Effects of Task Conditions on Spoken Performance in Retelling

A Dissertation

Submitted to the University of Tsukuba

In Partial Fulfillment of the Requirements for

the Degree of

Doctor of Philosophy in Linguistics

Yuichiro YOKOUCHI

2018

Abstract of the Dissertation

Effects of Task Conditions on Spoken Performance in Retelling

By

Yuichiro YOKOUCHI

English education in Japan is rapidly changing, and the new course of study published in 2017 (Ministry of Education, Science, and Technology in Japan, 2017) emphasizes the importance of integrating four skills. However, the teaching and assessment of speaking skills are typically avoided in Japan due to time and human resource constraints. Therefore, there is a need for developing practical speaking tests for use in classrooms. Retelling tasks are widely used to assess a learner's speaking proficiency; however, the effects of task conditions on the spoken performances have not been examined in detail. In studies conducted by Hirai and Koizumi (2008, 2009, 2013) and Koizumi and Hirai (2010), a practical speaking test called Story Retelling Speaking Test (SRST) and its associated EBB (empirically-derived, binary-choice, boundary definition) scales were developed. The authors were interested in the relationships between the conditions of the retelling tasks and the spoken performances. If task difficulty differs in a particular condition, teachers or test developers can choose the appropriate task condition to elicit test takers' spoken performance effectively.

To examine the effects of the task conditions on spoken performance, four studies and six experiments, including a pilot study, were conducted. Spoken performances were measured in terms of complexity, accuracy, and fluency (CAF) of utterance, and scores were given by trained raters and compared with different task conditions. In the current study, the differences in spoken performances, the length of the text, and the effects of task direction in both retelling and summarizing tasks were observed (Study 1). Subsequently, the effects of the input text length, the text difficulty based on the Flesch-Kincaid Grade Level (FKGL), and the input mode to spoken performance in retelling were compared (Study 2), and the effects of pre-task planning and the pre-task reading-aloud activities on the performance of retelling tasks as main-task were observed (Study 3). The task conditions covered in this series of studies were selected as to whether or not test developers or teachers could easily control the task condition or material. The author considered that the instruction of the task, length of text, difficulty of input material, style of material presentation, and the addition of pre-task planning or a pre-task activity were factors that were easy to change, allowing for a non-specialist of language tests to alter the task difficulty, so those task conditions were focused upon. During the last part of the study, all task conditions were compared in terms of task difficulty using the many-faceted Rasch measurement (MFRM), and the relation of other test facets such as participants, raters, topics of texts, information presentation modes, and criteria were also observed.

The results of Study 1 revealed that the neither the naming of tasks nor differences in task instruction affected the spoken performance. The spoken performances were not changed in retelling and summarizing tasks, but the participants of Study 1 could reproduce the expressions more accurately under the summarizing condition than the retelling condition. This result indicates that task direction affects the use of expressions shown in materials.

Study 2 aimed to observe the effects of text length and difficulty and the effects of presentation mode of input material on spoken performances in retelling (Experiment 2A). The purpose of Experiment 2B was to observe the effects of text length and input mode as well, since there were limitations of the experiment design in Experiment 2A. The results of Experiment 2A revealed that longer text and reading input mode elicited a longer performance. In Experiment 2B, participants were able to perform more words under the reading input condition, while under the listening condition, the participants performed their speaking skills more fluently.

The study 3 consists of two experiments: Experiment 3A aimed to observe the effects of pre-task planning time on spoken performance, and Experiment 3B proposed finding the effects of pre-task reading-aloud tasks on spoken performances. The results of those experiments revealed that one minute of pre-task preparation is effective in eliciting more fluent spoken performances. However, the accuracy and complexity of utterances was unchanged in both the control and experimental groups. This result partially matches previous studies, such as those performed by Skehan (2001, 2009), although the duration of preparation time differed. Experiments 3A and 3B both indicated that pre-task reading-aloud tasks partially affected the performance in terms of length of utterance and fluency.

