example for the semi-structured interview...95

Appendix 12: Profile of participants...96

Appendix 13: Consent form – English version...98

Appendix 14: Consent form – Japanese version...99

Appendix 15: Information sheet – English version 100

Appendix 16: Information sheet for participants – Japanese version....................................101

Appendix 17: Raters’ reason for holistic scoring 102

Appendix 18: Feature ranking of the difficulty of holistic score determination....................104

Appendix 19: Essay example...105

List of abbreviations

AR: Aptis rater

CEFR: Common European Framework of Reference for Languages: Learning, teaching, assessment

CLT: Communicative Language Teaching

EFL: English as a Foreign Language

ESL: English as a Second Language

JT: Japanese high school teacher

L1: first language

L2: second language

NES: native English speaker

NJS: native Japanese speaker

NNES: non-native English speaker

SLA: second language acquisition

Declaration

I declare that this dissertation has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where states otherwise by reference or acknowledgement, the work presented is entirely my own.

Acknowledgments

On the completion of this dissertation, I wish to thank the numerous people who have helped and supported me during my MA research. Firstly, I would like to say a special thank you to Dr. Emma Bruce, my supervisor, and to Professor Parvaneh Tavakoli, both of you have given me a great deal of guidance, constantly encouraging and believing in me. As a researcher, a teacher, and a working woman, I cannot help but respect you wholeheartedly. I cannot thank you enough.

I would also like to express my heartfelt thanks to all of my professors, teachers, tutors, study advisors, my proofreader, and my classmates, each of whom has contributed to my learning. Also my thanks to the University of Reading, where I was able to gain all of this valuable experience. Importantly, thank you to my study mates for the motivational boosters we have shared.

I am very grateful to all the Japanese teachers and Aptis raters who took part in this research and generously shared their experiences with me. I sincerely thank them for their time and commitment: without their voluntary participation and cooperation, this research would not have been possible.

I am deeply thankful to the members of small world cafe, and my friends for giving me such warm support during my dissertation journey.

Finally, I would like to thank my parents and family, who have supported me during this academic year in the UK. Thank you, Dinneer, for respecting and continuously supporting my decisions and for making my academic life a much happier one. Thank you!

Abstract

Both first language (L1) and second language (L2) English speakers are frequently employed and involved in English language teaching and assessment. The potential for bias in L2 writing assessment as a result of raters' linguistic and cultural backgrounds is the focus of this study. The purpose is to compare the rating behavior of two groups of raters with different backgrounds: Japanese high school teachers and trained Aptis test raters. 10 Japanese and 10 Aptis raters evaluated 20 Aptis essays written by Japanese teenagers using a 10-point scale, they then stated their three main reasons for allocating their overall score in order of importance.

Although there was no significant difference between the two groups in the overall scores, allocated, the differences in their reasons reveal that the Japanese teachers focused on content (task achievement and organization), while the Aptis raters focused on language use (grammar, sentence structure, and vocabulary). The implication of this study raises issues concerning validity in language testing. Although the overall scores of the two groups were almost consistent, the assessment processes and perspectives were markedly different. The findings from this study undertaken in the Japanese context find support among similar studies conducted in other countries (Rao & Chen, 2020; Shi, 2001). This suggests rater bias may be a global issue for language assessment.

1. Introduction

Writing skill is considered important in English as a second and foreign language learning (ESL and EFL) context, however, the assessment of writing is a more complex process than that required for multiple-choice questions. Writing performance is usually evaluated by one or more raters using a set of rating criteria. Unlike automated scoring systems, human raters have a range of experiences, values, and backgrounds. This may influence the quality of rating and a number of studies have investigated the issue (e.g., Cumming et al., 2002; Weigle et al., 2003). It seems that linguistic and cultural background plays an important role because both native English speaking (NES) and non-native English speaker (NNES) are frequently employed and involved in English language teaching and assessment in the EFL environment. There might be some concerns about whether NES teachers and NNES teachers assess consistently when using the same rating method as a result of cultural and linguistic background differences influencing their rating behavior. In this line of study, some studies examined raters' perceptions of speaking performance (e.g., Kim, 2009; Zhang & Elder, 2011), while others examined differences in raters' behaviors when assessing writing tests (e.g., Johnson & Lim, 2009; Shi, 2001).

These empirical studies have been conducted in various university contexts in various countries. In contrast, there are a limited number of studies that investigate the issue in high school contexts in EFL countries (Hughes & Lascaratou, 1982). In Japan, native Japanese speaker (NJS) English language teachers assess together with NES teachers in large-scale tests and classroom tests. With the growing demand for writing tests in Japan for both the national entrance examinations (Saito, 2019), and the national curriculum for English language education (Hosoki, 2011), the findings from this study are expected to contribute to the understanding of how linguistic background affects teachers' assessment of EFL students writing, to identify the processes Japanese high school teachers' use and their attitudes toward the writing assessments they make. While the study focused on Japanese high school teachers, it may also have implications for test administrators, training and recruiting raters, individual teachers, and raters who share similar EFL contexts around the world.

In summary, based on the studies mentioned above, the existence of rater linguistic and cultural background-related bias in second language (L2) writing assessment is the focus of this study. It concentrates on trained Aptis test raters and NJS high school English teachers to investigate if these two rater groups differ in their scoring and rating processes.

This study has six chapters. The following chapter reviews the theoretical framework for this study, which is based primarily on rater variance in L2 teaching and assessment, focusing on a particular socio-cognitive framework (Weir, 2005). Also, investigations into rater variance associated with the cultural and educational background are presented. Two research questions are derived from this literature review. The third chapter presents the rationale behind the research approach used and the procedure for the data collection and analyses. Chapter 4 then presents the results of the data analysis which are discussed in Chapter 5 along with a number of implications and limitations of the study. Finally, in the concluding chapter, the current research is summarized, and recommendations are made for further study.

2. Literature Review

2.1 Introduction

The quality of assessment in education is usually associated with validity and reliability. It has been discussed that some factors, such as rating scales and rater variance, might affect the validity and reliability of language tests (Knoch et al., 2021). The effect of rater variance on scores, for example, may be the result of differences in raters' educational and linguistic backgrounds, which are not relevant to a candidate's performance (Cumming et al., 2002). This chapter presents definitions of reliability and validity in writing assessment and focuses particularly on Weir's (2005) socio-cognitive framework. It then considers the different types of evidence that are significantly important for validating a writing assessment. Previous research on NES and NNES writing evaluation is discussed and a summary of the major findings is presented. The Japanese educational context in which the study is situated – e.g., the English language curriculum and entrance examination – is briefly explained to provide readers with a better insight into the setting. Finally, drawing on this review of the literature, the research questions are proposed.

2.2 Validity

Traditionally, validity has been considered in multiple ways, including those related to construct, content, and criteria. However, the concept of validity has shifted over the years (Chapelle, 2012) and expanded from a sole focus on test properties to investigations into their use. Views on validity drastically alerted after Messick's (1989) study which introduced an integrated view, suggesting validity is relevant not only to assessment and scores but also to inferences that may be made from test scores. Messick (1989) defines validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" (p.13).

Reliability has previously been considered a separate concept from validity, more recently, it has come to be regarded as a characteristic of validity. This more recent approach views reliability as evidence for validity. Following Messick (1989), Weir (2005) proposed his socio-cognitive framework for the validation of the test; this is described in the next section.

2.3 Weir's (2005) socio-cognitive framework

Weir (2005) proposed a socio-cognitive framework for test validation which has been extended, adapted, and modified in several recent publications (e.g., O'Sullivan, 2011; Shaw & Weir, 2007). Figure 1 presents a graphical representation of the framework as adapted from Shaw and Weir (2007). The framework comprises five basic elements of validity: contextual, theoretical, social cognitive, scoring, consequential, and criterion-related. Each of these is described below.

There are several advantages to using a framework that allows the validity of the used tests to be evaluated critically in the light of the educational context, while at the same time allowing data from learners' performance to be used to inform feedback decisions for learners. Applying this framework helps ensure that, at the development stage, teachers consider their approaches to test development and assessment validation by incorporating contextual, cognitive, and scoring parameters. The framework can also guide the generation of evidence for the successful operationalization of these features during the test implementation phase. Moreover, it covers both development stages, i.e., a priori and a posteriori. Context and cognition are considered a priori validity as evidence collected before the test while scoring, consequential, and criterion-related validities are considered a posteriori validity because the evidence is collected after the test.

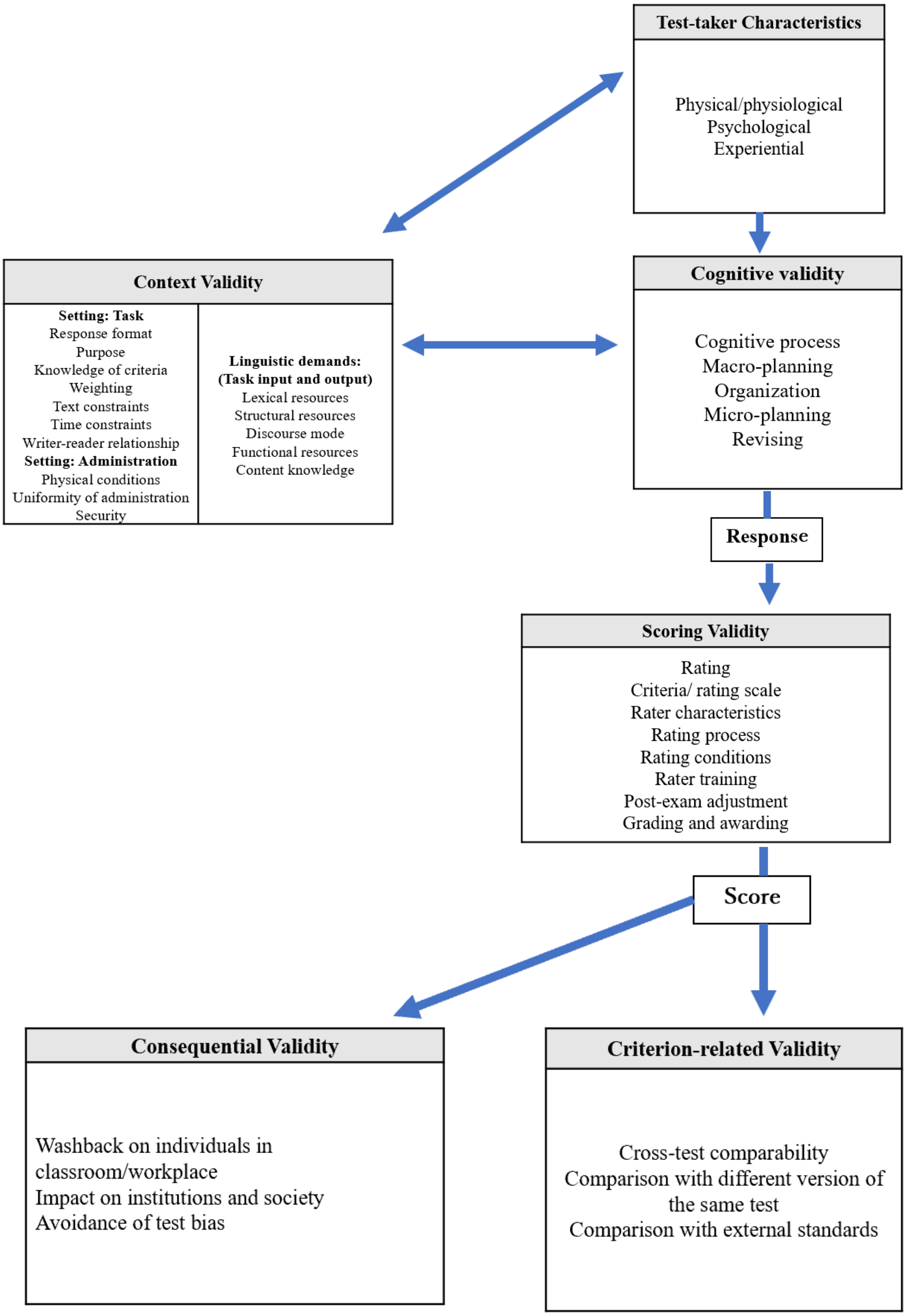


Figure 1:A socio-cognitive framework for conceptualizing writing test performance

The various types of validity are outlined as follows and defined in detail in the proceeding sections. Context validity is related to testing items, performance conditions, and operations. It is important to ensure that the task-based performance conditions are representative of real-world constructs. Meanwhile, cognitive validity is defined as a measure of how faithfully a test represents the cognitive processing involved in performing the same task in contexts other than the test itself, i.e., in real-world contexts. The third type is scoring validity, it concerns evaluation criteria and scales; rating procedures, training, and conditions; rater characteristics; the rating process; post-test adjustments; grades, awards, and penalties. According to Weir (2005), “scoring validity is concerned with all the aspects of the testing process that can impact the reliability of test scores” (p.48). Consequential validity relates to the extent to which test scores are interpreted and acted upon in an intended manner, the degree to which a test produces intended or unintended consequences, and if it causes washback on teaching and learning, and on society generally. Finally, criterion-related validity represents the comparison of a test score with an external source that was administered at the same time as the test and that measures the same competencies as the test taker (Shaw & Weir, 2007).

2.3.1.1 Construct validity

Shaw and Weir (2007) claimed a symbiotic relationship between contextual validity, cognitive validity, and scoring validity from construct validity. In the development of writing assessments, to define the construct, test developers might discuss which ability should be tested or given priority, abilities may relate to coherence, grammar, vocabulary, or accuracy. Clearly, writing assessment is subjective and scores depend on the raters’ subjective interpretation of the rating criteria. Therefore, construct validity in writing tests might be threatened by raters’ judgments. Many researchers have discussed the validity of the speaking and writing test (e.g., Bachman & Palmer, 1996; Lynch & McNamara, 1998; McNamara, 1996; Messick, 1994).

2.4 Reliability of writing assessment

Both the validity and reliability of language assessment have been investigated in second language acquisition (SLA) research. In Weir's framework, reliability of a language assessment is then classified into validity of the scoring. Firstly, reliability refers to the consistency of measurement over time, as they relate to rater and content (Fulcher, 2010). For instance, if the same person takes the same writing tests repeatedly in a short period of time, then if the test is reliable, they should receive a consistent score every time. In a good test, a variance in scores is due to different levels of ability being measured without being affected by other features. In a study by Lado (1961), test scores were affected by three major factors: the test itself, its administration, and how it was scored. The test itself is associated with what aspects of language to test. If the content of the test is quite different from what learners learn, reliability is reduced. Test administration is an important factor and includes appropriate timing and testing circumstances. To maintain high-quality, reliable tests, consistency of test time and test taker's conditions – such as the room used and the invigilators present – are vital.

2.4.1 Rater reliability

When a group of raters judges the same performance, it is almost impossible their scores will match with each other. The extent of the variation between the scores the raters award is known as inter-rater reliability (Green, 2014). Moreover, even when the same rater judges the same performance twice, the score may change slightly. This extent of the variation is a matter of inter-rater reliability; thus, it is related to the consistency of the judgment between multiple raters; whereas intra-rater reliability is related to the extent of variation of the same person's judgment.

Rater reliability attracts the most attention from researchers (Lumley & McNamara, 1995; Shohamy et al., 1992), especially in subjective or performance-based tests such as oral interviews and written compositions. In contrast, an objective or closed response test, e.g., multiple choice, is much easier to evaluate and assess consistently since generally, only one answer is correct. Unlike an objective test, a writing test is much more complicated to evaluate because the test taker's ability might be interpreted differently, and the score might be influenced by each rater's individual perspective. For example, Weigle et al. (2003) found that raters from the English department were more severe when rating language usage than raters from a different department.

2.5 Rater variance for writing assessment

Rater variance is the variation in scores awarded by raters with the same rating scale (McNamara, 2000). Raters bring their own experiences, backgrounds, and values to writing assessment, thus it is believed that one of the significant issues in writing assessment is raters' variability (Cumming et al., 2002).

2.5.1 Rater bias

Raters may develop bias toward a variety of factors in their evaluation. They may become more or less tolerant towards an essay with certain characteristics without perceiving such an action. An example is associated with the rhetorical patterns used in an essay. That is, American and Japanese university teachers rated essays written by students of their respective nationalities and found that they appreciated familiar rhetorical patterns (Kobayashi &

Rinnert, 1996). The results showed that raters reacted positively to essays with familiar rhetorical patterns. Du and Brown (1996) explored potential bias based on task type, ethnic group, and gender. The results showed that significant rater bias accounted for 11% of all interaction effects and that raters' biases were directed towards task types based on ethnicity and gender.

2.5.2 Rater background

It has also been investigated whether various rater background variables affect rating quality. Rater groups may differ in terms of their rater training experience, teaching experience, and their professional and linguistic backgrounds. Many studies found that rater training experience shows a difference in rater severity and consistency (Kang et al., 2019; Shohamy et al., 1992), and the quality of ratings differs depending on the content of the rater training and the duration of the training.

Professional background is seen as one of the key factors. Some studies have investigated the differences in evaluation between university teachers of English as a subject and other teachers (Cumming, 1990; Weigle et al., 2003). It is found that raters from different disciplines tend to base their assessment on conventions and highlighted discourse communities that influence the criteria by which they rate students' writing. For example, raters who are teachers of English focused more on language use, while raters from the psychology department emphasized content. In addition to research on how differences in teacher expertise affect scoring, there is also research on how teaching experience and rater experience affect the quality of scoring. Compared to other factors, the impact of teacher experience has been less well researched. Mainly, non-teacher and teacher evaluations have been compared. Huang (2013) found that teachers tend to focus on elements that represent their instructional goals. Significantly more studies have investigated the impact of rater experience compared to the impact of teaching experience. Barkaoui (2010) investigated the behavior of expert and novice raters and found that novice raters tended to be more lenient and were often affected by variations in the rating scale compared to experienced raters.

Furthermore, some researchers have raised questions about the construct validity of the rating of different linguistic background groups using the same rating scale when assessing L2 writing (Kobayashi, 1992; Lee, 2009; Rao & Li, 2017; Sheorey, 1986; Shi, 2001). Although many studies compare the quality of NES and NNES raters in their writing assessments (Connor-Linton, 1995; Hijikata-Someya et al., 2015; Kobayashi & Rinnert, 1996; Rao & Chen, 2020; Shi, 2001), this terminology tends not to be used now because some researchers argue that the distinction between NES and NNES is becoming increasingly unclear due to increased movement across borders and the trend that many people undertake education and work in English (Kachru, 1992). However, the results still provide useful insights. This research is meaningful since the results reveal significant differences in the writing assessment between the two different groups. This is expected to be helpful for language teachers and test administrators to better understand the rater's tendencies and behavior from the perspective of raters' linguistic and cultural backgrounds; this information can feed into rater training and recruitment.

2.5.3 Rating environment and conditions

The characteristics of the physical environment and working conditions in which the raters mark may degrade the quality of the raters. Raters may be distracted from their concentration by noise, time pressure, and physical fatigue. Ling et al. (2014) investigated the fatigue of aters rating the TOEFL iBT speaking test and found that the level of rating quality in their work would fluctuate during a long shift of marking.

2.6 Rating scales

Another important factor that may affect writing assessment is the nature of the rating scale (McNamara, 1996). The rating scale is the framework for rating which usually consists of several distinct levels. Rating scales are traditionally made from a set of levels defined by descriptors (Fulcher, 2010). A descriptor is defined as " a prose description of a level of performance on a scale" (Fulcher, 2010, p. 320). Rating scales also provide descriptions of candidates' expected discourse at the different levels of performance, typically rating scales have between three to nine levels. Once the statement is defined, raters choose the single descriptor that best describes the performance of the test-takers. Weigle (2002) categorized these into three scales: primary trait, holistic, and analytic. The primary trait scales are designed for specific writing tasks and evaluated within a narrowly defined range of discourse, e.g., explanation or persuasion; meanwhile, holistic and analytic scales can be applied to multiple tasks. Due to this limitation of primary trait scales, the value of holistic and analytic scales has become more important (Shaw & Weir, 2007). Certainly, although there are other scale formats in use for L2 performance assessment, these two key classifications for rating scales are highlighted in applied linguistic theory, as explained in the following sections.

2.6.1 Holistic scale

In holistic rating, a single score reflects the overall quality of the performance, rather than scores being given for specific features, e.g., separate scores for context and grammar (Davis, 2016; Shaw & Weir, 2007). Holistic scales are considered to have the practical advantage that they are less time-consuming than other formats, and therefore are more effective and less expensive to implement than analytic scales. It is much easier to rate with a single score than to rate several features, the latter requires reading repeatedly to evaluate each feature. As a result, holistic scoring is useful, especially for large-scale language tests, such as entrance examinations. Another advantage is that the holistic scale focuses the rater's attention on the strengths of the writing, rather than its single shortfalls (e.g., poor sentence structure). As a result, test-takers are rated by what they do well in their performance (Shaw & Weir, 2007; White, 1984, 1985).

On the other hand, holistic scoring has significant disadvantages, particularly in the second language learning context. It is a rank order process using just one score. It cannot provide specific feedback for each ability and learners might miss opportunities to understand their strengths and weakness in English writing. In a sense, a holistic scale might prevent second language learners from improving various aspects of their writing skills after receiving the score because they do not know which areas to focus on. Knoch et al. (2021) argue that foreign language learners tend to have uneven skills, so holistic scales might hide their weaknesses. For instance, even though the test taker's grammar use may be less advanced than their sentence structure, various elements of their language ability are evaluated with just one performance level, thereby hiding this weakness. Another disadvantage of holistic scales concerns the interpretation of the rating scale which might differ between raters. Cumming et al. (2002) criticize holistic scales since the specific nature of the assessment constructs remains uncertain.

2.6.2 Analytic scale

Many studies show that analytic scoring is usually more reliable than holistic scoring as rating criteria include several aspects of writing, as opposed to a one-score scale (Kudo &

Negishi, 2002). This makes it useful as raters are alerted to several aspects of test-takers performance in more detail. Criteria for analytic scales differ, depending on the test purpose, and may include such elements as content, organization, grammar, vocabulary, accuracy, and coherence.

Despite this being a more time-consuming process, analytic rating is supported by many raters because it allows test-takers to receive more specific feedback on different aspects of their writing. For instance, the result may show that the writer’s organization is strong, but their vocabulary use is weak. Also, an analytic scale has advantages for inexperienced raters, i.e., as the rating scale is separated into elements, raters tend to be more confident about the mark they give (Barkaoui, 2011).

However, practical drawback is that rating is more time-consuming and less economical than when using a holistic scale, making it more problematic to implement in large-scale tests (Marsh & Ireland, 1987).

2.7 Evaluation of writing

Numerous studies have examined how NES and NNES assess EFL learners’ writing performance (Connor-Linton, 1995; Hughes & Lascaratou, 1982; Kobayashi & Rinnert, 1996; Kobayashi, 1992; Lee, 2009; Marefat & Heydari, 2016; Shi, 2001; Tatsukawa, 2018; Zhang & Elder, 2011).

Among empirical studies which have collected data from both NES and EFL raters using writing samples written by ESL or EFL learners are Japanese university students (Connor-Linton, 1995; Hijikata-Someya et al., 2015; Kobayashi & Rinnert, 1996; Kobayashi, 1992); Greek high school students (Hughes & Lascaratou, 1982); Chinese university students (Rao & Chen, 2020; Santos, 1988); and Iranian university students (Marefat & Heydari, 2016).

2.7.1 NES and NNES evaluation of error

Much of the literature has investigated the differences between NES and NNES raters' reactions to L2 writing (Hughes & Lascaratou, 1982; Hyland & Anan, 2006; James, 1977; Kobayashi, 1992; Rao & Li, 2017; Sheorey, 1986). Some of the empirical studies which focused on the evaluation of errors are presented in Appendix 1. One line of study focused on how NES and NNES raters rated errors presented within individual sentences, not in longer texts. As one of the early researchers to attempt to explore the differences between NES' and NNES' writing evaluation, James (1977) asked raters to underline the writers' errors in each sentence and then rate the sentence from 1-5, where 5 represents the most severe errors and 1 the least severe. He found that NNES teachers were less tolerance toward errors than NES. Hughes and Lascaratou (1982), and Sheorey (1986) were inspired by James' (1977) investigation and used it as a blueprint for their research to examine the difference between NES and NNES in their evaluation of the writers' errors. Like James (1977), they found that NES raters NNES teachers showed a lower tolerance for linguistic errors than NES. In terms of the criteria for error gravity, NES focused on the criterion of intelligibility while NNES placed emphasis on rule infringement.

Most research has shown similar results；NNE raters were more severe toward written errors, generally. The drawback to this line of study is that it focused only on sentences, not more extended written texts. Kobayashi (1992) pointed out that this research does not examine the global features of writing, such as organization, coherence, and cohesion.

The next group of error gravity studies set out to examine the errors in an authentic piece of writing (Hyland & Anan, 2006; Kobayashi, 1992; Santos, 1988). Similarly, this line of the study found that NNES raters generally are more severe towards writers' errors than NES raters (Hughes & Lascaratou, 1982; Hyland & Anan, 2006; Rao & Li, 2017; Santos, 1988). On the other hand, research in this area reveals a mixed picture; Kobayashi (1992) has shown that NNES raters reacted less leniently toward grammatical errors. These contradictory results were based on guided rating criteria which might have influenced the rater's judgment (Kobayashi & Rinnert, 1996; Santos, 1988).

2.7.2 Evaluation of authentic written work

The studies discussed in the previous section have focused on raters' evaluations of error gravity in writing. There is a significant shift in the literature that focuses more on the content of writing than on linguistic correctness (Hyland, 2019). Instead of an assessment of sub-skills, many studies have focused on the assessment of authentic writing (Weigle, 2002). Some empirical studies focused on the evaluation of authentic written work are presented in Appendix 2.

Kobayashi and Rinnert (1996) investigated how NES and NNES evaluate EFL students' compositions containing different culturally influenced rhetorical patterns. Their participants were divided into four groups: one made up of NJSs who are university students and have received English writing instruction; the second comprised NJSs who are university students, but with no English writing instruction; the third group represents NES teachers, and the last represents NNES teachers.

Overall, it was found that NJS without English composition instruction rated Japanese rhetoric more positively, NJS students with English composition instruction rated features of both patterns positively, and NES teachers preferred American rhetoric.

Another focus of research using an authentic writing sample is to examine the difference between the two groups using holistic or analytic ratings. Much research has examined whether NES and NNES raters differ in their evaluation of students' writing performance and the evaluation process using a holistic scale, but the results are slightly contradictory. For example, Santos (1988) asked 144 NES professors and 34 NNES professors to rate compositions written by Chinese and Korean students. The holistic results indicate that the NNES professors were more severe in their rating than the NES. A similar study was conducted by Marefat and Heydari (2016); they investigated whether 72 NES teachers and 72 Iranian teachers differed in their writing assessment. Similarly, they found that NES teachers were more lenient than Iranian teachers in their holistic rating.

On the other hand, Shi (2001) found that the NES teachers gave lower marks than NNES teachers. She investigated twenty-three Chinese and NES raters rating ten essays using a holistic scale, then raters gave three reasons for their rating. Due to the qualitative reasons for the holistic rating, Shi (2001) found significant differences in the feature analysis and the Chinese teachers rated content and organization severely whereas English-background teachers rated content and language use positively. NNES gave fewer comments on language use than NES while NNES gave more comments on the organization than NES. Similarly, Brown (1991) found that NES focused more on sentence-level features, while NNES teachers focused more on structure. Connor-Linton (1995) also discovered the different rater behaviors; NES raters tended to focus on both inter-sentential features of the discourse and specific intra-sentential grammatical features, while JNS tended to focus on matters of accuracy (word choice, grammar, and content).

As a result of the conflicting results in research, a number of researchers (e.g., Connor-Linton, 1995; Rao & Li, 2017; Shi, 2001) have suggested the use of an analytic scale to investigate how raters with different language backgrounds react to using more specific rating criteria. Kobayashi (1992) used a 10-point scale analytic rubric with four criteria: grammaticality, clarity of meaning, naturalness, and organization to ask NES and NNES raters to evaluate essays written by Japanese learners. He found that NES raters gave more positive evaluations on organization and clarity of meaning than Japanese-speaking groups (Kobayashi, 1992).

Lee (2009) investigated the rating behavior between NES and Korean English teachers using both holistic and analytic rating criteria and questionnaires. She found that Korean raters were more severe in features to do with sentence structure, organization, and grammar, while the NES group was more severe on content and the overall score.

Rao and Chen (2020) argue that most previous studies need further research. Even though they provide evidence of the impact of raters' linguistic backgrounds on their writing evaluation perceptions, most previous studies have used only a quantitative rating protocol. The use of a holistic scale and analytical reasons for the holistic scores given would be more useful. Thus, the collection of data for quantitative and qualitative analysis would provide insights into raters' rationale for how they allocate scores.

2.8 Japanese context

In the Japanese education system, school levels are divided into primary (6-12 years old), junior high school (13-15), high school (16-18), and university. Primary and junior high school are mandatory and entrance examinations are conducted for all high schools and universities regardless of public or private status. The study of EFL is mandatory from the third year of primary (age eight or nine) to the third year of junior high school (age fourteen or fifteen), based on the national curriculum created by the Ministry of Education (MEXT, 2019).

In 1987, the Japanese government began to hire NES teachers as language assistants for public schools (Sugimoto & Yamamoto, 2019). Currently, almost all public and private institutions from primary schools to universities, carry out team-teaching with NES and NJS English teachers. The national entrance examination for universities does not currently evaluate writing performance. However, in preparation for a major change in the 2021 national university entrance exam, the government sought to adopt speaking and writing tests

(Nakatani, 2019). This plan has been postponed due to various practical issues to do with the introduction of writing and speaking tests (Butler et al., 2020) and is yet to be implemented. Preparing students to pass the national university entrance examination is one of the main goals for high school teachers who teach English. English teachers tend to focus on teaching reading and listening (Bailey, 2018) because they are measured in the national entrance examination, so the teaching of English writing tends to be neglected. As a result, writing instruction and evaluation have not been developed well in Japan. This is the washback effect of the university entrance examination on teaching pedagogy (Sudo, 2020). Due to the lack of knowledge about writing assessments, many teachers avoid evaluating students' writing performance directly (Numata, 2006). According to Kowata (2015), 30 % of teachers do not assess writing skills even though they teach English writing in high schools. In addition, a limited number of studies have examined Japanese high school teachers' rating behavior for English writing, although it is acknowledged that writing ability and assessments are becoming more important for L2 learners. Due to this gap in knowledge about the evaluation of writing assessments in Japan, the current study aims to better understand Japanese teachers' rating tendencies and perceptions toward writing assessments. By comparing professional raters' writing evaluations, limitations and problems for Japanese teachers may be revealed. Thus, this study aims to contribute to the adaptation of writing tests and large-scale English tests in Japan.

2.9 Summary

This chapter has provided an overview of the important factors that may affect writing assessments. Certainly, rater variance and different rating scales might affect raters' judgment for writing assessments (Knoch et al., 2021). Especially, cultural and linguistic backgrounds might threaten the validity and reliability of writing assessments. Many studies have been conducted to investigate how raters from different linguistic backgrounds differ in their rating of L2 writing (e.g., Shi, 2001; Lee, 2009). These studies have found some differences between NES and NNES raters. In the EFL context, teachers require cooperation for teaching and assessment with teachers who have proficient English language ability. This study aims to demonstrate the difficulties that arise when raters from different backgrounds evaluate essays holistically. This consideration could be particularly important for policymakers, language curriculum developers, and individual teachers in EFL settings where highly proficient English speakers and EFL teachers need to work together in their teaching activities. However, little research has been conducted to investigate raters' perceptions of writing assessment from a high school teacher's viewpoint. This research aims to investigate this issue by focusing on the Japanese context. It investigates high school teachers' ratings and trained Aptis raters using Aptis for Teen's writing written by Japanese learners of English. Therefore, the questions that this research project aims to answer are as follows:

Research Questions:

1. To what extent do trained Aptis raters and Japanese teachers differ in their severity and consistency during unguided holistic rating of Aptis writing performance?
2. How do trained Aptis raters and Japanese teachers differ in their qualitative judgment and analytical reasons for their holistic rating of Aptis writing performance?

3. Methodology

3.1 Introduction

This chapter illustrates how the study addresses the two research questions presented at the end of the previous chapter. A mixed method approach was adopted. Data were gathered for quantitative analysis using scores allocated by the different rater groups and their answers to a post-marking questionnaire. The latter was also used to gather data for qualitative analysis along with semi-structured interviews.

The following sections present the design approach which provides the rationale for choosing a mixed method approach. The three instruments of data collection are described in detail and a brief description is given of the participants and the ethical procedures carried out. Finally, the methods of data analysis are illustrated to show how data were organized for analysis and presentation which follows in Chapter 4.

3.2 Design Approach

The decision to employ mixed method research is based on previous studies about the evaluation of writing that have inspired this study. Mixed method research further develops investigations in the field of applied linguistics research (Dörnyei, 2007); it allows the researcher to verify findings by looking at them from different perspectives to gain a greater understanding of the findings.

This mixed approach allows the study to investigate deeply how Japanese high school teachers (henceforth JTs) and trained Aptis test raters (henceforth ARs) differ when allocating the overall scores for the essays they mark, and to uncover their analytic reasons (via quantitative analysis) with retrospective interviews (via qualitative data) to allow the researcher to elicit the meanings behind their scoring and rating process more deeply.

An added advantage of a mixed method approach compared to a single method (quantitative or qualitative alone), is that it can compensate for the weakness of one approach with strengths of the other, thereby reducing the biases which may occur when using one approach (Paltridge & Phakiti, 2015). Certainly, as Paltridge and Phakiti (2015) point out, their advantages and disadvantages are to both qualitative and quantitative research. Quantitative research tends to be more objective, reliable, and replicable since it involves numbers and statistics. Meanwhile, it may not be representative of subjective aspects, it cannot evaluate individual perspectives and opinions because it works with the averages of the samples. Even though qualitative research requires a small sample size due to the time-consuming process of data analysis, it may counter the above disadvantage of quantitative research by providing data on the participants' experiences, thoughts, and knowledge. In this study, while the overall score awarded to each essay might reveal how two groups of raters differ when scoring essays, individual interviews provide a deeper understanding as to why they differ, or which specific sentence may have influenced their overall score. This triangulated process can provide more valid and reliable results (Dörnyei, 2007).

Thus, to answer the research questions, the following data collection instruments were used : evaluation sheet, a questionnaire, and a semi-structured interview. The holistic score and the reasons for the holistic score were collected through an evaluation sheet in order to analyze the severity and consistency of each group's score and to identify on which aspect each group focuses. The questionnaire was designed to collect raters' background information and their perspectives on their writing assessment. The semi-structured interview was designed to elicit deeper insights into the reasons for the holistic score and process of rating. More detailed descriptions of the research instruments used in this study are given in the following sections.

3.3 The data collection instruments

This study employs three research instruments which are described in the following sections. Table 1 shows how this research was operationalized to answer the research questions.

The evaluation sheet aimed to collect two data: holistic score of 20 essays and reasons for raters' holistic scores. These led to the foundation of this study and are explained in Section 3.3.1. A 10-point scale holistic score allocated to each essay by ARs and JTs, using unguided criteria, provided the necessary data. These scores revealed how the two groups differed in their severity and consistency. Reasons for their holistic scores for each essay identified how the two groups differ in their focus when they evaluated essays.

In addition, a questionnaire was completed by the participants after they rated the 20 essays. Its aim was to collect information about the participants' backgrounds and their attitudes toward writing assessments. Semi-structured interviews aimed to elicit information about the evaluation process and to identify why the two groups differ. Data collection was conducted in English for the AR group, but in Japanese for the JT group as this is the first language (L1) of the researcher and that group. This helped the participants understand the instructions and answer the questionnaire and interview more deeply.

Table 1: Operationalization of this study

Research Questions	**Data collection**	**Type of data and analysis**
RQ1 To what extent do trained Aptis raters and Japanese teachers differ in their severity and consistency during	Holistic scoring of 20 essays	Quantitative: descriptive inferential statistics

unguided holistic rating of Aptis writing performance?		
RQ2 How do trained Aptis raters and Japanese teachers differ in their qualitative judgment and analytical reasons for their holistic rating of Aptis writing performance?	Reasons for the score (Adapted from Shi (2001))	Quantitative: frequency
	Rater's rank order of features for the rank-type questions (Adapted from Lee (2009))	Quantitative: descriptive
	Semi-structured interviews with the raters	Qualitative: Thematic analysis

3.3.1 Evaluation sheet

This research used the Aptis for Teen's Part Four essay which requires the ability to write a for and against style essay of 220-250 words. The Aptis Testing System is an English language proficiency test developed by the British Council (British Council, 2022). The broad purpose of the tests is to examine the English language proficiency of users of ESL/EFL based on the scale of the Common European Framework of Reference for Languages (CEFR) (North, 2000). From the Aptis for Teen's data warehouse of the British council, 23 essays, written by Japanese learners and officially scored by British Council raters, were retrieved. 20 essays were used for the live research and three for the rater training. The Aptis test takers were given 40 minutes to answer the four tasks in the writing section. Essays were written by test-takers who scored 26-40 overall in the writing section, this represents level B of the CEFR. The purpose of choosing this level is twofold: (1) to ensure the essays are of sufficient length to evaluate both language and content, and (2) because helping students to achieve B level is the educational goal of high school teachers in Japan.

The chosen essays were typed on a Microsoft Forms evaluation sheet to guarantee uniformity in appearance and were presented to participant raters as originally written, including all errors. The task prompt for the essay was:

> **"Every month we run a competition on our website. Why not enter? You might win one of our fabulous prizes! The theme this month is Sport.**
>
> **Write your argument in response to this statement:**
>
> **International sports competitions such as the Olympics help to bring countries together.**
>
> **Remember to include an introduction and a conclusion.**
>
> **Write your competition entry below in 220-250 words."**

Prior to the data collection, three essays were chosen for the rater training and given a benchmark score by the researcher and applied to each essay's overall score. A one-hour rater training session was held for each group of raters via Microsoft Teams and Zoom in their L1, Japanese, or English. During the rater training, the raters were asked to give a score to each of the three essays, using a 10-point scale (where 10 is the highest and 1 is the lowest), to represent overall performance, and then to state the three main reasons for the rating in order of importance (see Appendices 3 and 4). Once raters submitted their training sheet online, the researcher provided the benchmark score for each essay, i.e., 2,5, and 8 out of 10. Such provision of level-specific essays was expected to be helpful for raters when interpreting band descriptors and follows both Lowie et al. (2010), and Weigle (2007).

The evaluation sheet for the rater training was adapted from Shi (2001), who argues that the predetermined evaluation criteria would restrict or mandate raters' judgments. Both Connor-Linton (1995), and Shi (2001) compared holistic scores and self-reported reasons from raters with different cultural backgrounds. However, their open-ended questions have one concern that would be difficult to code the unexpected reasons and ambiguous reasons raters may have given. For instance, if raters choose accuracy as a reason, it is not sure it could be related to grammar, sentence structure, or vocabulary. From this perspective, a checklist of reasons was compiled for the raters to choose from in the live study. Moreover, the rater has a different reason from those featured in the checklist, they can choose "other" and state this reason. The features on the checklist were drawn from a combination of Rao and Chen (2020) and the Aptis for Teens writing marking scale, which is a confidential document. It includes arguments, organization, grammar, vocabulary, and sentence structure. To revise this first version of the checklist, raters were asked to state three reasons for their holistic score during the rater training. That is, from the rater training, 180 reasons were collected and were revised, and categorized into the first checklist and, although most of them matched the features in the draft checklist, the reason categorized as "task achievement" was seen eighteen times in the pilot and, therefore, was added to the checklist for the live study. Even though the reason related to "length" was noted seven times for a 36 word-essay, no participant gave length as a reason for scores allocated to Essays 1 (77 words) or 2 (159 words). As a result, length was not included as a reason because the length of essays for the live study ranged from 73-229 words.

The rater's severity and consistency were explored by analyzing the holistic score given on the evaluation sheet (see Appendices 5 and 6). To examine an inter-rater reliability index, SPSS version 27 was employed to check the reliability coefficient. In terms of the analysis of the qualitative data, the frequencies of reasons for their scoring were computed to show the differences between the two groups.

3.3.2 Questionnaire design

Questionnaires are one of the major instruments used for data collection in the field of Applied Linguistics because they are multifaceted, they can create data relatively easily, and can collect different kinds of information in a short period of time (Dörnyei, 2007; Paltridge & Phakiti, 2015).

After considering its advantages, this research employed an online questionnaire. That is, because the study compared two groups, one in Japan and the other with members spread around the world. An online questionnaire would allow collection of data from much larger and more diverse populations than face-to-face interviews (Dörnyei & Taguchi, 2009). In addition, because it does not require any face-to-face interaction between participants and the researcher, participants are less likely to feel any pressure, this may enhance the level of honesty in their responses (Dörnyei, 2007).

The questionnaire contained questions about their background and their perceptions of the difficulty of grading various features when making a holistic decision; this question asked raters to rank features in order and by their preference.

The questionnaire was divided into four parts, with a total of 10 questions overall. One section comprised three open-ended questions that aimed to identify the participants' basic backgrounds. The second section aimed to investigate the participants' educational background, and the third section related to their professional backgrounds, such as their degree level and type, years of teaching experience, and amount of rater training received. The fourth part contained a question about the difficulty of grading the components when allocating holistic scores and followed Lee’ s (2009) instrument, which includes rank-order type questions about the important and difficult components of writing assessment. This information can explain the raters’ behavior with the importance of components in rating holistic scores revealed from the data from the evaluation sheet. The questionnaire can be seen in Appendices 7 (in English) and 8 (in Japanese). Meanwhile, Table 2 shows the organization of the questions.

Table 2: Organization of questions 1 to 10 from the questionnaire

No.	**Question type**	**Expected answers**	**Data**	**Purpose of collection**
Q1-4	Close and open-ended	Participants’ basic background	Quantitative Qualitative	Provide extra information
Q5-6	Close and open-ended	Participants’ educational background	Quantitative Qualitative	Provide extra information
Q7-9	Close	Participants’ professional background	Quantitative	Provide extra information
Q10	Rank order	Participants’ perceptions about the writing assessment	Quantitative	Research question 2

Once the online questionnaire was completed, it was uploaded to the Microsoft Forms website and was sent for pilot tests via e-mail. The decision to use Microsoft Forms was taken due to its accessibility for the data collection and its familiarity with both the researcher and participants. The majority of the participants use it for business purposes. After piloting, minor changes were made in terms of the instructions given (they were made more precise), and the presentation of the questions. Furthermore, the information details which explained the research project were followed by the method of consent. That is, it was stated explicitly in the introduction that: “By completing and submitting this online questionnaire I understand that I am giving consent for my answers to be used for the purposes of this research project”. Ultimately, the final version of the questionnaire was uploaded to Microsoft Forms; a link sent to participants at the beginning of June was available for them to complete the questionnaire within seven days.

3.3.3 Semi-structured interview

In addition to the data collected through a post-marking questionnaire, individual interviews were conducted to allow the participants to express themselves freely by asking open-ended questions (Paltridge & Phakiti, 2015). Furthermore, interviews can elicit more personalized thoughts and answers about their evaluations. In mixed-method studies such as this, quantitative analyses of the questionnaire are supplemented by interviewing a subset of respondents to investigate the issues raised by the findings more deeply (Polio & Friedman, 2016). With the data collected from all raters, individual semi-structured interviews were conducted with three raters from each group. Semi-structured interviews are flexible due to their compromise between open and structured interviews (Dörnyei, 2007). Truly open questions can allow the participants to bring up points important to them and to pursue a line of interest that may result in deviation from the aim of the questions; thus, a prepared guide can help maintain consistency regarding the topic in question (Polio & Friedman, 2016).

Interview schemes from Rao and Chen’s (2020) study in a Chinese context were adapted in this study.

On completion of the interview questions, three pilot interviews were conducted in both English and Japanese using Microsoft Teams and Zoom, with which both the researcher and the participants are familiar. In addition, Microsoft Teams allows the entire interview to be recorded and transcribed in whichever language is used, in this case, English and Japanese. This allowed the researcher to concentrate on the interviews without concerns about note-taking.

After three pilot interviews, some ambiguous instructions were found and corrected. The order of the questions was prepared in advance as the first stage of the pilot interview; however, it was found that not preparing the order in which questions were asked allowed more flexibility in eliciting the opinions of the raters.

Then, interviews were conducted, in English for the ARs and in Japanese for JTs, approximately 15-30 minutes after the scoring procedure. The questions were mostly related to the writing features participants chose for their overall rating, focusing particularly on the results which stand out from others, such as where there is a large difference between the two groups, or where one rater gives a different score from another rater (see Appendix 9 for example of interview questions).

The role of the interview was to identify the differences between the two groups’ evaluations and elicit the process they apply when scoring the essays. In the process of the thematic analysis of the semi-structured interview, firstly, coding schemes were discussed, and the categorization of the data was decided upon (see Appendix 10). Then, the interview transcripts were read repeatedly, and salient comments related to the criteria for their scoring were marked. Once the coding system was finalized, the interview data were coded using Nvivo (see Appendix 11).

3.4 The participants

Empirical research has compared the differences between native and non-native-speaking raters' perspectives on English writing assessment (Rao & Chen, 2020; Shi, 2001). This study also compared the participant raters' perspectives between the two groups. The first group comprised 10 ARs working for the British Council with ages ranging from 37-59 years old. Trained ARs have high levels of English proficiency, the nationalities of the participating group are mostly NES, 4 British, 1 Australian, 1 South African, and 1 American. Nine (90%) raters have a Master's degree, but only six of these are related to language teaching or applied linguistics. The second group of participants is Japanese, they work as English language teachers at high schools in various Japanese districts. All have undergraduate degrees with English language-related majors, e.g., linguistics, English literature, and education, but only 2 (20%) have a Master's degree (see figure2).

Appendix 12 provides more details, including that 6 teachers were male and 4 female. All 10 ARs have received rater training, whereas eight (80%) of the JTs stated that they had received no such training (see figure 3). All but one participant (a JT) had a minimum of five years of experience teaching English (see figure 4).

Of the 20 raters, one (5%) had taught English for more than five years, 10 (50%) for more than 10 years, six (30%) for more than 20 years, and two (10%) for more than 30 years. In general, it is clear that the ARs are more highly educated and have more teaching experience than the JT group. The following figures provide a visual representation of these numbers.

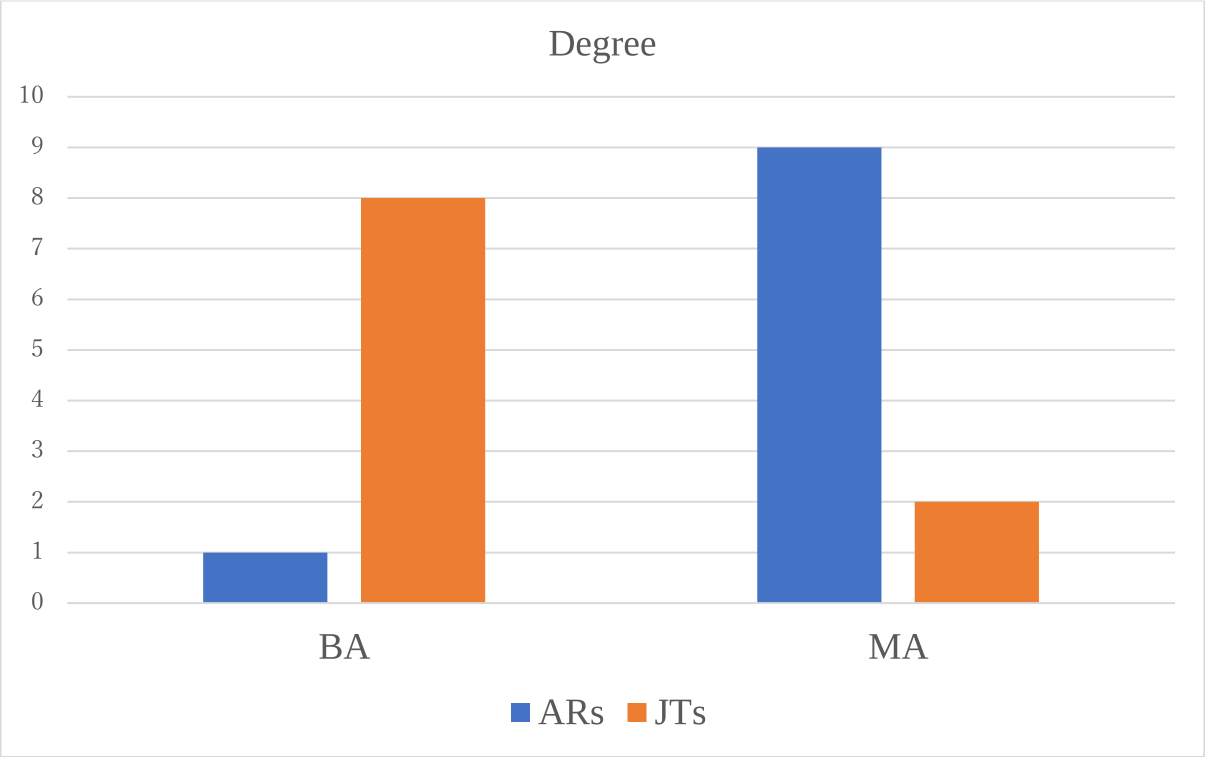


Figure 2: Degree of participants

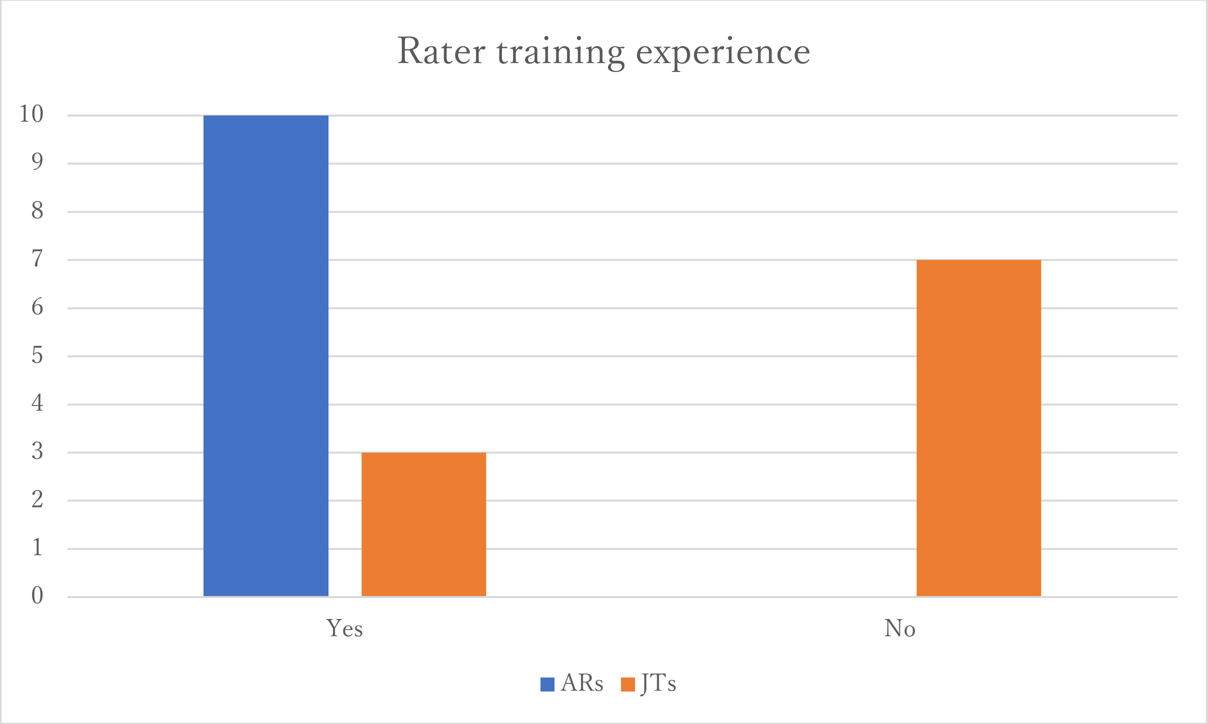


Figure 3: Rater training experience

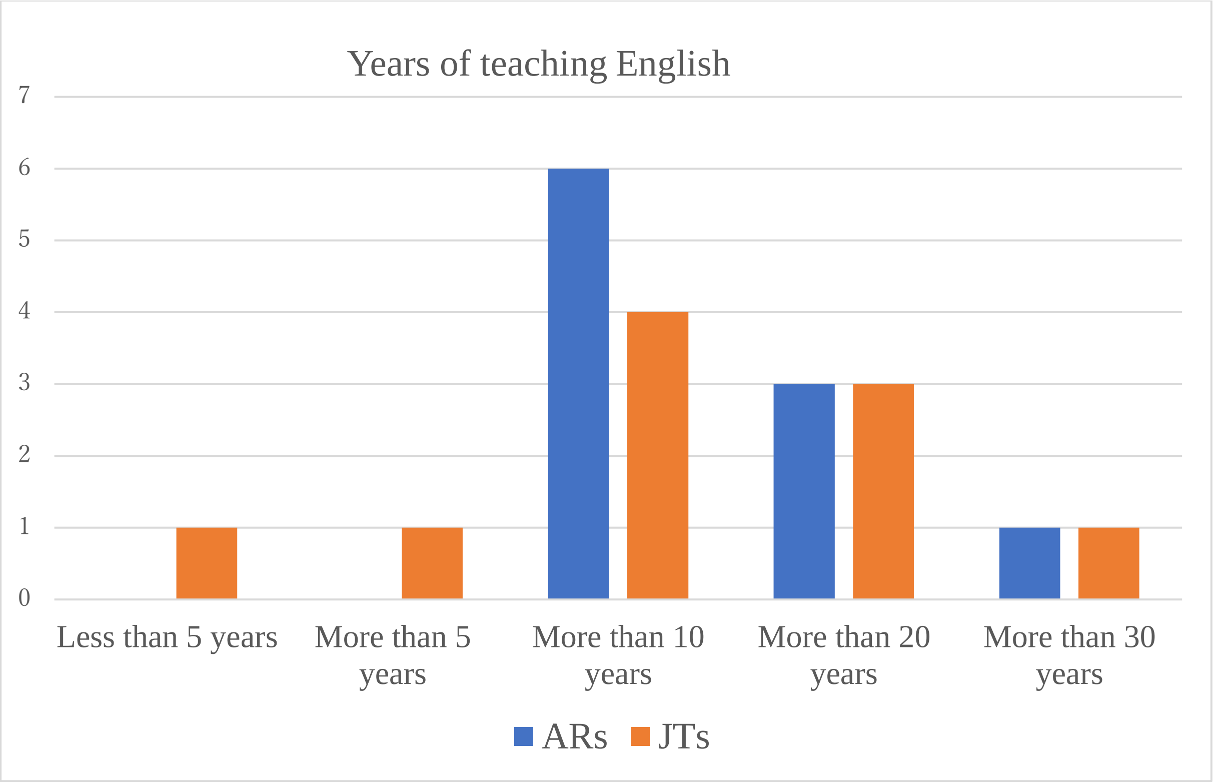


Figure 4: Years of teaching English

3.4.1 Interview participants

Three participants were chosen for semi-structured interviews from each group for a total of six interviews. Similarly to Rao and Chen (2020), this study followed the maximum variation sampling proposed by Patton (1990) to select these interviewees. This type of sampling is a method that allows the researcher to collect data from a maximum variation in the participants' genders, ages, years of teaching, and nationalities. First, all participants' background information was collected by questionnaire and their evaluations of 20 essays was analyzed, then interviewees were chosen to represent the 20 participants. Tables 3 and 4 provide details for these 6 participants (3 ARs and 3 JTs) who agreed to take part in the semi-structured interviews.

Table 3: Details about the interviewees from the AR group

Pseudonym	Gender	Age	Years of teaching experience	Nationality
AR-A	Male	59	More than 30 years	British
AR-B	Male	43	More than 20 years	American & German
AR-C	Female	37	More than 10 years	Tanzanian

Table 4: Details about the interviewees from the JT group

Pseudonym	Gender	Age	Years of teaching experience	Nationality
JT-1	Male	55	More than 30 years	Japanese
JT-2	Female	34	More than 10 years	Japanese
JT-3	Male	27	More than 5 years	Japanese

3.5 Ethical considerations

As this project involved human participants, an ethics consent form was required to be approved by the Department of English Language and Applied Linguistics Ethics Committee. The Ethics form, the information sheet (see Appendices 13-16), a description of the project, and other requested materials were submitted to the committee at the end of April 2022. The approved consent form and information sheet stated explicitly the purpose of the research and the instructions associated with the questionnaire, the interview, and the essay evaluation. Participants were assured that their data would remain anonymous by the use of pseudonyms and that they withdraw themselves and their data from the study at any time. An information sheet illustrated the aim of the project and the researcher and supervisor's contact. All documents were submitted to the Japanese participants in Japanese but are presented in English in this report.

3.6 Data Analysis

Once all data were collected, both from the essay evaluations and the questionnaires, the quantitative data were transferred from Microsoft Forms to Excel and SPSS, where they were analyzed to answer Research Questions1. Then, the qualitative data derived from the questionnaire were analyzed using Excel, and the qualitative data derived from the semi-structured interviews were analyzed thematically using Nvivo. Before the analysis, the interview data from the JTs group were first transcribed in Japanese by Microsoft Teams, then translated into English for coding by the researcher. The data from open-ended questions about the essay evaluations and the questionnaires were also translated from Japanese to English for analysis.

It is important to note that the qualitative data were analyzed by the researcher as this process requires personal interpretation of the answers. To support the subjectivity of the qualitative inquiry, data were compared with the quantitative results to validate the analysis.

This chapter has described the methodology applied to the study. The following chapter reports the results as they address the research questions seen at the end of Chapter 2.

4. Results

4.1 Introduction

This chapter presents the results of the data analysis. Firstly, differences in the consistency and severity of holistic scores between ARs and JTs were examined. Secondly, qualitative data collected from questionnaires and interviewees were used to identify differences in the rating process and the emphasis placed on various elements. This chapter is organized around the research questions, as seen in Table 5, below.

Table 5: Breakdown of data used to answer research questions

Research questions	Source of data and data analysis
RQ1: Consistency and severity of the holistic scores	400 Raters' holistic scores - 20 essays rated by 10 ARs and 10 JTs
RQ2: Analytical reasons and qualitative judgments for the holistic scores	1200 rating reasons - Three reasons for each essay - 20 essays rated by 10 ARs and 10 JTs 20 Questionnaires - 10 questions in total - 10 ARs and 10 JTs 6 Semi-structured interviews - 3 ARs and 3 JTs

The quantitative data were analyzed using SPSS and Excel; the qualitative data were coded and analyzed using Nvivo. Due to the small number of raters (N=20) and interviewees (N=6), frequency is used to show the findings of the qualitative analysis.

4.2 Research Question 1

To what extent do trained Aptis raters and Japanese teachers differ in their severity and consistency during unguided holistic rating of Aptis writing performance?

4.2.1 Reliability

In order to compare consistency among the scores awarded by ARs and JTs when rating the same scripts, inter-rater reliability was estimated using Cronbach's coefficient alpha (Bachman, 2004) which indicates reliablitity of the measurements used. The coefficient score varies between 0 and 1, where 0 represents no consistency and 1 the highest level of consistency, or reliability. The results for the two groups indicate ratings made by the JTs show greater reliability (coefficient α=0.93) than the ARs (coefficient α=0.89). This indicates that the group of JTs were more consistent in their rating than the ARs when assess the 20 Aptis essays involved in the study's rating process.

4.2.2 Comparison of scores

Table 6 summarizes the means and standard deviation for the holistic scores of the two groups for the 20 essays. To examine the severity and consistency of the holistic scores, the mean, standard deviation, and rank order of the means were compared. Except for Essays 10 and 17, the difference between the two groups' respective essays is within 1.3 and the rank ordering of the 20 essays correlates highly (r =.85). The overall agreement of the ARs and

JTs' ratings of the 20 essays is the most striking result in this comparison. This similarity may in part be due to the small range of rating scales (1-10), limiting the expression of distinction.

However, there are some differences in their ratings. While the group means score overall essays rated by JTs was 5.3, for the AR it was 5.5 It seems that JTs gave a wider range of scores to the best five, from 8.9 to 6.3, while the ARs scores ranged from 7.7 to 7.1. Even though both groups agreed that Essay 5 rated most highly, and Essay 14 rated the lowest among 20 essays and selected the same five essays to represent their top five, the rest of the essays experienced a somewhat greater difference in their ranking between the ARs and JTs. For illustration, JTs rated Essay 2 13th, and ARs 6th, while JTs rated Essay 10 17th and ARs 10th. The possible reasons for these differences are discussed in depth in the next chapter.

Although the coefficient shows that individual raters in the JT group are slightly more reliable than AR group, the standard deviation does not support this. Rather, it indicates that the group of ARs maintains higher reliability (SD=1.80) than the JT group (SD=2.36). Scores provided by the JTs were spread more widely than those provided by ARs; e.g., the JTs gave Essay 14 the lowest score with the least spread at 0.88, while other essays had a spread of over 1.20. This means all JTs agreed that this essay should be rated the lowest.

In terms of means, in total 11 essays were scored higher by the ARs and in total 7 essays were scored higher by the JTs; differences in the means between the two groups vary from -1.2 to +1.8. There is a distinct disagreement in the rating for Essay 10, revealing the largest difference between the rater groups, i.e., the means for the ARs is 5.3 and JTs, 3.5.

However, to reveal differences between the two groups' ratings, a t-test was run. No significant difference in holistic scores awarded by the JTs (M= 5.5, SD=2.36) or the ARs (M=5.3, SD=1.8; t (398) =8.32,p=.4, two-tailed) was found.

Table 6: Comparisons of ARs and JTs' holistic scores on the 20 essays

	ARs (n=10)		JTs (n=10)		Ranking of essays	
Essay	Mean	SD	Mean	SD	ARs	JTs
1	4.1	1.37	4.8	1.69	18	13
2	4.9	1.29	5.8	2.15	13	6
3	4.7	0.95	5.4	1.78	15	10
4	5.7	0.95	5.4	2.07	8	10
5	7.7	1.49	8.9	1.20	1	1
6	7.2	1.03	7.6	1.78	4	2
7	5.6	1.07	5.6	2.22	9	9
8	6.1	0.88	5.8	1.55	7	6
9	4.5	0.85	4.4	2.72	16	15
10	5.3	1.34	3.5	2.12	10	17
11	3.3	0.95	2.6	1.17	19	19
12	4.4	1.26	3.4	1.26	17	18
13	7.4	0.70	7.6	1.51	3	2
14	2.2	1.14	2.1	0.88	20	20
15	6.8	0.92	5.7	1.64	6	8
16	4.9	1.20	4.6	1.35	13	14
17	5.2	0.79	3.8	1.23	12	16
18	7.1	1.45	7.2	1.87	5	4
19	7.6	0.97	6.3	1.83	2	5
20	5.3	1.25	5.3	1.49	10	12
GroupM	5.5	1.80	5.3	2.36		

4.3 Research Question 2

How do trained Aptis raters and Japanese teachers differ in their qualitative judgment and analytical reasons for their holistic rating of Aptis writing performance?

4.3.1 Reasons for holistic rating

To identify the differences between the two groups’ emphasis on writing when scoring, the three reasons for scoring in order of importance were first examined according to each and all reasons. Then, from the questionnaire and interview data, the reasons for their weighting and their perspectives on the evaluation were analyzed.

Figures 5 and 6 show the six criteria most frequently mentioned as the groups first, second, and third reasons for giving the scores allocated, and the level of frequency of those reasons. In total, the evaluation of 20 essays by 20 raters generated approximately 1200 reasons for their holistic rating. As in previous studies (Shi, 2001) (see Section 2.7.2), the elements of writing in this study were analyzed separately in terms of language use (grammar, vocabulary, and sentence structure) and content (task achievement, organization, and argument). As can be seen from Table 7, the greatest proportion of reasons for the score given by the ARs related to grammar (21.6%), followed by sentence structure (14.9%) and vocabulary (12.6%), each of which related to language use, totaling 49.1%. It is interesting to note that the JTs show a somewhat different trend with less focus on language use, totaling 23% (see Figure 6).

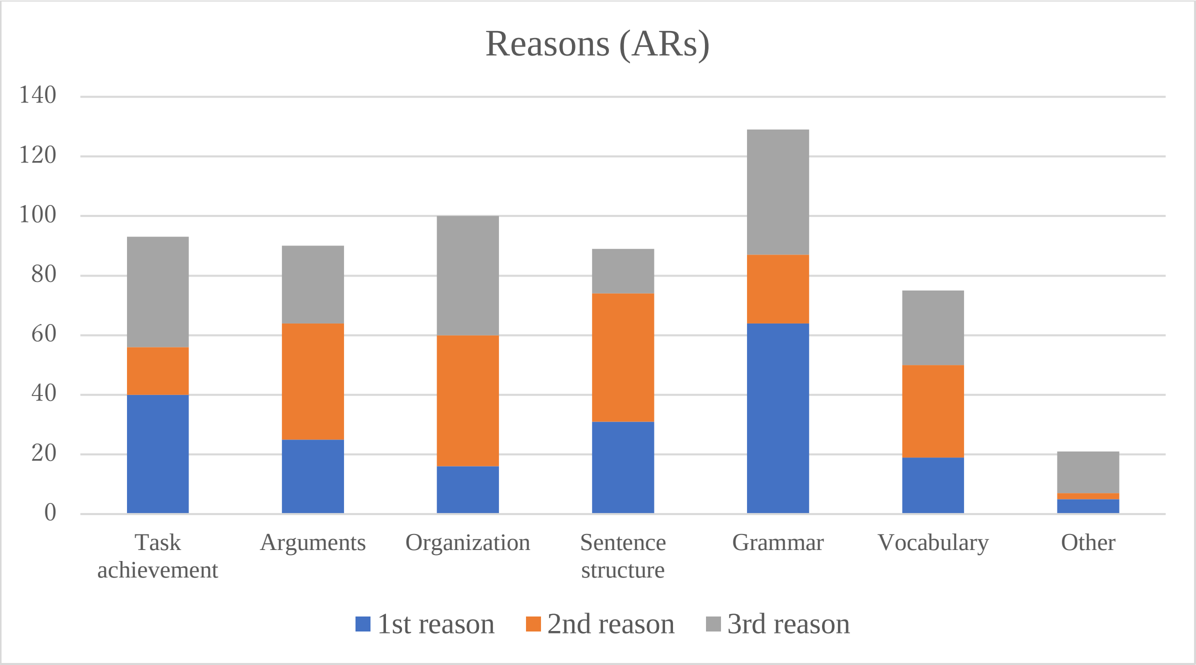


Figure 5: Reasons for holistic scores (ARs)

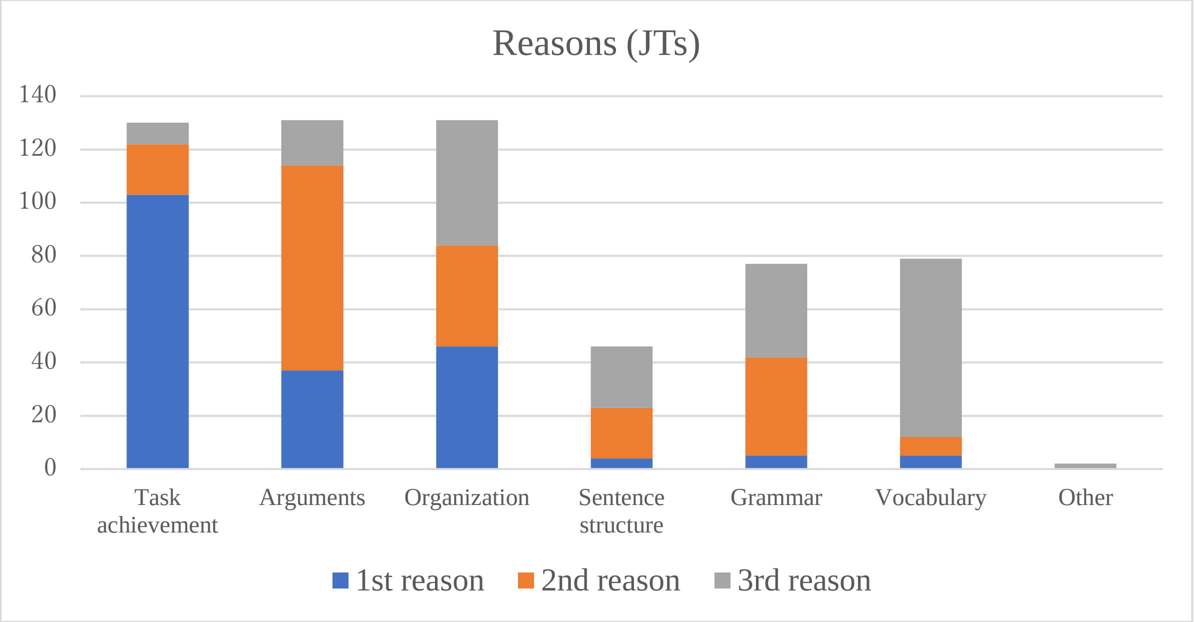


Figure 6: Reasons for holistic scores (JTs)

As can be seen in Table 7 below, the most striking difference among the reasons given is the rank order for grammar, the more frequent reason for ARs but the second least for JTs. Moreover, JTs chose grammar (12.9%) and sentence structure (7.7%) about half as many times as the ARs and paid more attention to the content (65.8%) than language use (32.9%).

They paid most attention to organization (22.0%) and argument (22.0%), followed by task achievement (21.8%). Compared to the JTs, the ARs chose each reason at about the same rate, and there appears to be no difference between their ratings of the content and the language elements used.

Table 7: Reasons for holistic scores (ranking and percentage)

Rank	ARs	%	Rank	JTs	%
All reasons					
1	Grammar	21.6	1	Organization	22.0
2	Organization	16.8	1=	Argument	22.0
3	Task achievement	15.6	3	Task achievement	21.8
4	Argument	15.1	4	Vocabulary	13.3
5	Sentence structure	14.9	5	Grammar	12.9
6	Vocabulary	12.6	6	Sentence structure	7.7
7	Other	3.5	7	Other	0.3
First reasons					
1	Grammar	32.0	1	Task achievement	51.5
2	Task achievement	20.0	2	Organization	23.0
3	Sentence structure	15.5	3	Argument	18.5
4	Argument	12.5	4	Grammar	2.5
5	Vocabulary	9.5	4=	Vocabulary	2.5
6	Organization	8.0	6	Sentence structure	2.0
7	Other	2.5	7	Other	0.0
Second reasons					
1	Organization	22.2	1	Argument	39.1
2	Sentence structure	21.7	2	Organization	19.3
3	Argument	19.7	3	Grammar	18.8
4	Vocabulary	15.7	4	Task achievement	9.6
5	Grammar	11.6	4=	Sentence structure	9.6
6	Task achievement	8.1	6	Vocabulary	3.6
7	Other	1.0	7	Other	0.0
Third reasons					
1	Grammar	21.1	1	Vocabulary	33.7
2	Organization	20.1	2	Organization	23.6
3	Task achievement	18.6	3	Grammar	17.6
4	Arguments	13.1	4	Sentence structure	11.6
5	Vocabulary	12.6	5	Arguments	8.5
6	Sentence structure	7.5	6	Task achievement	4.0
7	Other	7.0	7	Other	0.0

A significantly different picture from that shown in Figures 5 and 6, is revealed in the further analysis of reasons given between the two raters groups in terms of the ordering of their three reasons for each of the 20 essays. Raters were asked to expand on their reasons for allocating specific scores to better understand how they weighted different writing components in their evaluation.

The most striking difference for the first reason given is that the JTs put little emphasis on language use: vocabulary (2.5%), grammar (2.5%), and sentence structure (2%), while they paid much more attention to task achievement (51.5%), organization (23%), and argument (18.5%) (see Table 7). Data from the interviews supports these findings. All JTs reported that they first checked whether the content met the assignment requirements before evaluating other factors, such as language use. This is exemplified by JT-2:

> *I usually work from the point of view that no matter how beautiful the writing is or how solid the content is, if the student does not answer the question, he/she will not be able to get a good score.*

In contrast, the ARs tended to focus on language use, especially grammar (32%) and sentence structure (15.5%), (see Table 7) and in their interviews they all reported that they focused on language use more than essay content. AR-A explained the reason why he emphasized language use:

> *I think you can't have the task achievement without the vocabulary, so that's I think in a way why I'd be focusing on the vocabulary and the grammar would be focusing more on the level.*

All ARs reported that their evaluation processes differ depending on the test – they all work as raters for different examinations. Compared to academic essays assessment, Aptis aims to measure general writing ability and all three AR interviewees indicated that this is the reason they tend to focus more on language use than on task achievement, i.e., it is a test of language ability.

Figures 5 and 6 illustrate a change of tendencies in the reason ranked second in importance. JTs remained focused on the content of the essays, and the ARs displayed a balanced focus on both language use and content. As Table 7 shows, the JTs turned their attention to organization (19.3%) and arguments (39.1%), surprisingly around 70 % of reasons ranked as second in importance correspond to the content category.

In terms of the reason ranked third in importance (see Table 7), the majority of the JTs shift focus from content to language use at which time vocabulary becomes the most frequent reason given (33.7%), with grammar (17.6%) and sentence structure (11.6%) relatively high among the six components. It is important to note that this shift in focus suggests that these raters consciously place different weightings on writing components. The ARs also chose grammar (21.1%) as the third reason most often but moved from language use to content (totaling 51.8%).

Previous research asked raters to state three reasons for allocating their scores in the way they do, regardless of the importance they place on them (Connor-Linton, 1995). When considering all the reasons given (see Appendix 17), there is a clear difference when each group’s first reasons are compared. The research method used in this study allows representation of how participants would express their process more precisely if they were asked to state their reasons in order of importance, as in Shi’s (2001) study (see Section 2.7.2).

4.3.2 Rater perceptions of grading analytic features

Question11 in the post-marking questionnaire asked all participants (AR=10, JT=10) for their perceptions of analytical features which might have influenced their marking method. Their answers to rank-order type questions were analyzed. Table 8 reveals the most difficult feature and Table 9 is the least difficult. Further, Appendix 18 shows how each participant ranked the writing features in the terms of the difficulty of holistic evaluation.

Table 8: The most difficult criterion

Criterion	ARs (n10)	JTs (n10)
Sentence structure	3	2
Argument	3	3
Vocabulary	2	0
Grammar	1	2
Organization	1	0
Task achievement	0	3
Total	10	10

Table 9: The least difficult criterion

Criterion	ARs (n10)	JTs (n10)
Task achievement	4	6
Grammar	4	3
Organization	1	1
Sentence structure	0	0
Vocabulary	1	0
Argument	0	0
Total	10	10

As can be seen in Table 8, no JTs and only one AR perceived organization as the most difficult feature. Interestingly, both groups placed more importance on organization than other features, although both groups commonly perceived organization as a less difficult feature when deciding on an overall score. The statistical results are consistent with the interviewees' comments. In terms of essay organization, the majority of raters work as English language teachers and report familiarity with the basic organizational style of English essays, i.e., introduction, body, and conclusion.

AR-B perceives his evaluation process to be influenced significantly by his working experience as an academic writing teacher. In the evaluation of an academic essay, he focuses on the organization. Furthermore, it is clear that JTs have more confidence in their evaluation of organization than in content. JT-3 stated that he could not judge which argument is better. He claims:

> *I don't think I'm qualified to judge whether I'm doing well or not at all in the argument. I thought I would be able to judge my ability in terms of grammar, vocabulary, and organization based on my past studies.*

This confidence might affect raters' behaviors. In terms of evaluation of essay organization, two out of three JTs explained this component in more detail than the other components. They specified not only the importance of general structure but also details such as "spaces" or "indentation of paragraphs". JT-1 stated:

> *The paragraphs were not perfect. The graphs are composed with space between the lines, but I think the graphs start too far to the left.*

Similarly, JT-2 stated:

I think the paragraphs were not indented or something like that. I did not find it easy to read.

Four out of the 10 ARs and six out of the 10 JTs thought task achievement was the least difficult for determining an overall score. Figure 6 shows, that the majority of Japanese raters thought that task achievement was one of the most important reasons for their holistic scoring, and Table 9 shows that they thought it was one of the easiest to grade. In terms of vocabulary and sentence structure, some raters from both groups perceived it as a difficult criterion and others thought it was easy.

The semi-structured interviews revealed some striking differences in the evaluation process between the two groups. The salient finding relates to the rating processes coded "can do" and "reduction and adding". When participants decide on the overall quality of an essay, the ARs determine the level at which – the learners "can do", according to their internal scoring criteria; while JTs added, or subtracted points based on the learners' strengths and weaknesses. According to the interviews, all JTs often specify the areas where they deduct points when stating the reason for the score they allocated. Surprisingly, this was also true for the high-scoring essays.

JT-1 reflected on his evaluation process by using an adding and reducing evaluation style:

> *Even if the same rubric is used, the scoring for grammar, vocabulary, and discussion is negative for oddities and mistakes. On the other hand, I have the impression that achievement and other items are graded with points being added concerning other items.*

JT-3 explained the evaluation process for an essay in which he scored 9 out of 10.

> *I think the one point I took away from this essay is that "a international" is a mistake for "an international" and it is unclear where the flags in the third line modifies it.*

In contrast to this reducing and adding process, the ARs included mention of the positive elements of the essay when they describe the reason for the mark allocated.

AR-C stated:

> *I decided if I'm gonna uh and also we had to give reasons why we gave that mark, so I had to, for me how I interpreted it is I chose three things which are positive, the most positive things in the essay.*

AR-A reflected on his evaluation process, claiming:

> *I think now my focus more is on the actual trying to define the level rather than things like task achievement.*

Another interesting discrepancy in their perspectives on the evaluation of language use in the essays between the ARs and JTs concerns their focus. The latter are more likely to focus on accuracy, while the former is equipped with a wider scope of evaluation criteria, not only accuracy but also other aspects, such as range and complexity. All JT interviewees reported grammatical and spelling errors and these findings suggest they were greatly influenced by grammatical, structural, and spelling errors (see Appendix 10).

This is demonstrated in comments from JT-3:

> *Grammatically, well, there are errors here and there. For example, in line 4. The first one, "sport and communication has" should be "have", "play very good" should be "very well", the spelling mistake in line 8, "I wass" has extra S.*

JTs-1 evaluated the linguistic details by focusing on students' errors in their essays, he said:

> *As for vocabulary, there are many errors in the use of "close" and other words, and the sentences are somewhat incomprehensible if spoken colloquially.*

In addition, it is interesting to note that the JTs also paid meticulous to more specific mistakes concerning such areas as writers' punctuation and space use. In contrast, ARs did not point out only these specific errors, rather they judged the language use from a wider perspective, including tense use and complexity. AR-B stated:

> *Some basic mistakes like subject-verb agreement and hopes. . . I think this writer doesn’t have full control of punctuation, but other than that I thought it's good to control tenses. There's uh model construction.*

AR-A reported:

> *There was some good sort of good mix of structures there sort of trying to use some more sort of complex stuff, not always successfully.*

5. Discussion

5.1 Introduction

In summary of the results presented in the previous chapter, the findings from the analysis of the holistic scores given to 20 essays rated by the ARs and the JTs shows that both rater groups achieved relatively high reliability when giving a holistic rating. Although the ARs assigned higher scores to the 20 essays than the JTs, no significant differences were found in the holistic scoring of the essays.

The following sections compare the qualitative judgment and analytical reasons for their ratings using data gathered from 20 essay evaluations, the post-marking questionnaire, and the semi-structured interviews. It was found that the JTs focused on task achievement when determining their holistic rating of the overall quality of essays, while the ARs firstly placed emphasis on language use and then moved their attention to language content.

In addition, the raters' perspectives concerning their approach to evaluation. Differed considerably. The ARs decided the score for the level of language use based on what the learners can do considering accuracy and complexity, while the JTs focused on accuracy and learners' errors, adjusting the mark accordingly, similar to negative marking.

These findings are explained in more detail in the following sections, drawing on examples from the data and considering each research question individually. The chapter concludes with the implications and limitations of this study.

5.2 Research Question 1

To what extent do trained Aptis raters and Japanese teachers differ in their severity and consistency during unguided holistic rating of Aptis writing performance?

The results show that the two groups of raters achieved reasonably high reliability, the means of the marks they awarded and their rankings are almost the same in the evaluation of the 20 essays. The difference in the inter-rater reliability between the ARs and the JTs’ holistic scores was not significant. Similarly, Connor-Linton (1995) reported the same levels of reliability (0.75) for the NES and NNES raters, while most of the previous research (described in Chapter 2) found NNES teachers to be less consistent than NES raters (e.g., Johnson & Lim, 2009; Lee, 2009; Marefat & Heydari, 2016; Rao & Chen, 2020; Shi, 2001). This study illustrates that the two groups achieved fairly high reliability, despite not being provided with any predetermined evaluation criteria. All ARs are professionally trained and have some work experience as Aptis test raters, however, the majority of the JTs (70%) have never been professionally trained as raters. Before commencing the marking of the 20 essays for this study, a one-hour rater training session, using benchmark essays, which held online for participants; this may account for the high rater reliability. According to Shohamy et al. (1992), untrained raters given guidelines provided high-reliability coefficients of over 0.80 to 0.90. Shi (2001) suggests that NES raters use a wider range of average scores than NNES raters, ranging from 5.14 to 8.14 out of 10. In her study, she concluded that the NES took more risks giving a wide range of scores to make distinctions among the 10 essays assessed. This contrasts with the current study where the scores awarded by the JT group were more broadly spread than those awarded by the AR group; however, individual members within the former were more consistent, as seen in Section 4.2.2. Furthermore, this study revealed that the average scores were much wider than Shi (2001) reported, ranging from 3.2 to 8.5 given by the JTs and 2.2 to 7.4 given by the ARs. This could be due to the possibly more varied levels of essays used in this study.

5.3 Research Question 2

How do trained Aptis raters and Japanese teachers differ in their qualitative judgment and analytical reasons for their holistic rating of Aptis writing performance？

In this section, the analysis of reasons raters gave for their holistic rating are discussed. This includes data collected about the rating process and raters' attitudes resulting from the post-marking questionnaires and the semi-structured interviews are discussed. The results of the analysis presented in Section 4.3.1 showed that the ARs paid more attention to language use than JTs who first considered content, and then shifted their focus to language use.

This is in contrast to previous studies which suggest that NES teachers tend to focus on the quality of essays and their content, rather than on language use (see Section 2.7.1). However, the findings show that JTs seem to pay less attention to the language quality of the essays than ARs and this supports the results of previous research (Brown, 1991; Shi, 2001). Exploring the differences between NES and NNES teachers' ratings of native and NNES students' essays, Brown (1991) found that NES teachers focused more on sentence-level features, while NNES teachers focused more on the structure. Similarly, NNES gave fewer comments on language use than NES while NNES gave more comments on the organization than NES (Shi, 2001).

Firstly, this difference between the two groups may be a result of differences in their understanding of the purpose of the Aptis essay test. The JTs regarded the essays they were asked to mark as a single text on a single given topic, and therefore placed more importance on task achievement and organization. On the other hand, the ARs saw the essays as tests of language use. An analysis of Essay 10 found it weak in content, it was incomplete and the argument was unreasonable but strong in language use. To illustrate this point, Essay 10 was compared with Essay 9 to identify why the holistic ratings differed significantly between the

ARs and the JTs (Essays 9 and 10 can be seen in Appendix 19). According to Text Inspector (Bax, 2012), Essay 10 is less lexically diverse than Essay 9, while Essay 9 can be said to demonstrate better language use. Essay 9 is completed, following the basic essay format, while Essay 10 is incomplete, the writing stops in the middle of the body section of the writing. The two groups differ in their evaluations of these two essays; the ARs scored Essay 10 higher than Essay 9, while the JTs scored it lower than Essay 9. Even though Essay 9 is much shorter than Essay 10, it follows the standard essay structure which includes an introduction, a body, and a conclusion, which JTs, in general, expect from their students. All the JT interviewees negatively pointed out that Essay 10 was incomplete. This indicates that their judgment may be influenced by the fact that they perceive task achievement to be the first element to apply to their evaluation. As their first reasons for their rating for Essay9, 9 JTs chose content components (task achievement, organization, and argument), and for Essay 10, 6 JTs chose content components (task achievement and organization). In contrast, more ARs focused on language use (grammar, vocabulary, and sentence structure) for Essays 9 (5) and 10 (6) than JTs. These reasons could explain the difference in the average scores allocated by the two groups for complete and incomplete essays.

The interview data provides a further explanation for the findings expressed in Table 7. JTs-1 reported that he gives priority to task achievement, no matter how good the content, he may not give a high score if the argument is not complete:

> *It's not completed. There is no conclusion. Organization is my first reason for rating this essay.*

Similarly, JTs-2 reported:

> *This essay did not complete, and this is the main reason to give it 4 out of 10.*

In contrast, it seems that incomplete work does not have a significant impact on ARs' decisions in their overall evaluations. According to the interviews, and supporting the empirical findings, the ARs focused more on language use (grammar range and accuracy), than on task achievement.

AR-A reported:

> *They saying we're going to give two reasons, but then only gave one reason. But there were some quite good constructions in there . . . There's an interesting vocabulary.*

As can be seen in Table 6, the average scores allocated to Essay 10 were more than two points apart between the JTs and the ARs. It has been reported above that the JTs score incomplete essays more severely than the ARs; the latter gave higher marks to incomplete essays that do not answer the task provided the language use is good. Repeatedly in the interviews, the ARs stated that the evaluation criteria would lead them to focus more on content when teaching academic writing. Consequently, it can be said that the JTs rate essays on a macro level, while ARs rate them on a micro level, i.e., language use.

The second point to make about the differences between the two groups of raters concerns their cultural and educational backgrounds. Japanese society, and thus schools emphasize the importance of following rules (Ninagawa, 2017). In Japan, it is common knowledge that no matter how good your ability is, you will not receive a high score if you deviate from what you are instructed to do, following instructions is very important for Japanese people in general. In his interview, JT-3 explained his instructional goal and the reason for it, saying:

> *I often witnessed students giving different answers to questions and was aware of the more serious problem. One of my educational goals is to be able to answer a given question in a confrontational way.*

The instructions participants followed for the essays used in this study included: “Remember to include an introduction and a conclusion”. This instruction may have influenced the JTs’ evaluations given that many of them pointed to organization (22%) and task achievement (21.8%) as reasons for their ratings. Some may have reported negative task achievement as their reason for their evaluation of an incomplete essay like Essay 10, others may report organization as their reason because the essay does not include an introduction, a body, and a conclusion.

Furthermore, a third explanation for the differences between the groups concerns their professional background. The reason for the JTs focus on content is thought to be the result of the recent shift in Japan from grammar translation and reading methods to Communicative Language Teaching (CLT) methodology (Abe, 2013; Kitahara, 2010). Since 2003, the national curriculum has been revised to encourage CLT and communicative skills in English to nurture Japanese who can use English (MEXT, 2003) and teachers' instructional focus has gradually changed from language use to content. With the revision of the national curriculum, the number of textbooks approved by the government increased from reading-based ones to those with a focus on communicative activities. In schools, many teachers have shifted focus from grammar to content, they invariably have a team-teaching lesson with NES and teach from English textbooks that emphasize communicative activities (Yuasa, 2010). Meanwhile, ARs are considered to have a broader scope and knowledge of ratings as they have undergone professional rater training and tend to have more practical rating experience. This may account for their understanding of Aptis as a language test, for them, language use is more important than content. AR-B had taught academic writing and clearly stated that content was more important when teaching academic writing. Others explained that they shift their focus when they evaluate the different examinations such as Cambridge Assessment English and IELTS.

AR-B stated:

> *I used to be an academic writing teacher. If I'm marking for example essays in my University of undergraduate students, then it's the task achievement is a little bit valued more than, for example, Aptis is more language here.*

AR-A stated:

> *I started as an IELTS examiner and then moved to Aptis. The focus is more task achievement. I found it hard at first with Aptis to focus more on the grammar levels, so I have been trying to look more at the structures.*

It is possible that having received more training as a rater and working as a professional rater, the ARs are more flexible in terms of their focus, depending on the test. On the other hand, it is also possible that the majority of the JTs have never been professionally rater trained, certainly many teachers do not study language testing when at university studying for their English teaching qualifications in Japan, which is not a compulsory subject (MEXT, 2014a). As Table 8 shows, no JTs perceived organization to be the most difficult feature, and 6 JTs perceived task achievement as the least difficult feature. For the JTs, content is much easier than language use when evaluating, therefore their lack of knowledge as a rater may lead to a tendency to focus on content (task achievement and organization).

In terms of differences revealed in each group's focus on the evaluation of language use, the JTs focused on grammatical and lexical accuracy whereas the ARs paid greater attention to wider aspects, such as general accuracy and complexity. The interviews made it possible to notice differences not only in the items raters looked at during their evaluations but also in the process by which the evaluation was conducted. The ARs evaluated language use based on what test-takers were able to do, i.e., positive performance, however, the JTs focused on negative performance features, such as learners' errors, and adjusted marks accordingly, similar to negative marking.

According to the interviews, all three JTs mentioned grammatical errors as a reason for their holistic scores. JT-3 spoke about Essay 9, in which he gave a score of 9 on a 10-point scale. Surprisingly, despite this high score, he focused on weaknesses, rather than strengths, and pointed out errors.

> *"A" international competition is an error for "an" international competition.*

This suggests he focused only on grammatical accuracy, not grammatical complexity and range. The fact that JTs pinpointed grammatical and structural errors may suggest they are less tolerant of students' linguistic errors. Similar findings have been presented in previous research (e.g., Connor-Linton, 1995; Lee, 2009; Rao & Chen, 2020) (see Section 2.7.2).

One possible interpretation of this phenomenon may lie in the JTs lower level of English language proficiency. JTs are more confident about checking for errors they can detect within their own knowledge, certainly, they are familiar with common errors in EFL learners' writing having themselves gone through the process of developing their English in the context of a foreign language learner (Rao & Chen, 2020). By way of contrast, the ARs frequently referred to CEFR levels when judging grammar and vocabulary used in a particular essay, rather than drawing on personal experiences of learning.

AR-C stated:

> *I was focusing on the positive and I was looking at the mark from one to 10.*

AR-A referred to the CEFR level saying:

> *I would say it's like a B1 level, so something on this scale, four or five.*

As professional raters with a high level of English proficiency, the ARs have a wider scope of criteria when marking language use than EFL raters. These findings support the results in Rao and Li's (2017) and Shi's (2001) studies in which NES were shown to have a wider scope of criteria in marking language use than EFL raters. With limitations in their grammatical and lexical knowledge, it could be challenging for EFL raters to assess these criteria by drawing on their knowledge to decide the level of language use.

Another possible reason for the differences found may be related to the language learning and teaching patterns of the raters. Methods of most Japanese teachers of English are dominated by teacher-centered and text-based approaches. According to the Japanese national curriculum, high school students take mainly “communication English” and “English expression” (MEXT, 2019). These titles differ completely from the content of the textbooks used; the former is based on reading materials and the latter on grammar guidance. Even so, teachers are instructed to use the textbook-centered methodology in Japan. They endeavor to make the students understand the textbooks, firstly explaining the meaning of new words and phrases, making their students interpret each sentence, and then introducing grammar knowledge (Butler & Iino, 2005). Although English teachers try to teach communicative lessons, in reality, the emphasis is on reading comprehension and grammar. The objective tends to be to see whether learners can complete appropriate sentences and whether they have acquired the grammatical knowledge they have learned. In other words, the teacher’s emphasis tends to be on accuracy. Many teachers in Japan focused on negative performance features, such as learners’ errors, and adjusted marks accordingly (Nogami, 2016; Omiyama, 2019), and the JTs in this study were no different when evaluating language use. They have not developed the habit of checking the complexity of grammar and vocabulary. Should writing classes and rater training become better developed in Japanese high schools, teachers’ concerns may extend to complexity. The current teaching pattern seemed to influence this study’s JTs’ judgment and rating processes which continue to pay special attention to the accuracy of the language form. This also accords with earlier observations (Rao & Chen, 2020), which show that EFL Chinese raters focus on accuracy in the evaluation of language use and pinpoint grammatical and structural errors, a finding which is related to the Chinese teacher-centered teaching pattern.

Finally, the importance of rater training needs to be considered. It has been made clear that all ARs received professional rater training, while the JTs received no such training. This suggests the JTs may not know to score essays other than by reducing the number of marks for a single error, thus allocating scores on language use to score negatively.

5.4 Implications and limitations of the study

5.4.1 Implications

Based on the present findings, this study examined cases where difficulties arise when raters from different backgrounds evaluate essays on a holistic scale. This consideration is particularly important for policymakers, language curriculum developers, and individual teachers in EFL settings where highly proficient English speakers and EFL teachers need to work together in teaching activities.

Firstly, as pointed out by Shi (2001), is the issue of scoring validity in language testing. This study has shown that the evaluation process and perspectives were very different between the two groups of raters. It is important to note that these differences, revealed in the qualitative analysis, were not reflected in the holistic ratings allocated to the 20 essays which actually showed no significant differences. The discrepancy between the two rater groups with different backgrounds underlines the disadvantages of holistic evaluations. Certainly, the study raises questions about the extent to which holistic scoring reflects analytical or qualitative competence. There is a clear risk that the construction of the test can be interpreted differently by different groups due to the different writing components each group aims to evaluate.

In addition, although holistic scoring seems to be effective at first glance – there is no difference between the two groups in the holistic scores allocated – there may be scoring validity issues with the feedback pupils receive (Shi, 2001). Certainly, it is difficult to understand what traits are evaluated because students cannot see where they are being assessed, which is generally seen as a disadvantage of holistic scoring (Connor-Linton, 1995). Even if the same score is given and raters provide some form of feedback, the feedback they receive on the content of the work can differ significantly, a source of confusion for learners. For example, the study suggests that NJS teachers would make comments related to task achievement and organization. On the other hand, professionally-trained raters would comment on grammar and vocabulary. The type of feedback and the score they receive for their English writing might influence students' behavior and writing skills.

Thirdly, such variations in scoring methods can have a washback effect on student's attitudes towards learning and examinations, which is concerned with consequential validity. The JTs' tendency to emphasize accuracy when evaluating language use, seems to encourage students to use simple phrases or grammar. Also, when marking is based on errors, there is a tendency for writers to become defensive as they try to avoid making mistakes as much as possible (Xie, 2015). Learners tend to make simple word choices at the levels with which they are confident when the breadth of grammar and vocabulary is not assessed. On the other hand, if teachers assess on a can-do basis, like the ARs in this study, learners' attitudes towards language usage may also change so they become more confident to take risks and use more difficult vocabulary. Thus, teaching goals and instruction received as a result of evaluation have a strong impact on students' attitudes towards writing.

To remedy these possible issues, rater training sessions must consider the differences between raters from different backgrounds and arrive at a consensus about evaluation criteria before tests are taken. L1 English raters in an EFL context should understand the criteria that local Japanese students need to meet because students usually need to pass the examination that local teachers create and rate (Kobayashi & Rinnert, 1996; Shi, 2001). In contrast, Japanese teachers should be aware of the different evaluation processes of professional raters of international English language tests because approximately 60,000 Japanese people study abroad every year (MEXT, 2014b), and most need to take an international test. Japanese teachers endeavor to fill the gap between their criteria and the international standard.

5.4.2 Limitations and suggestions for further research

Although this study has contributed to deepening current knowledge about rater variance among raters with different backgrounds, there remain limitations. One limitation is that the number of raters in each group is relatively small and cannot be generalized to the entire population of the two groups.

In addition, the study revealed differences in the assessment process and judgments made by the ARs and the JTs but these remain ambiguous because we do not know the individual’s perspective or their definition of the assessment criteria. For example, it is not clear whether a reference to grammar reflects accuracy or complexity, or whether task achievement refers to the communicative perspective or the basic format of the assessed work. Further study could investigate these issues in detail; e.g., research that investigates the decision-making process might be applicable to these issues. For instance, Cumming et al. (2002) recorded the decision-making process of raters by using verbal reports collected at the time the raters were evaluating essays; such an approach would give deeper insight into the personal decisions made by individual raters. A final concern relates to the complexity of coding for a single researcher at a Master’s degree level who cannot manage large amounts of qualitative data given time restrictions, however, the study shows that, in spite of this, a small-scale study can be effective in gaining a deeper insight into the decision-making process of the raters.

6. Conclusion

The research presented in this paper examined to what extent ARs and JTs differ in their assessment of 20 essays written by Japanese teens. Data were collected for both quantitative and qualitative analysis and included student essays that provided holistic scores along with their reasons for allocating their scores. A questionnaire and individual interviews allowed exploration of not only how the two groups assess Japanese teens' writing holistically, but also how the two groups differed in their qualitative judgment and analytical reasons given for their scores. The results from the holistic scoring show that the ARs and the JTs achieved reasonably high reliability in their evaluation of the 20 essays and no significant differences were found in their holistic scoring.

However, the qualitative data, which drew out the reasons given for the ratings, showed considerable differences in the frequency of the different reasons given for the holistic scores. The analysis of these reasons reveals that the JTs were most concerned with task achievement from among the other marking criteria when rating the overall quality of essays; meanwhile, the ARs firstly put emphasis on language use and then moved their attention to language content. Furthermore, the analysis of the post-marking questionnaires and semi-structured interviews provided more detailed insights into the differences in the perspectives and processes of the raters. An interesting finding is that the JTs placed more emphasis on organization and tended to evaluate an essay more positively when it followed their familiar structure of introduction, body, and conclusion. In contrast, they tended to more negatively evaluate incomplete essays than the AR group. In addition, the raters' perspectives also differed considerably. The ARs evaluated language use based on what test-takers were able to do or positive performance; however, JTs focused on negative features and tended to narrow in on learners' errors, adjusting marks accordingly, similar to negative marking.

Based on the present findings, this study examined cases where difficulties arise when raters from different backgrounds evaluate essays on a holistic scale. The issue of scoring validity in language testing is one of the most serious problems and the findings in this study suggest appropriate rater training in which raters with different backgrounds have a consensus on their rating before the evaluation. These considerations are particularly important for policymakers, language curriculum developers, and individual teachers in EFL settings where highly proficient English speakers and EFL teachers need to work together in teaching activities.

7. References

Abe, E. (2013). Communicative language teaching in Japan: Current practices and future prospects: Investigating students' experiences of current communicative approaches to English language teaching in schools in Japan. *English Today, 29*(2), 46-53.

Bachman, L. F. (2004). *Statistical analyses for language assessment book*. Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice : designing and developing useful language tests*. Oxford University Press.

Bailey, J. L. (2018). A study of the washback effects of university entrance examinations on teaching pedagogy and student learning behaviour in Japanese high schools. *British Journal of Education, 6*(6), 50-72.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*(1), 54-74.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice, 18*(3), 279-293. https://doi.org/10.1080/0969594X.2010.526585

Bax, S. (2012). *Text Inspector*. In [Online text analysis tool]. https://textinspector.co

British Council. (2022). *Aptis - English language test*. Retrieved September 5 from https://www.britishcouncil.org/exam/aptis

Brown, J. D. (1991). Do English and ESL faculties rate writing samples differently? *TESOL quarterly, 25*(4), 587-603.

Butler, Y. G., & Iino, M. (2005). Current Japanese reforms in English language education: The 2003 “action plan”. *Language Policy, 4*(1), 25-45.

Butler, Y. G., Lee, J., & Peng, X. (2020). Failed policy attempts for measuring English speaking abilities in college entrance exams: Cases from China, Japan, and South Korea. *English Today*, 1-7.

Chapelle, C. A. (2012). Conceptions of validity. In G. Fulcher, & Davidson, F. (Ed.), *The routledge handbook of language testing*. Routledge.

Connor-Linton, J. (1995). Crosscultural comparison of writing standards: American ESL and Japanese EFL. *World Englishes, 14*(1), 99-115.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*(1), 31-51.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern language journal (Boulder, Colo.), 86*(1), 67-96. https://doi.org/10.1111/1540-4781.00137

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117-135.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford university press.

Dörnyei, Z., & Taguchi, T. (2009). *Questionnaires in second language research: Construction, administration, and processing*. Routledge.

Du, Y., & Brown, W. L. (1996, April). *Differential facet functioning detection in direct writing assessment* [Paper presentation]. Annual Conference of American

Educational Research Association, New York. https://files.eric.ed.gov/fulltext/ED400293.pdf

Fulcher, G. (2010). *Practical language testing*. Hodder Education. http://reading.eblib.com/patron/FullRecord.aspx?p=615889

Green, A. (2014). *Exploring language assessment and testing: language in action*. Routledge. https://doi.org/10.4324/9781315889627

Hijikata-Someya, Y., Ono, M., & Yamanishi, H. (2015). Evaluation by native and non-native English teacher: Raters of Japanese students'　summaries. *English Language Teaching, 8*(7), 1-12.

Hosoki, Y. (2011). English language education in Japan: Transitions and challenges (I). *Journal of International Relations, Kyushu International University, 6*(1/2), 199-215. http://id.nii.ac.jp/1265/00000272/

Huang, B. H. (2013). The effects of accent familiarity and language teaching experience on raters' judgments of non-native speech. *System, 41*(3), 770-785. https://doi.org/https://doi.org/10.1016/j.system.2013.07.009

Hughes, A., & Lascaratou, C. (1982). Competing criteria for error gravity. *ELT journal, 36*(3), 175-182.

Hyland, K. (2019). *Second language writing*. Cambridge university press.

Hyland, K., & Anan, E. (2006). Teachers' perceptions of error: The effects of first language and experience. *System (Linköping)*, *34*(4), 509-519. https://doi.org/10.1016/j.system.2006.09.001

James, C. (1977). Judgements of error gravities. *English language teaching journal, 31*(2), 116. https://go.exlibris.link/ChxNQN3g

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing, 26*(4), 485-505.

Kachru, B. B. (1992). *The other tongue: English across cultures*. University of Illinois press.

Kang, O., Rubin, D., & Kermad, A. (2019). The effect of training and rater differences on oral proficiency assessment. *Language Testing, 36*(4), 481-504.

Kim, Y.-H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing, 26*(2), 187-217.

Kitahara, N. (2010). *Eigo jugyo no miki o tsukuru hon (gekan) [A book that forms the core of English classes (second volume)]*. Benesse Holdings.

Knoch, U., Fairbairn, J., & Jin, Y. (2021). *Scoring second language spoken and written performance : Issues, options and directions*. Equinox Publishing Ltd.

Kobayashi, H., & Rinnert, C. (1996). Factors affecting composition evaluation in an EFL context: Cultural rhetorical pattern and readers' background. *Language learning, 46*(3), 397-433.

Kobayashi, T. (1992). Native and nonnative reactions to ESL compositions. *TESOL quarterly, 26*(1), 81-112.

Kowata, T. (2015). *Washback Effects of University Entrance Examination Writing Tasks on Learning and Teaching* [Doctoral dissertation, Tokyo University of Foreign Studies, JAPAN].

Kudo, H., & Negishi, M. (2002). Jiyu sakubun no saiten hoho ni yoru saiten-sha-kan shinrai-sei ni tsuite [About reliability between graders by scoring method of free composition]. *Zenkoku eigo kyoiku gakkai kiyo, 13*, 91-100.

Lado, R. (1961). Language testing: The construction and use of foreign language tests-A teacher's book.

Lee, H.-K. (2009). Native and nonnative rater behavior in grading Korean students' English essays. *Asia Pacific education review, 10*(3), 387-397. https://doi.org/10.1007/s12564-009-9030-3

Ling, G., Mollaun, P., & Xi, X. (2014). A study on the impact of fatigue on human raters when scoring speaking responses. *Language Testing, 31*(4), 479-499.

Lowie, W. M., Haines, K. B., & Jansma, P. N. (2010). Embedding the CEFR in the academic domain: Assessment of language tasks. *Procedia-Social and Behavioral Sciences, 3*, 152-161.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54-71.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*(2), 158-180.

Marefat, F., & Heydari, M. (2016). Native and Iranian teachers' perceptions and evaluation of Iranian students' English essays. *Assessing writing, 27*, 24-36.

Marsh, H. W., & Ireland, R. (1987). The assessment of writing effectiveness: A multidimensional perspective. *Australian Journal of psychology, 39*(3), 353-367.

McNamara, T. (2000). *Language testing*. Oxford University Press.

McNamara, T. F. (1996). *Measuring second language performance*. Longman.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher, 18*(2), 5-11.

Messick, S. (1994). Validity of psychological assessment: Validation of infrences from persons' responses and performance as scienfitic inquiry into score meaning. *ETS Research Report Series, 1994*(2), i-28.

MEXT. (2003). *Eigo ga tsukaeru nihonjin no ikusei no tame no kodo keikaku [Action plan for nurturing Japanese who can use English]*. Retrieved from https://www.mext.go.jp/b_menu/shingi/chukyo/chukyo3/004/siryo/04031601/005.pdf

MEXT. (2014a). *Kyoin menkyojo shutoku ni kakaru hitsuyo tanisuto no gaiyo [Overview of the number of credits required for obtaining a teacher's license]*. Retrieved from https://www.mext.go.jp/b_menu/shingi/chukyo/chukyo3/002/siryo/__icsFiles/afieldfile/2014/04/23/1347091_03.pdf

MEXT. (2014b). *Wakamono no kaigai ryūgaku o torimaku genjo ni tsuite [The Current Situation Surrounding Young People Studying Abroad]*. Retrieved from https://www.cas.go.jp/jp/seisaku/ryuugaku/dai2/sankou2.pdf

MEXT. (2019). *Gaikokugokatsudo gaikokugohen monbukagakusho [Foreign language activities].* Retrieved from https://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2019/03/18/1387017_011.pdf

Nakatani, Y. (2019). Eibun essei no jidoreberu hantei shisutemu to shudo saiten kekka no hikaku kensho: CEFR - J raitingu tesutotasuku kochiku tame no yobi chosa [Comparative verification of English essay automatic level judgment system and manual scoring results: CEFR-J preliminary research for building a writing test task]. *Keizai shirin*, *87*(1), 21-50.

Ninagawa, Y. (2017). Koto gakko ni okeru dotoku kyoiku no arikata ni kansuru kenkyu-kokosei ni totte no dotoku kihan ishiki to kyoin no dotoku kyoiku-kan [Study on the ideal way of moral education in high school-awareness of moral norms for high school students and teachers' view of moral education]. *Tohogakushi*, *46*(1), 127-139.

Nogami, I. (2016). Kokosei no eisakubun ni okeru "bunsho no matomari" ni shoten o ateta pia fidobakku katsudo no koka [Effect of peer feedback activities focusing on "cohesion of sentences" in English composition of high school students]. *Eiken kenkyu houkokusho, 28.*

Numata, K. (2006). Chugakko ni okeru jiyu eisakubun shido no koka raitingusukiru to gakushu-sha ishiki ni ataeru eikyo
[Effect of teaching free English composition in junior high school-Impact on writing skills and learners' awareness]. *Iwate University English Education Journal*(8), 1-19.

O'Sullivan, B. (2011). *Language testing : Theories and practices*. Palgrave Macmillan. http://reading.eblib.com/patron/FullRecord.aspx?p=1812299

Omiyama, K. (2019). *Nihonjin no kokosei no raitingu-ryoku no hattatsu ni okeru ekusutenshivu raitingu no koka ni kansuru jissho-teki kenkyu [In the development of writing skills of Japanese high school students
empirical research on the effects of extended writing]* [Doctoral dissertation, Kyoto University of Foreign Studies]. https://core.ac.uk/download/288296277.pdf

Paltridge, B., & Phakiti, A. (2015). *Research methods in applied linguistics : A practical resource*. Bloomsbury Academic.

Patton, M. Q. (1990). *Qualitative evaluation and research methods*. SAGE Publications, inc.

Polio, C., & Friedman, D. (2016). *Understanding, evaluating, and conducting second language writing research*. Routledge.

Rao, Z., & Chen, H. (2020). Teachers' perceptions of difficulties in team teaching between local- and native-English-speaking teachers in EFL teaching. *Journal of multilingual and multicultural development, 41*(4), 333-347. https://doi.org/10.1080/01434632.2019.1620753

Rao, Z., & Li, X. (2017). Native and non-native teachers' perceptions of error gravity: The effects of cultural and educational factors. *The Asia-Pacific education researcher, 26*(1-2), 51-59. https://doi.org/10.1007/s40299-017-0326-5

Saito, Y. (2019). Impacts of introducing four-skill English tests into university entrance examinations. *The Language Teacher, 43*(2), 9-14.

Santos, T. (1988). Professors' reactions to the academic writing of nonnative-speaking students. *TESOL quarterly, 22*(1), 69-90.

Shaw, S. D., & Weir, C. J. (2007). *Examining writing : Research and practice in assessing second language writing*. Cambridge University Press.

Sheorey, R. (1986). Error perceptions of nativespeaking and non-nativespeaking teachers of ESL. *ELT journal, 40*(4), 306-312.

Shi, L. (2001). Native-and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing, 18*(3), 303-325.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal, 76*(1), 27-33.

Sudo, s. (2020). Hakyu koka ni kansuru risachidezain no kento: Daigaku nyushi kaikaku o ebidensu nimotozuite giron suru tame ni [Examination of research design on ripple effect: To discuss university entrance examination reform based on evidence]. *Gakushuindaigaku eibun gakkaishi, 2020*, 67-87.

Sugimoto, H., & Yamamoto, Y. (2019). Nihon ni okeru Firipinjin gaikokugo shidojoshu (ALT) no koyomondai -gaikoku seinen shochi jigyo (JET) nado o chushin ni [Employment issues for filipino foreign language teaching assistants (ALTs) in Japan-focusing on foreign youth invitation program (JET)]. *Kyotodaigaku daigakuin kyoikugakukenkyuka kiyo, 65*, 179-200.

Tatsukawa, K. (2018). The expected oral proficiency level for Japan's secondary school English teachers: Analysis of the eiken pre-1st grade interview exam.

Journal of pan-pacific asscociation of applied linguistics, 22(1), 89-104. https://doi.org/10.25256/PAAL.22.1.5

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

Weigle, S. C. (2007). Teaching writing teachers about assessment. *Journal of Second Language Writing, 16*(3), 194-209.

Weigle, S. C., Boldt, H., & Valsecchi, M. I. (2003). Effects of task and rater background on the evaluation of ESL student writing: A pilot study. *TESOL quarterly, 37*(2), 345-354. https://doi.org/10.2307/3588510

Weir, C. J. (2005). Language testing and validation. *Hampshire: Palgrave McMillan, 10*, 9780230514577.

White, E. M. (1984). Holisticism. *College Composition and Communication, 35*(4), 400-409.

White, E. M. (1985). *Teaching and assessing writing*. Jossey-Bass Inc.

Xie, Q. (2015). “I must impress the raters!” An investigation of Chinese test-takers' strategies to manage rater impressions. *Assessing writing, 25*, 22-37. https://doi.org/10.1016/j.asw.2015.05.001

Yuasa, K. (2010). English textbooks in Japan and korea. *Journal of pan-pacific asscociation of applied linguistics, 14*(1), 147. https://go.exlibris.link/W05dYRLP

Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing, 28*(1), 31-50. https://go.exlibris.link/7yV84hgH

8. Appendices

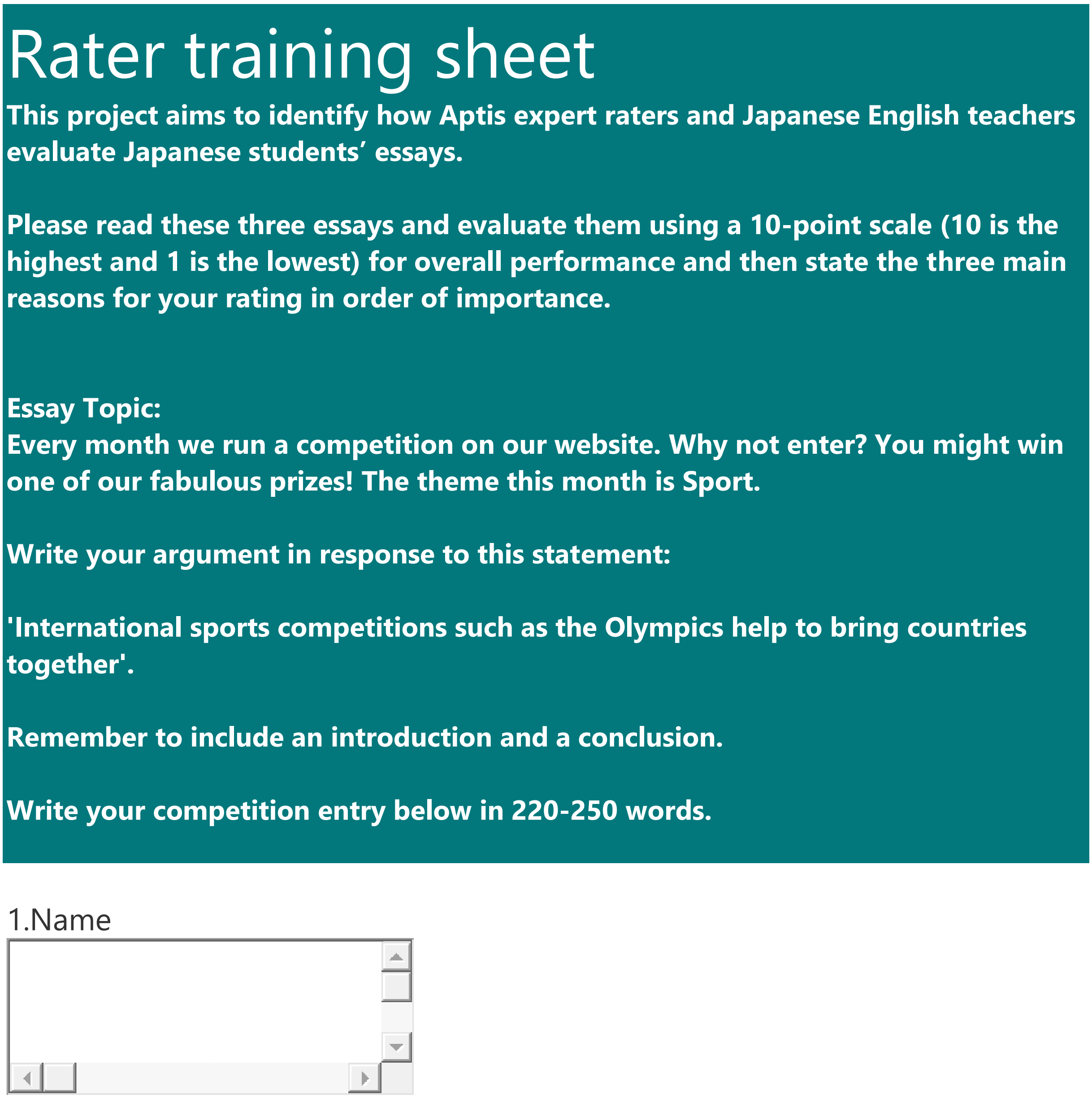
Appendix 1: Type of research: NES and NNES error gravity

Researcher (s)	Year	Context	Number of raters	Aim	Major findings
Hughes and Lascaratou	1982	Greek	10 NES teachers 10 NES students 10 Greek	・To identify the different attitudes toward error	・NES was more lenient than Greek-speaking raters.
Santos	1988	Korea China	144 NES 34 NNES	・To identify differences in perspective	・NES was more tolerant of their judgment than NNES.
Kobayashi	1992	Japan	145 NES 125 Japanese	・To identify the different attitudes toward error	・NES provided far more fixes than NNES and fixed the errors more accurately.
Hyland and Anan	2006	Japan	16 NES teacher 16 NES 16 Japanese teachers	・To identify the different attitudes toward error	・NNE were more lenient scores for grammar while NES are more lenient scores for clarity, naturalness, and organization. ・NES corrected more errors than NNES.
Rao and Liu	2017	China	22 NES 25 Chinese	・To identify the different attitudes toward error ・To identify the factors that may lead to the differences in the evaluation of error gravity	・NES are more lenient of student errors than Chinese raters. ・NES relies more on intelligibility than rule infringement.

Appendix 2: Type of research: NES and NNES evaluation of authentic written work

Researcher(s)	Year	Context	Number of raters	Main research methods		Aim	Major findings
				Holistic	Analytic		
Connor-Linton	1995	Japan	26 NES 29 Japanese	✔		To identify differences in severity and assessment perspective	· NES focused on discourse and grammar whereas NNES focused on matters of accuracy.
Kobayashi and Rinnert	1996	Japan	106 NES teachers 104 Japanese teachers 255 Japanese students	✔	✔	· To identify the different effects of rhetorical patterns and errors	· NNES have significantly higher scores than NES.
Shi	2001	China	23 NES 23 Chinese	✔		· To identify differences in severity and assessment perspective	· NES was more positive about content and language, while NNES was negative about organization and length.
Lee	2009	Korea	5 NES 5 Korean		✔	· To identify differences in severity and assessment perspectives	· NES has been stricter in content and overall, and NNES has been stricter in grammar, sentence pattern, and structure.
Marefat	2016	Iran	12 NES 12 Iranian	✔		· To identify differences in severity and assessment perspectives	· Iran's evaluators were stricter than NES in essay evaluation.
Rao and Liu	2020	China	25 NES 28 Chinese	✔		· To identify differences in severity and assessment perspective	· NES was more tolerant of grammar and sentence structure. · NNES were less severe in ideas and arguments.

Appendix 3: Rater training sheet – English version



Rater training sheet

This project aims to identify how Aptis expert raters and Japanese English teachers evaluate Japanese students' essays.

Please read these three essays and evaluate them using a 10-point scale (10 is the highest and 1 is the lowest) for overall performance and then state the three main reasons for your rating in order of importance.

Essay Topic:
Every month we run a competition on our website. Why not enter? You might win one of our fabulous prizes! The theme this month is Sport.

Write your argument in response to this statement:

'International sports competitions such as the Olympics help to bring countries together'.

Remember to include an introduction and a conclusion.

Write your competition entry below in 220-250 words.

1.Name

3. Essay 1

Essay1 (238 words)

Overall Score

○ 1 (the lowest)

○ 2

○ 3

○ 4

○ 5

○ 6

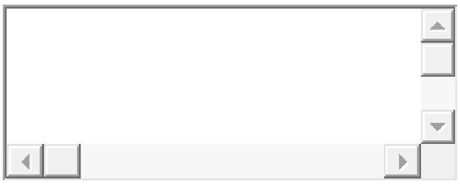
○ 7

○ 8

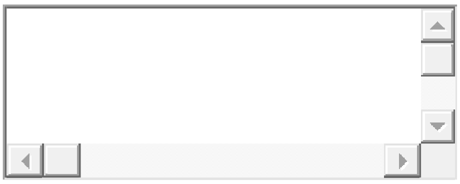
○ 9

○ 10 (the highest)

3.Rank three reasons for your rating in order of importance: First reason (Essay1)



4.Second reason (Essay1)



5.Third reason (Essay1)

6.Essay 2

Overall Score

Essay2 (77 words)

○ 1 (the lowest)

○ 2

○ 3

○ 4

○ 5

○ 6

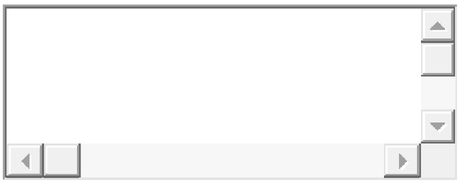
○ 7

○ 8

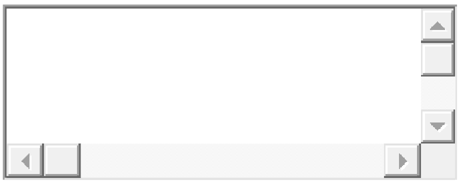
○ 9

○ 10 (the highest)

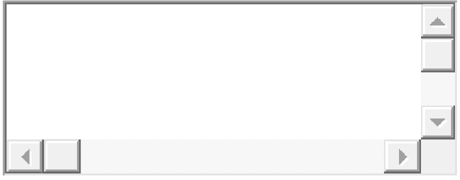
7.Rank three reasons for your rating in order of importance: First reason (Essay2)



8.Second reason (Essay2)



9.Third reason (Essay2)



10.Essay 3

Overall Score

Essay3 (36 words)

○ 1 (the lowest)

○ 2

○ 3

○ 4

○ 5

○ 6

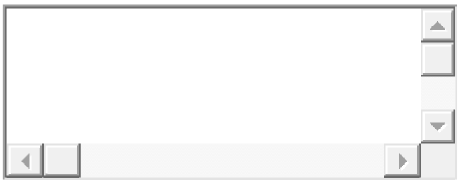
○ 7

○ 8

○ 9

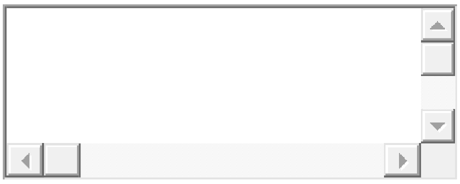
○ 10 (the highest)

11.Rank three reasons for your rating in order of importance: First reason (Essay 3)

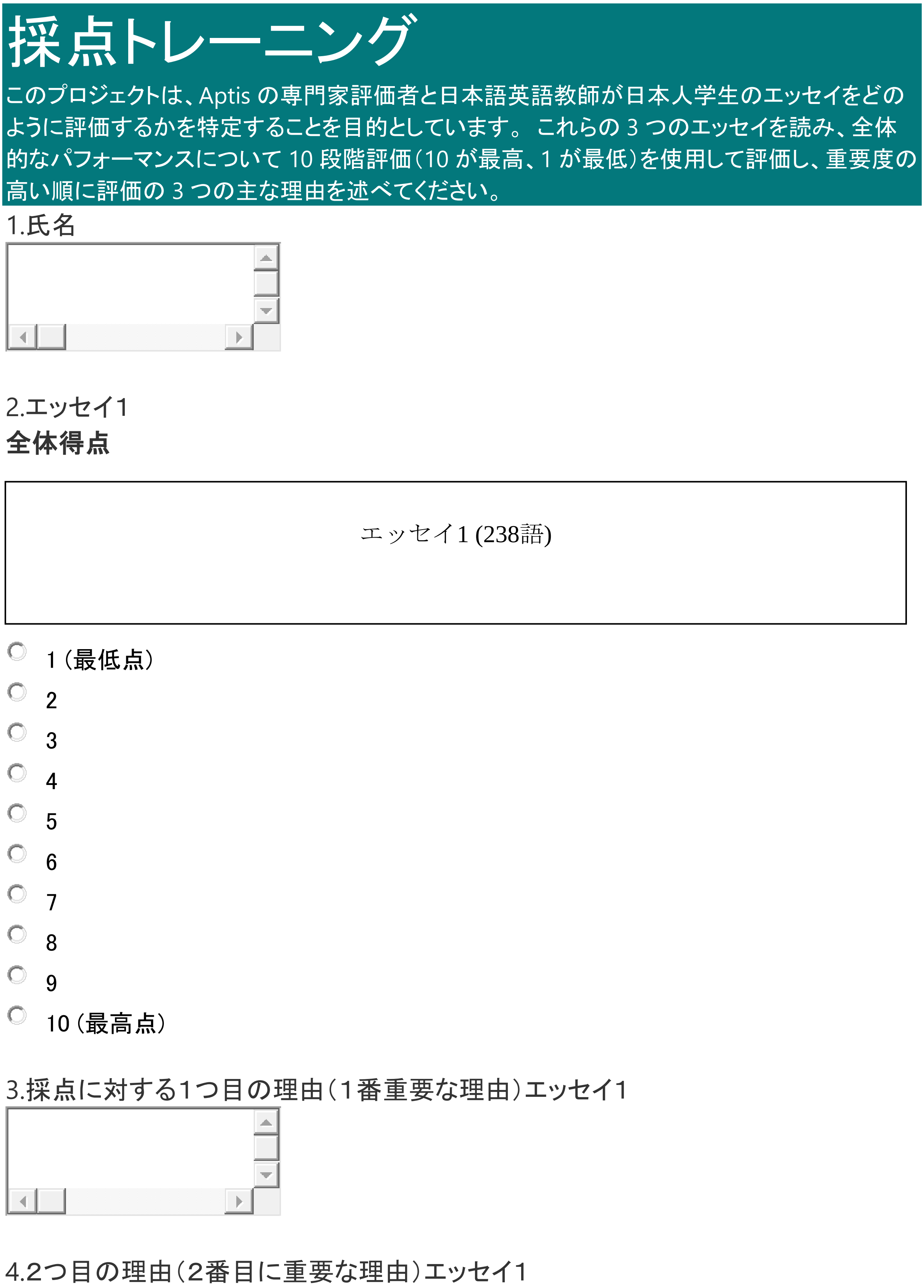


12.Second reason (Essay3)

13.Third reason (Essay3)



Appendix 4: Rater training sheet – Japanese version



採点トレーニング

このプロジェクトは、Aptisの専門家評価者と日本語英語教師が日本人学生のエッセイをどのように評価するかを特定することを目的としています。 これらの3つのエッセイを読み、全体的なパフォーマンスについて10段階評価（10が最高、1が最低）を使用して評価し、重要度の高い順に評価の3つの主な理由を述べてください。

1.氏名

2.エッセイ1

全体得点

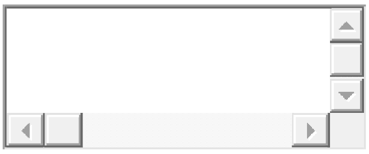
エッセイ1 (238語)

○ 1（最低点）
○ 2
○ 3
○ 4
○ 5
○ 6
○ 7
○ 8
○ 9
○ 10（最高点）

3.採点に対する1つ目の理由（1番重要な理由）エッセイ1

4.2つ目の理由（2番目に重要な理由）エッセイ1

5.3つ目の理由（3番目に重要な理由）エッセイ1



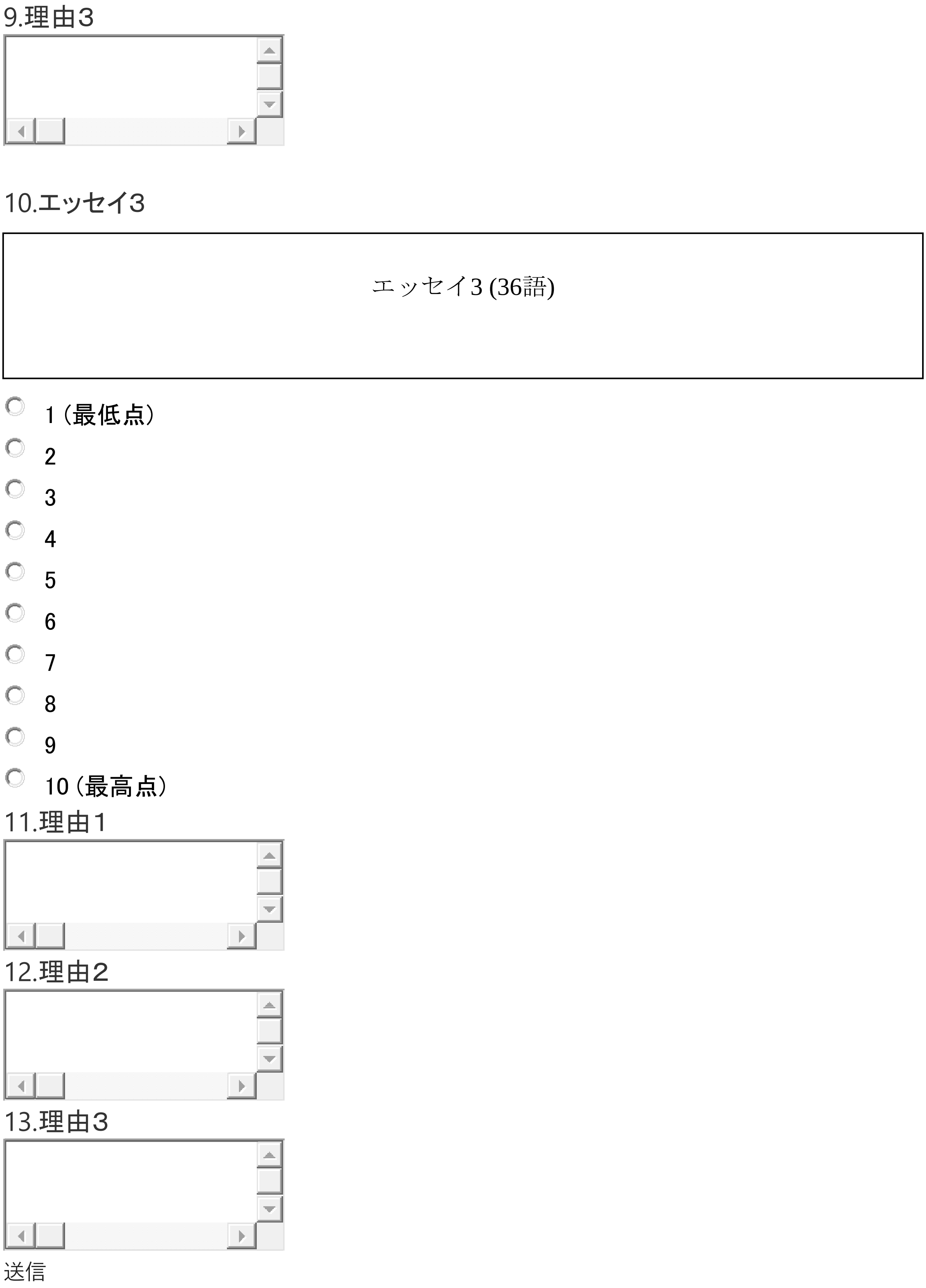
6.エッセイ2

エッセイ2 (77語)

○ 1（最低点）
○ 2
○ 3
○ 4
○ 5
○ 6
○ 7
○ 8
○ 9
○ 10（最高点）

7.理由1

8.理由2



9.理由3

10.エッセイ3

エッセイ3 (36語)

○ 1（最低点）
○ 2
○ 3
○ 4
○ 5
○ 6
○ 7
○ 8
○ 9
○ 10（最高点）

11.理由1

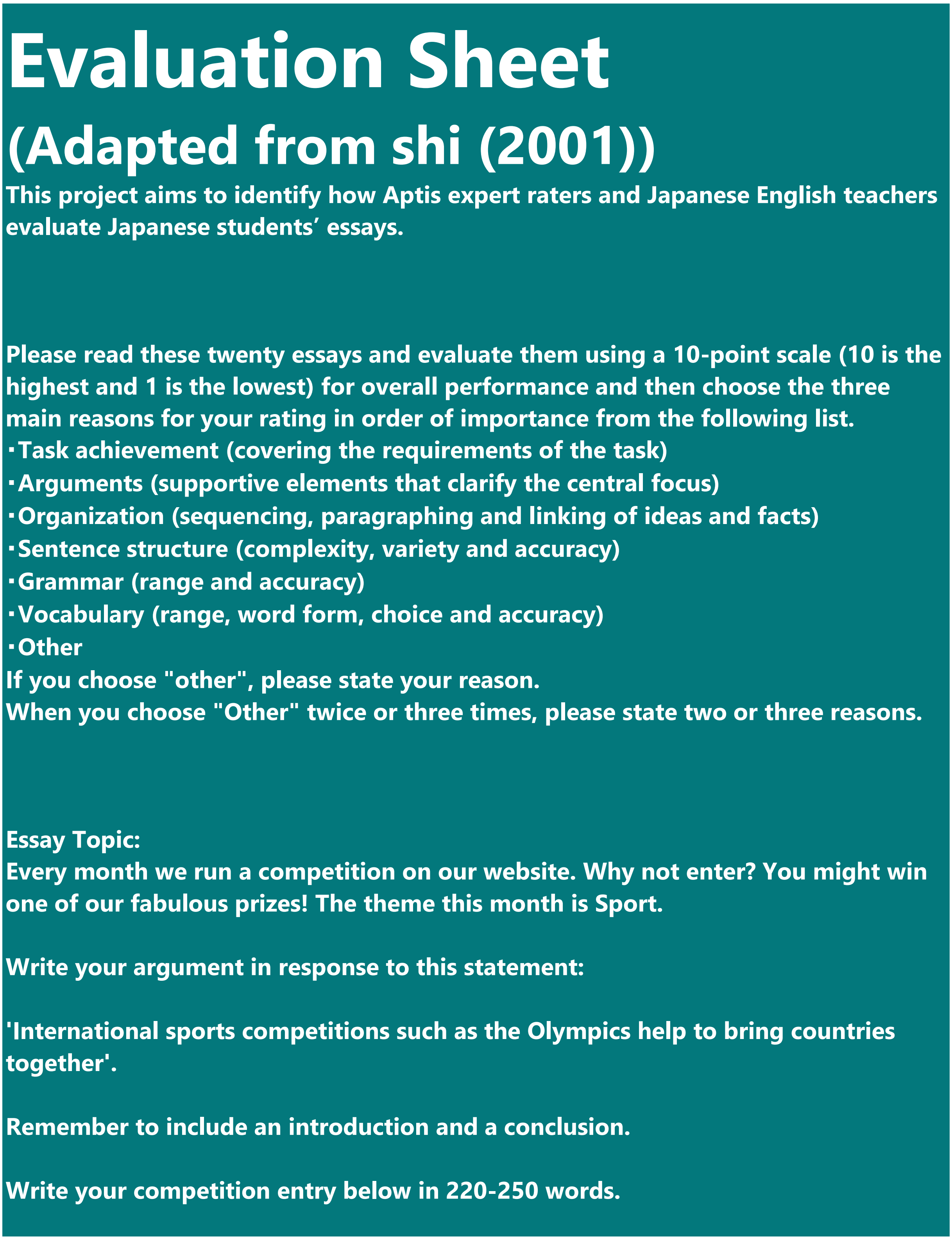
12.理由2

13.理由3

送信

Appendix 5: Evaluation sheet – English version

*These questions are used for each of 20 essays



Evaluation Sheet

(Adapted from shi (2001))

This project aims to identify how Aptis expert raters and Japanese English teachers evaluate Japanese students’ essays.

Please read these twenty essays and evaluate them using a 10-point scale (10 is the highest and 1 is the lowest) for overall performance and then choose the three main reasons for your rating in order of importance from the following list.

･Task achievement (covering the requirements of the task)

･Arguments (supportive elements that clarify the central focus)

･Organization (sequencing, paragraphing and linking of ideas and facts)

･Sentence structure (complexity, variety and accuracy)

･Grammar (range and accuracy)

･Vocabulary (range, word form, choice and accuracy)

･Other

If you choose "other", please state your reason.

When you choose "Other" twice or three times, please state two or three reasons.

Essay Topic:

Every month we run a competition on our website. Why not enter? You might win one of our fabulous prizes! The theme this month is Sport.

Write your argument in response to this statement:

'International sports competitions such as the Olympics help to bring countries together'.

Remember to include an introduction and a conclusion.

Write your competition entry below in 220-250 words.

Your name

1.Name

Essay1

2. Overall Score (Essay1)

Essay1

○ 1 (the lowest)

○ 2

○ 3

○ 4

○ 5

○ 6

○ 7

○ 8

○ 9

○ 10 (the highest)

3.Choose the three main reasons for your rating in order of importance from the following list.
＊Do not choose the same reasons.

・Task achievement (covering the requirements of the task)

・Arguments (supportive elements that clarify the central focus)

・Organization (sequencing, paragraphing and linking of ideas and facts)

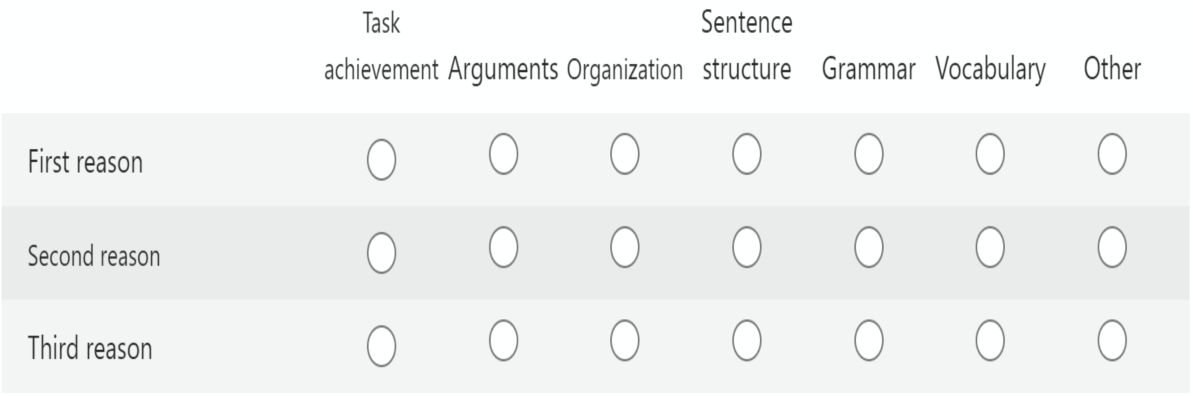
・Sentence structure (complexity, variety and accuracy)

・Grammar (range and accuracy)

・Vocabulary (range, word form, choice and accuracy)

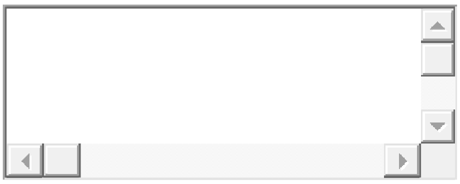
・Other

If you choose "other", please state your reason on the next question.



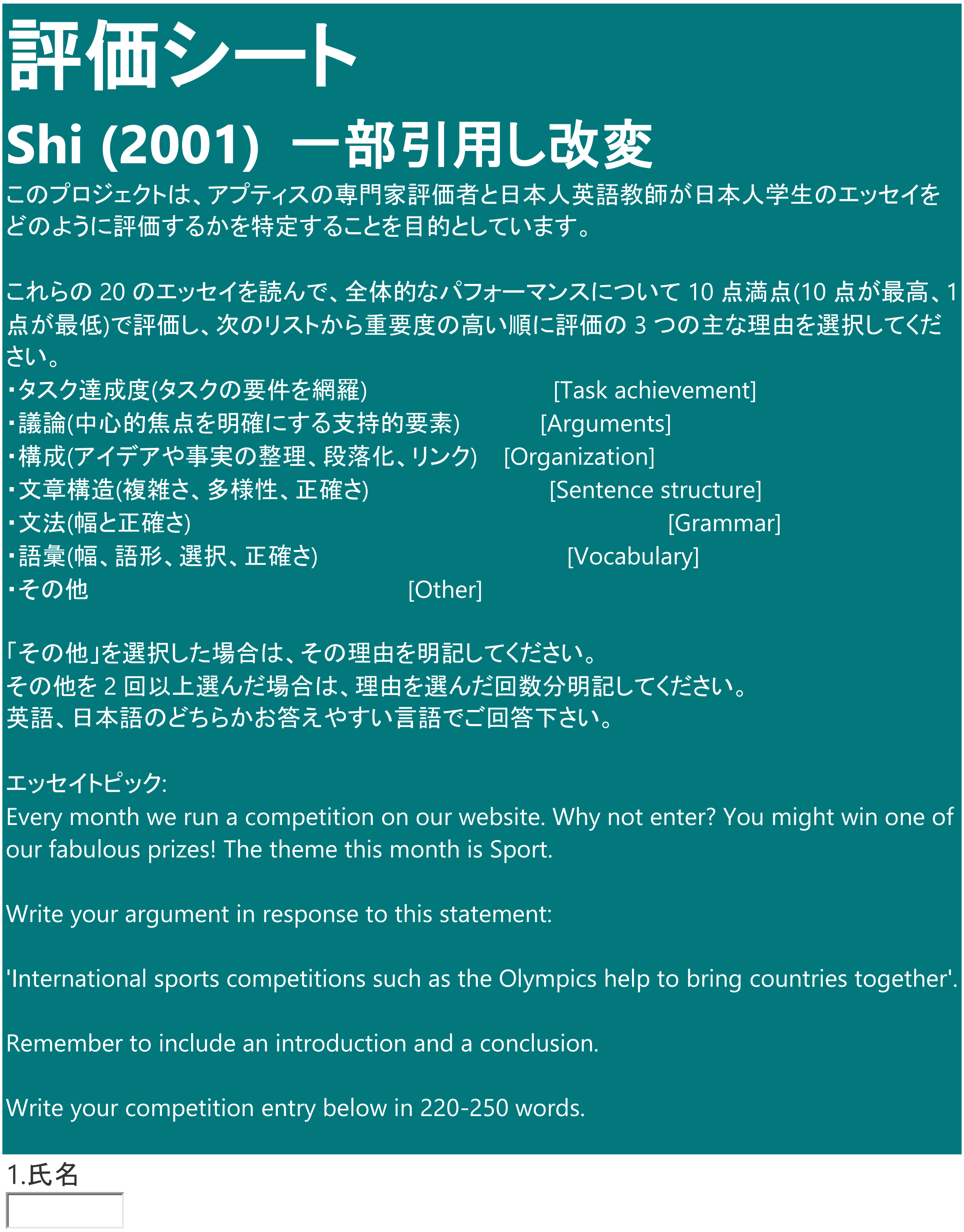
	Task achievement	Arguments	Organization	Sentence structure	Grammar	Vocabulary	Other
First reason	○	○	○	○	○	○	○
Second reason	○	○	○	○	○	○	○
Third reason	○	○	○	○	○	○	○

3. If other, please state your reason.
When you choose other twice or three times, please state two or three reasons.



Appendix 6: Evaluation sheet – Japanese version

*These questions are used for each of 20 essays



評価シート

Shi (2001) 一部引用し改変

このプロジェクトは、アプティスの専門家評価者と日本人英語教師が日本人学生のエッセイをどのように評価するかを特定することを目的としています。

これらの 20 のエッセイを読んで、全体的なパフォーマンスについて 10 点満点(10 点が最高、1 点が最低)で評価し、次のリストから重要度の高い順に評価の 3 つの主な理由を選択してください。

- ・タスク達成度(タスクの要件を網羅) [Task achievement]
- ・議論(中心的焦点を明確にする支持的要素) [Arguments]
- ・構成(アイデアや事実の整理、段落化、リンク) [Organization]
- ・文章構造(複雑さ、多様性、正確さ) [Sentence structure]
- ・文法(幅と正確さ) [Grammar]
- ・語彙(幅、語形、選択、正確さ) [Vocabulary]
- ・その他 [Other]

「その他」を選択した場合は、その理由を明記してください。
その他を 2 回以上選んだ場合は、理由を選んだ回数分明記してください。
英語、日本語のどちらかお答えやすい言語でご回答下さい。

エッセイトピック:
Every month we run a competition on our website. Why not enter? You might win one of our fabulous prizes! The theme this month is Sport.

Write your argument in response to this statement:

'International sports competitions such as the Olympics help to bring countries together'.

Remember to include an introduction and a conclusion.

Write your competition entry below in 220-250 words.

1.氏名

エッセイ1

4. **全体評価 (エッセイ 1)**

エッセイ1

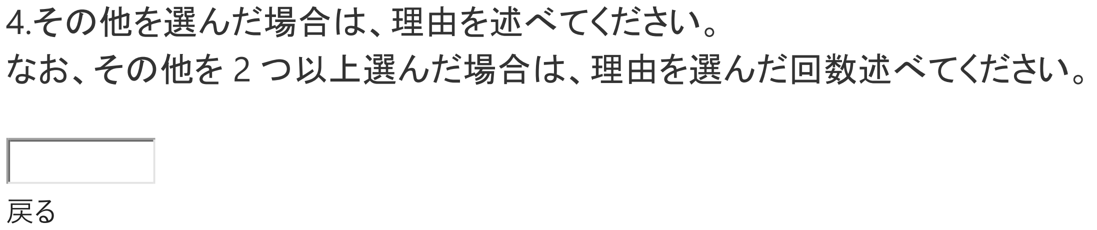
○ 1（最低）
○ 2
○ 3
○ 4
○ 5
○ 6
○ 7
○ 8
○ 9
○ 10（最高）

3.次のリストから重要度の高い順に評価の 3 つの主な理由を選択してください。
＊同じものを選ばないでください

・タスク達成度(タスクの要件を網羅)	[Task achievement]
・議論(中心的焦点を明確にする支持的要素)	[Arguments]
・構成(アイデアや事実の整理、段落化、リンク)	[Organization]
・文章構造(複雑さ、多様性、正確さ)	[Sentence structure]
・文法(幅と正確さ)	[Grammar]
・語彙(幅、語形、選択、正確さ)	[Vocabulary]
・その他	[Other]

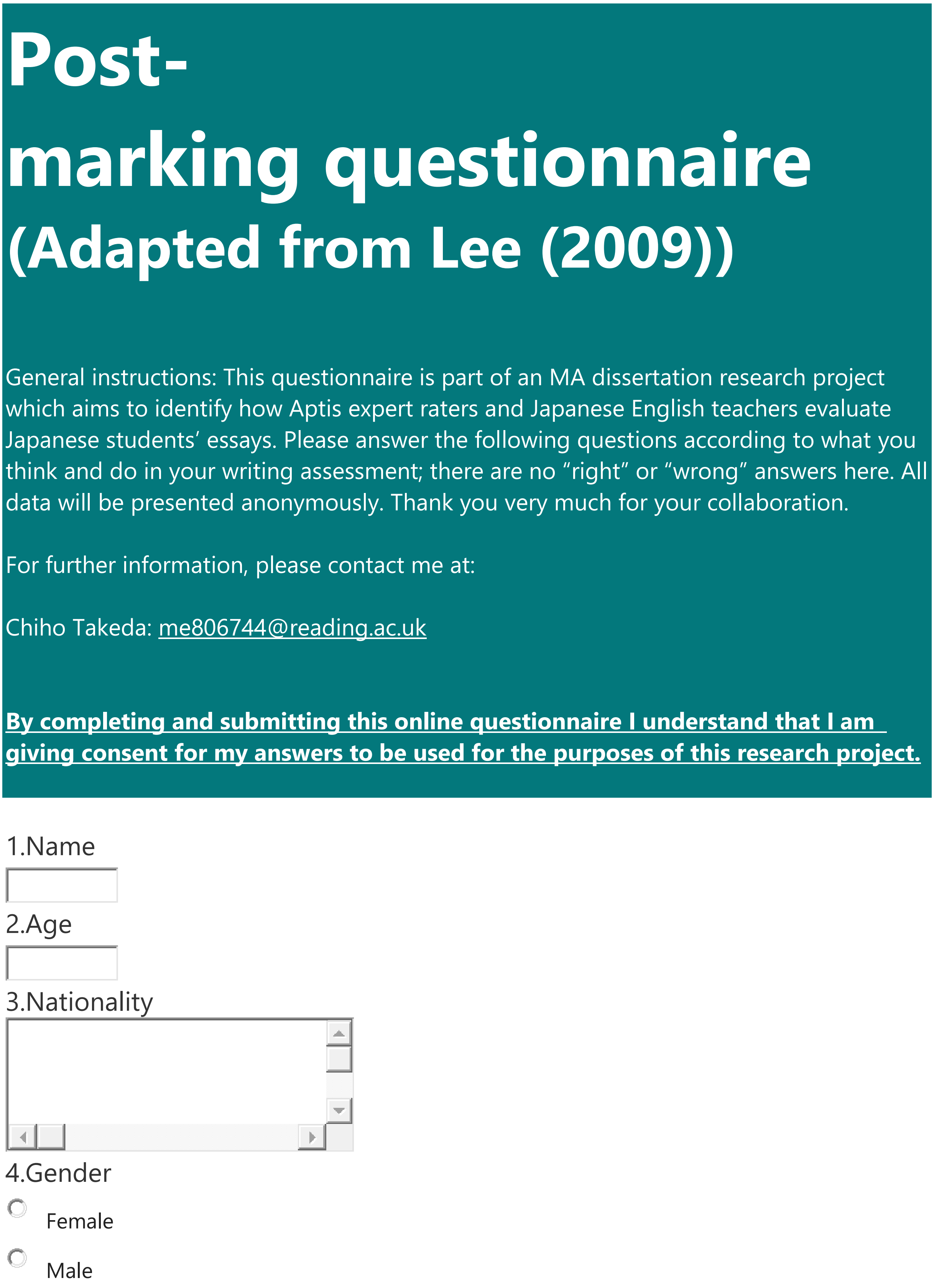
その他を選択した場合は、その理由を次の質問で明記してください。

	タスク達成度	議論	構成	文章構造	文法	語彙	その他
理由 1	○	○	○	○	○	○	○
理由 2	○	○	○	○	○	○	○
理由 3	○	○	○	○	○	○	○



4.その他を選んだ場合は、理由を述べてください。
なお、その他を 2 つ以上選んだ場合は、理由を選んだ回数述べてください。

Appendix 7: Questionnaire – English version



Post-marking questionnaire (Adapted from Lee (2009))

General instructions: This questionnaire is part of an MA dissertation research project which aims to identify how Aptis expert raters and Japanese English teachers evaluate Japanese students' essays. Please answer the following questions according to what you think and do in your writing assessment; there are no "right" or "wrong" answers here. All data will be presented anonymously. Thank you very much for your collaboration.

For further information, please contact me at:

Chiho Takeda: me806744@reading.ac.uk

By completing and submitting this online questionnaire I understand that I am giving consent for my answers to be used for the purposes of this research project.

1.Name

2.Age

3.Nationality

4.Gender

○ Female

○ Male

○ Non-binary

○ Prefer not to say

5.[Educational background]
Your highest degree

○ BA

○ MA

○ PhD

6.Subject area of your major
E.g., BA: English literature/ MA: Education/ PhD: Applied Linguistics

7.[Professional background]
English Teaching experience

○ More than 30 years

○ More than 20 years

○ More than 10 years

○ More than 5 years

○ Less than 5 years

○ No experience

8.English writing teaching experience

○ More than 30 years

○ More than 20 years

○ More than 10 years

○ More than 5 years

○ No experience

9.Have you ever taken part in any writing rating training?

○ Yes

○ No

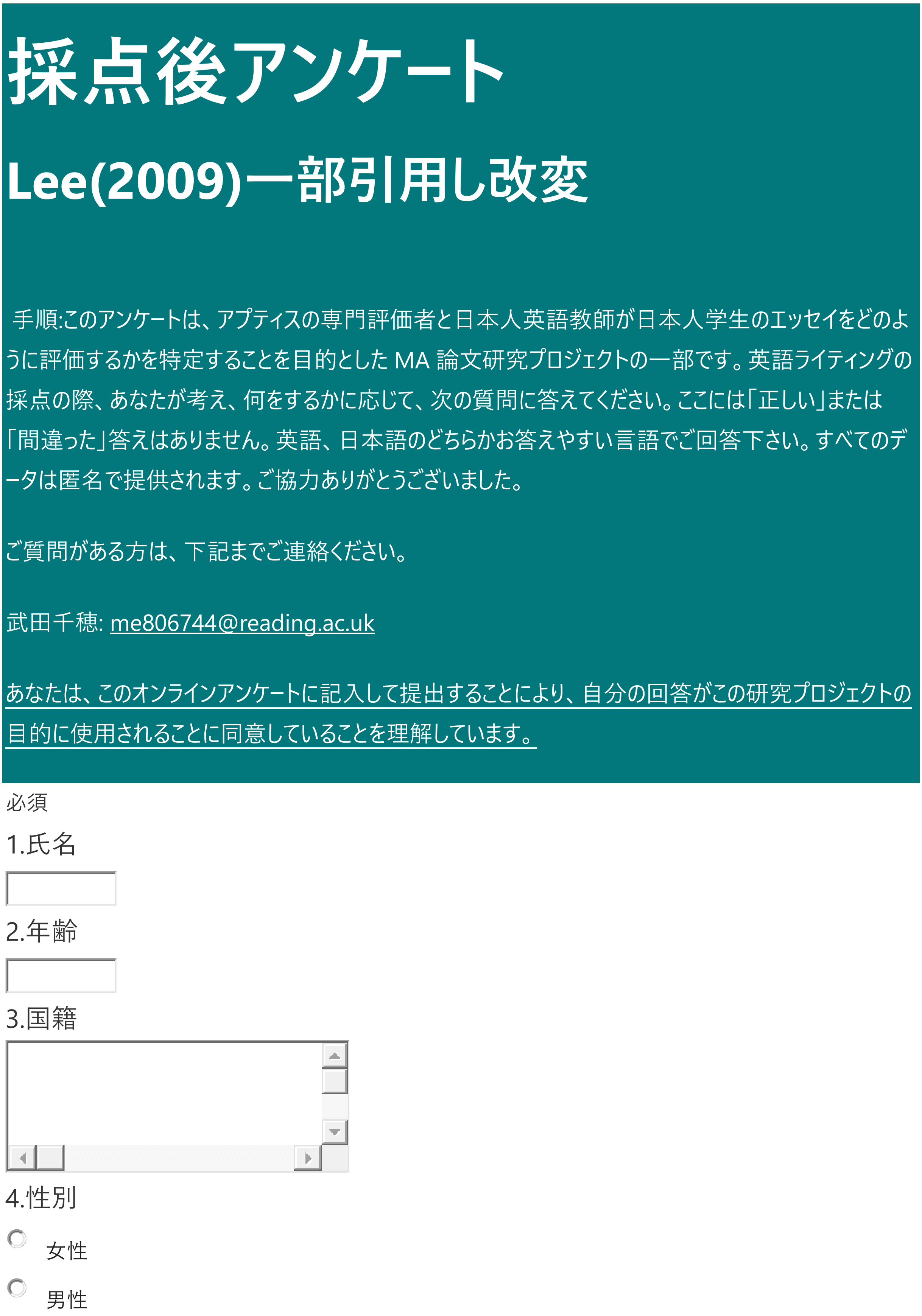
10.**In general,** which element is the most difficulty in deciding on a holistic score for writing?

Rank them from the easiest 1 to the most difficult 6.

Argument
Task achievement
Organization
Vocabulary
Sentence structure
Grammar

Submit

Appendix 8: Questionnaire – Japanese version



採点後アンケート

Lee(2009)一部引用し改変

手順:このアンケートは、アプティスの専門評価者と日本人英語教師が日本人学生のエッセイをどのように評価するかを特定することを目的とした MA 論文研究プロジェクトの一部です。英語ライティングの採点の際、あなたが考え、何をするかに応じて、次の質問に答えてください。ここには「正しい」または「間違った」答えはありません。英語、日本語のどちらかお答えやすい言語でご回答下さい。すべてのデータは匿名で提供されます。ご協力ありがとうございました。

ご質問がある方は、下記までご連絡ください。

武田千穂: me806744@reading.ac.uk

あなたは、このオンラインアンケートに記入して提出することにより、自分の回答がこの研究プロジェクトの目的に使用されることに同意していることを理解しています。

必須

1.氏名

2.年齢

3.国籍

4.性別

○ 女性

○ 男性

○ どちらでもない

○ 記載しない

5.最高学歴

○ 学士

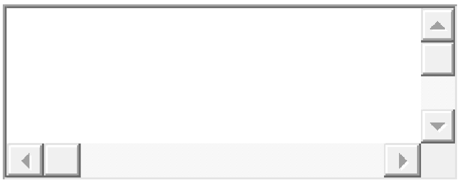
○ 修士

○ 博士

6.専攻

記入例

学士：英文　修士：教育　博士：応用言語学



7.[職歴]

英語指導年数

○ ３０年以上

○ ２０年以上

○ １０年以上

○ 5年以上

○ 5年以下

○ 経験なし

8.[職歴]

英語ライティング指導年数

○ ３０年以上

○ ２０年以上

○ １０年以上

○ 5年以上

○ 5年以下

○ 経験なし

9.採点の研修の受講歴の有無

○ 有

○ 無

10.一般的に、ライティングの全体評価をする際どの項目が採点するのが難しいですか？

1が最もやさしい、6が最も難しいで並べかえてください。

文章構造

タスク達成度

文法

語彙

議論

構成

Appendix 9: Semi-structured Interview “Question Examples”

Items
Questions for general writing assessment
How do you evaluate English essays? Could you illustrate the difference between the criteria used for this research and the ones you usually use?
What challenges did you face when you evaluated essays?
When you mark essays, what do you look for in the essay?
What writing aspect is the most important for you when you evaluate essays?
If you categorize writing components into two, content and organization, and language use, what percentage would you give to each?
What criteria do you usually use to distinguish between serious and minor errors when marking an essay?
Questions for evaluation in this research
Could you explain your evaluation process? What did you look at at first?
In this research, what difficulties did you face?
Your chose grammar was the most difficult to evaluate. How difficult it is?
What characteristics of this essay made you give it this score?
Could you illustrate why you assign like this?
Why did you choose “grammar” as the first reason? Could you explain your first reason more?
After evaluation, what did you think about writing the assessment?

Appendix 10: Coding categories for the semi-structured interview

Themes	Codes	*AR-A	AR-B	AR-C	*JT-1	JT-2	JT-3
Language use	Vocabulary	✔	✔	✔			✔
	Grammar	✔	✔	✔		✔	✔
	Sentence structure	✔	✔	✔			
	Punctuation		✔		✔		✔
	Spelling error			✔	✔	✔	✔
Content	Organization	✔	✔	✔	✔	✔	✔
	Task achievement						
	On topic /off-topic	✔			✔	✔	
	Argument			✔			
	Incomplete	✔	✔	✔	✔	✔	✔
Evaluation tendency	Can do	✔		✔			
	Reduction/Adding	✔	✔	✔	✔	✔	✔
	General evaluation	✔	✔	✔	✔	✔	✔
	L1 influence	✔					
	Bias				✔		

***AR: Trained Aptis rater**
***JT: Japanese high school teacher**

Appendix 11: Coding example for the semi-structured interview

Code		Transcript
	1	I=Interviewer
	2	A=AR-C
	3	I= OK you give seven for this essay and you first reason was grammar, the second was sentence structure and third was arguments. Could you explain this essay?
Code in Green: Can do Code in yellow: Grammar Code in blue: Sentence structure Code in red: Argument	4	A:I gave seven OK, so one second I'm just gonna note this down because. That's how I really. The seven based on grammar and sentence structure arguments. So OK, my because it's not Aptis. So yeah, my rationale for giving the marks from one to 10 reading is. It's not Aptis, so I decided if I'm gonna uh and also we had to give reasons why we gave that mark so I had to for me how I interpreted it is I chose three things which are positive, the most positive things in the essay because I think the other things were vocabulary. The vocabulary something else. No grammar was. I chose grammar. So that the way I chose the top three first of all was. Out of all the possible options. These ones were the ones that were OK, so the grammar I thought. Um, it's simple grammar, but it was. More or less accurate, with regard to the sentence structure is the same I actually found quite difficult to diffirenciate grammar and sentence structure but anyway that’s the another thing and the arguments, even though the arguments were not very strong, you could still see their arguments, whereas I could not give them vocabulary and umm,
	5	I: Can you tell me what the other elements are?
Code in pink: Organization Code in orange: Vocabulary	6	A:I think very and task achievement. Yeah it was origination. Yeah it wasn't well organized because yeah, I can't. OK, but it's not like it doesn't have paragraphs. It's not that great. Our vocabulary was not. It's very simple as well. So like great and Umm, the arguments. No argument was one of the things I showed so that that was my rationale.

Appendix 12: Profile of participants

Trained Aptis raters

No	Age	Nationality	Gender	Degree	Major	Teaching English	Teaching English writing	Previous rater training
1	44	Australian	Female	MA	Applied Linguistics	More than 20 years	More than 20 years	Yes
2	59	British	Male	MA	Cultural Studies,TESOL	More than 30 years	More than 30 years	Yes
3	39	Singaporean	Male	MA	English,Mechanical Engineering	More than 10 years	More than 10 years	Yes
4	43	American & German	Male	MA	TESOL	More than 20 years	More than 10 years	Yes
5	50	South African	Prefer not to say	MA	Psychology	More than 10 years	More than 10 years	Yes
6	38	Polish/British	Female	MA	English Studies	More than 10 years	More than 10 years	Yes
7	48	India	Male	MA	Human Resources	More than 10 years	More than 10 years	Yes
8	40	British	Male	BA	Journalism	More than 10 years	More than 10 years	Yes
9	51	British	Male	MA	English and American studies	More than 20 years	More than 20 years	Yes
10	37	Tanzanian	Female	MA	Applied Linguistics	More than 10 years	More than 10 years	Yes

Japanese high school teachers

No	Age	Nationality	Gender	Degree	Major	Teaching English	Teaching English writing	Previous rater training
1	44	Japan	Female	BA	English literature	More than 20 years	More than 20 years	No
2	27	Japan	Male	BA	International Culture	More than 5 years	More than 5 years	No
3	48	Japan	Male	BA	Economics, English	More than 10 years	More than 10 years	Yes
4	55	Japan	Male	BA	Education	More than 30 years	More than 20 years	No
5	44	Japan	Female	BA	English literature	More than 20 years	More than 20 years	No
6	23	Japan	Male	BA	International Communication	More than 5 years	More than 5 years	No
7	34	Japan	Female	BA	Education	More than 10 years	More than 10 years	Yes
8	39	Japan	Male	MA	TESOL	More than 10 years	No	No
9	33	Japan	Male	MA	Applied Linguistics	More than 10 years	More than 10 years	No
10	49	Japan	Female	BA	Applied Linguistics	More than 20 years	More than 20 years	No

Appendix 13: Consent form – English version

School of Literature and Languages
Department of English Language and Applied Linguistics

ETHICS COMMITTEE

Consent Form

Project title: EFL rater's evaluation of Japanese students' English writing

I understand the purpose of this research and understand what is required of me; I have read and understood the Information Sheet relating to this project, which has been explained to me by Chiho Takeda. I agree to the arrangements described in the Information Sheet in so far as they relate to my participation.

I understand that my participation is entirely voluntary and that I have the right to withdraw from the project at any time.

I have received a copy of this Consent Form and of the accompanying Information Sheet.

Name: ____________________

Signed: ____________________

Date: ____________________

Appendix 14: Consent form – Japanese version

School of Literature and Languages
Department of English Language and Applied Linguistics

University of Reading

倫理委員会

同意書

プロジェクト名: 日本人学生の英作文に対する EFL 評価者による評価

私はこの研究の目的を理解し、私に何が求められているのかを理解しています。研究者の説明があったプロジェクトに関する情報シートを読んで理解しました。私は、私の参加に関連する限り、情報シートに記載されている取り決めに同意します。

私は、私の参加は完全に自発的であり、いつでもプロジェクトから撤退する権利があることを理解しています。

この同意書と添付の情報シートのコピーを受け取りました。

名前：

____

署名：

____

日付：

____

Appendix 15: Information sheet – English version

University of Reading

School of Literature and Languages
Department of English Language and Applied Linguistics

Researcher:
Chiho Takeda
Phone: 0118 378 8141
Email: me806744@reading.ac.uk

Supervisor:
Emma Bruce
phone: +44 (0)7594 647489
Email: emma.bruce@britishcouncil.org

Department of English Language and Applied Linguistics School of Literature and Languages
Edith Morley Building
Whiteknights
Reading RG6 6EL

Phone 0118 378 8147
+44 (0)118 378 6472 +44 (0)118 975 6506
Email appling@reading.ac.uk

INFORMATION SHEET

The purpose of this study is to identify how Aptis expert raters and Japanese English teachers evaluate Japanese students' writings including rating behaviors, tendencies, and processes. This research will help me in my dissertation for a master's degree in TESOL at the University of Reading.

If the chosen participants agree to participate in this research after reading the information sheet, you will rate twenty sample writings using a 10-point scale for overall score and submit it online. Then you will be asked about your rating and background in an online survey. The rating and survey will probably take two hours to complete. After submitting them, selected raters will have an individual online interview about your rating. The interview will take approximately thirty minutes. The interview will be recorded for this research analysis.

Your participation is completely voluntary. Your name will never be mentioned. Instead, use a pseudonym such as "Rater A" to explain. If you wish to withdraw from this study, you can always do so by emailing me.

When the researcher receives the results, they are stored on a password-protected computer. Only the researcher and my supervisor can access the data. All data is used as confidential information only for academic research. When the investigation is complete, it will be destroyed at the end of the year.

This project has been subject to ethical review by the School Ethics Committee and has been allowed to proceed under the exceptions procedure as outlined in paragraph 6 of the University's *Notes for Guidance* on research ethics.

If you have any queries or wish to clarify anything about the study, please feel free to contact my supervisor at the address above or by email at emma.bruce@britishcouncil.org

Signed

Appendix 16: Information sheet for participants – Japanese version

School of Literature and Languages
Department of English Language and Applied Linguistics

研究者:
武田千穂
電話番号: +07413117665
メール: me806744@reading.ac.uk

指導教員:
エマ・ブルース
電話: +44 (0)7594 647489
Eメール: emma.bruce@britishcouncil.org

英語学科 文学・言語学部
エディス・モーリー・ビルディング
ホワイトナイト
RG6 6EL レディング

Phone 0118 378 8147
+44 (0)118 378 6472 +44 (0)118 975 6506
Email appling@reading.ac.uk

研究概要

この研究の目的は、アプティスの専門家と日本人の英語教師が、評価行動、傾向、プロセスを含む日本の学生の文章をどのように評価するかを特定することです。この研究は、レディング大学で TESOL の修士号を取得するための私の修士論文の研究として行われます。

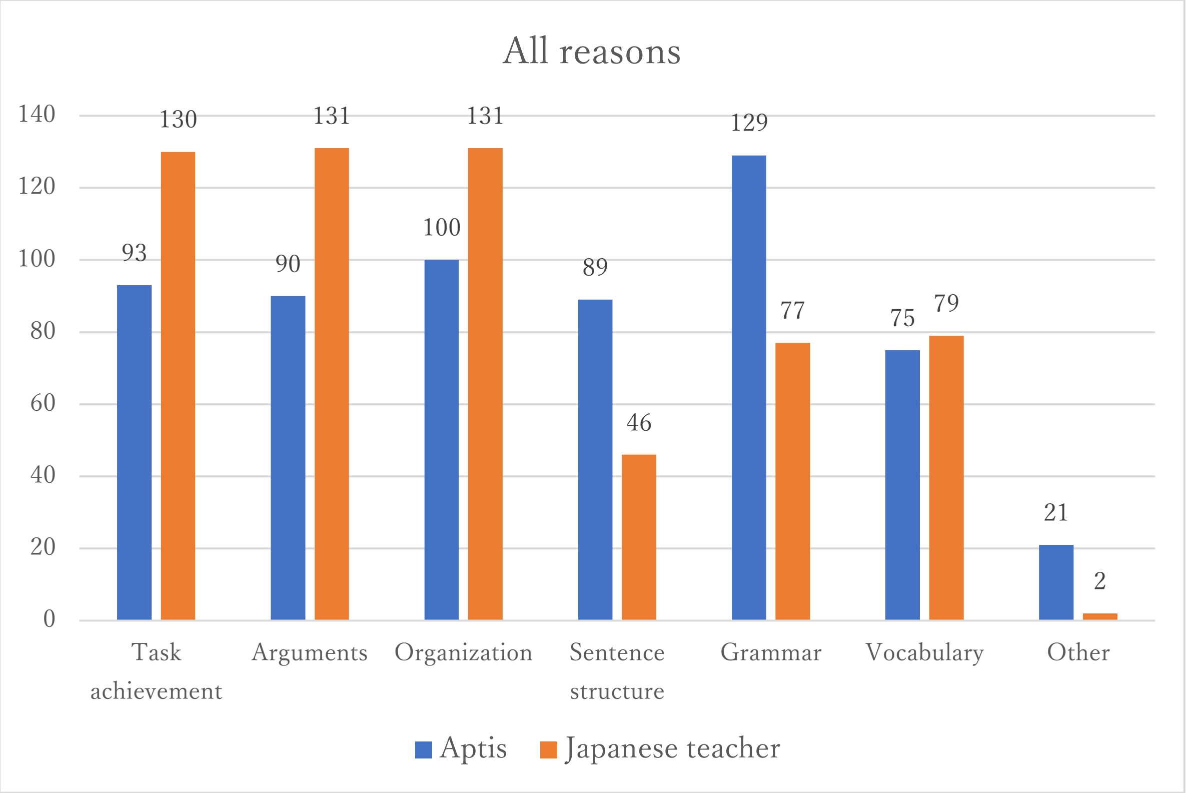
情報シートを読んだ後、選ばれた参加者がこの研究に参加することに同意した場合、あなたは総合スコアの 10 点満点を使用して 20 のサンプルエッセイを評価し、 それをオンラインで提出します。 その後、オンライン調査であなたの評価と背景について尋ねられます。評価と調査が完了するまでにおそらく 2 時間かかります。それらを提出した後、選ばれた評価者はあなたの評価について個々のオンラインインタビューを受けます。面接時間は約 30 分です。インタビューは、この研究分析のために記録されます。

あなたの参加は完全に任意です。あなたの名前は言及されません。代わりに、「採点者 A 」などの仮名を使用して説明してください。あなたがこの研究から辞退したい場合は、いつでも私に電子メールを送ることによって辞退することができます。

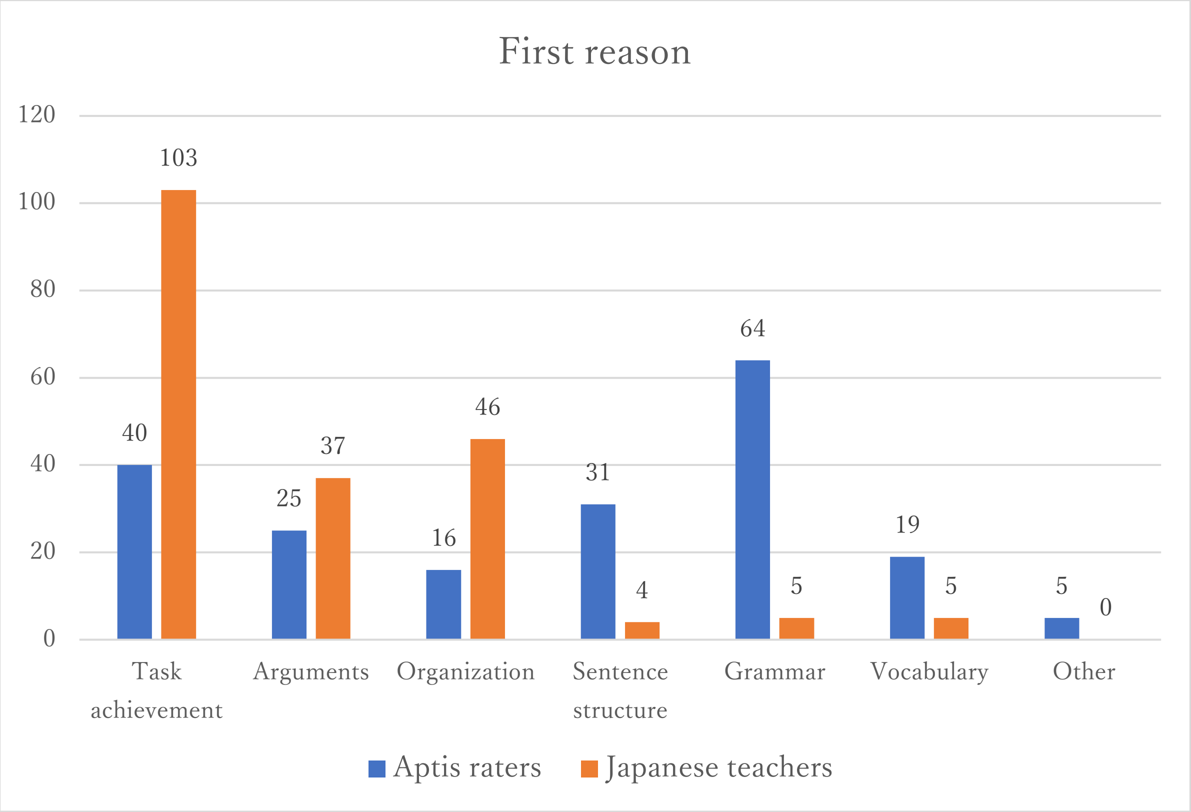
研究者が結果を受け取ると、パスワードで保護されたコンピュータに保存されます。研究者と私の上司だけがデータにアクセスできます。すべてのデータは、学術研究のためにのみ機密情報として使用されます。調査が完了すると、年末に破棄されます。
本事業は、学校倫理委員会の倫理審査の対象となり、本学の研究倫理に関する指導要項第 6 項に定める例外手続の下で進めることが認められています。
ご質問がある場合、または研究について何か明確にしたい場合は、上記のアドレスまたは電子メールで私の指導教員に連絡することができます emma.bruce@britishcouncil.org

署名

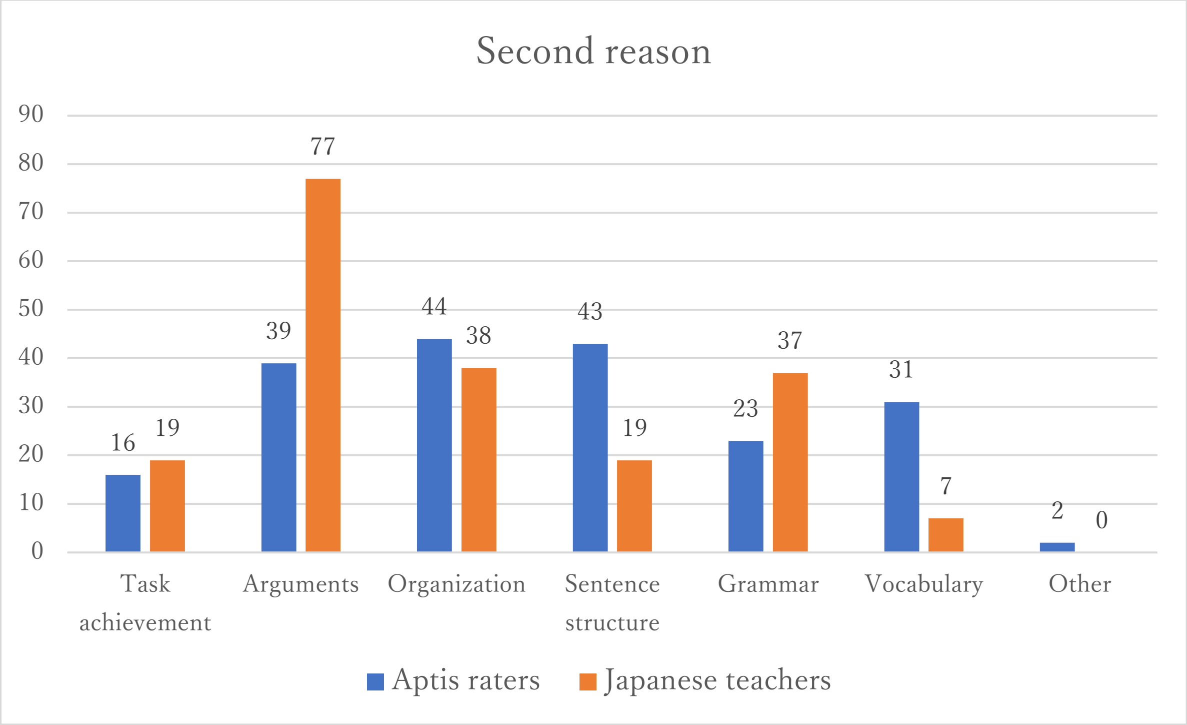
Appendix 17: Raters’ reason for holistic scoring



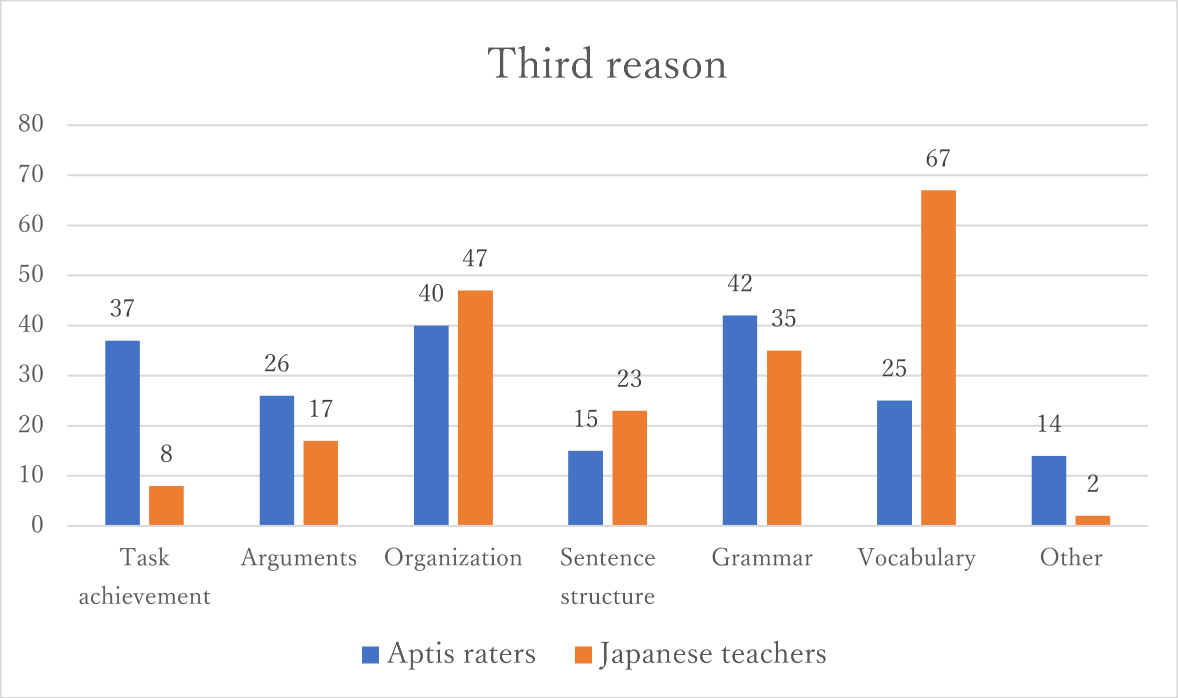
All reasons



First reason



Second reason



Third reason

Appendix 18: Feature ranking of the difficulty of holistic score determination

Aptis raters

	A-A	A-B	A-C	A-D	A-E	A-F	A-G	A-H	A-I	A-J
TA	1	1	4	4	1	5	3	2	1	3
Org	3	2	5	6	2	3	4	1	5	5
SS	4	5	1	2	4	6	5	5	4	6
Arg	2	6	3	5	3	4	6	3	6	4
Gram	5	4	2	1	5	1	1	6	2	1
Voca	6	3	6	3	6	2	2	4	3	2

*1 means “the easiest”, and 6 “the most difficult”

Japanese raters

	J-1	J-2	J-3	J-4	J-5	J-6	J-7	J-8	J-9	J-10
TA	6	1	1	5	6	1	1	1	1	6
Org	2	2	2	6	4	3	5	3	2	4
SS	3	4	4	4	2	4	3	6	6	3
Arg	4	3	1	1	5	5	6	5	3	5
Gram	1	6	6	3	1	6	2	2	5	1
Voca	5	5	5	2	3	2	4	4	4	2

*1 means “the easiest”, and 6 “the most difficult”

Appendix 19: Essay example

Essay 9

I think International sports competitions such as the Olympics help to bring countries together. There are three two I think so that.

First, This is because people all over the world love sports. If we could not talk with people in other countries, we can be excited with them through sports. Regardless of country, we come to want to cheer athlete who are making effI believe sports has a power.

Second, we are able to learn about other countries by going sports competitons taken part in many countries.

In conclusion, I believe international sports competitions help to bring countries together.

Essay10

I think that the International sports competitions such as the Olympic helps to bring the countries together. This thinking is based on my strong belief, which I have learned through doing and facing sport nearly all day. I have two reasons to suport my idea.

First of all, sport and comunication has a really close things to each other. I am going to talk about my experience to discribe this idea more clearly. One day at the tennis club, I was having a hard mach against my friend, who is also important for my private life.First, I was wining but he changed his gears and started to play very good. Soon I became loosing. The macht score had only a tinnny difference, but I lost at the end with my unforced error. Affter the match, I waas really stressed out and had nothing to do. I was half like refusing the handshake after the match. The head teacher got mad to see that. I got a big punishment and have realized how important is to take a polite.