To synthesize the results of the experiments conducted in this series of studies, MFRM was conducted. The purpose of Study 4 was to examine the difficulties of tasks and text topics. In the first step of Study 4, the validity of the criteria used in this study was examined. The EBB and EBB2 scales (Hirai & Koizumi, 2008, 2009, 2013; Koizumi & Hirai, 2010) were used to assess participant's English speaking performance. The original EBB and EBB2 scales ranged from 1 (low) to 5 (high), but some participant's utterances were so short that the validity of the score 1 was questioned because some participants spoke very few words; namely, their spoken performance was unworthy to be assessed. Therefore, a value of 0 was added to the original EBB scale, and the validity of the new scale was examined. The result of the validity test determined that the new scale was not reliable; therefore, the original EBB scale was applied to this analysis. The result of a three-faceted Rasch measurement (participants × raters × text topics) revealed that the difficulty of text topics split into four groups. On the other hand, another MFRM (participants × raters × tasks) revealed that the most difficult task was the 150-word reading, and the easiest task was the 100-word reading. In Study 2, in which participants responded to a listening stimulus, their spoken performances were lower than when a reading stimulus was given. Therefore, the listening type tasks were expected to be more difficult than the reading type tasks; however, the results of Study 4 revealed that the listening-speaking style tasks were in the middle, being neither the hardest nor the easiest tasks given. Another finding of this research was all tasks used in this series of studies were somewhat difficult for the participants.

Major Abbreviations and Acronyms

ANOVA	Analysis of variance
CAF	Complexity, accuracy, and fluency
DM	Disfluency marker
Er/AS	Errors per AS-unit
FKGL	Flesch-Kincaid Grade Level
FRE	Flesch Reading Ease
GI	Guiraud Index
JACET 8000	JACET List of 8,000 Basic Words
MEXT	Ministry of Education, Culture, Sports, Science, and Technology
MFRM	Many-faceted Rasch measurement
NP	Number of one second or longer pauses
PL	Pause Length
SPS	Syllables per second
To	Number of word tokens
TTR	Type token ratio
Ty	Number of types words
Wd/AS	Words per AS-unit
WPM	Words per minute
95% CI	95% confidence interval

Acknowledgements

I received considerable help from many people during the process of writing this dissertation, providing me with the right support when I felt that I could not complete this study. I learned a lot from their advice, and their encouraging words became anchors to the study. I therefore extend my deepest appreciation to all who supported and guided me. I will be forever grateful.

I would first like to express my deepest gratitude to my supervisor, Professor Akiyo Hirai. She mentored me for six years, and I learned a great deal from her. She also persistently motivated me to complete this study. I started studying speaking tests because I was highly influenced by her work. I learned many statistical methods through her classes, and my research skills were improved through the research activities in the SLAA (the Second Language Acquisition & Assessment Research Group) study group. I am highly indebted to her.

I offer special thanks to the committee members for reviewing this dissertation. I am grateful to Professor Hirosada Iwasaki of the University of Tsukuba, who gave me many comments regarding the weak points of my study. I am also grateful to Professor Akira Kubota of the University of Tsukuba. I became aware of my ignorance in his class, but he attentively guided my study. Without these professors' invaluable comments, I would not have completed this study.

I express my gratitude to my external advisor Professor Hidetoshi Saito of Ibaraki University. His advice was very critical, and his viewpoints as a language testing professional were instrumental to me. His advice and comments on this study were of great help in revising the research design and analysis.

I gratefully acknowledge Professor Yuji Ushiro of the University of Tsukuba. His keen advice regarding my research design became the base of my study. I also express my appreciation to Professor Yuichi Ono of the University of Tsukuba. He encouraged me in many ways, and I benefitted from new points of view about my study, thanks to his comments.

I extend special thanks to Dr. Rie Koizumi, who shared with me her knowledge of statistics, and she encouraged me to complete this study. Her warm comments were surely connected to this achievement. I thank my precious friends, Dr. Shuichi Takaki and Dr. Akira Hamada. They always provided me with important comments about my study, and they encouraged me throughout. Also, the members of the Hirai seminar helped me with my study in many ways. In particular, Dr. Ryoko Fujita always shared important information with me, and her kind advice cheered me up. Thanks to all members of the Hirai seminar, who participated in my study as transcribers or raters and gave me considerable advice in the process. I would not have been able to complete this study without their support; therefore, I am very grateful for their help.

I am highly appreciative of all the participants of this study as well as the colleagues of my current and previous workplaces. Everyone was understanding of my situation, and they kindly encouraged me.

Finally, I would like to show my greatest appreciation to my parents and sisters. They always motivated me to finish this dissertation, and their support was necessary to get to this point. My father showed understanding regarding my will to undertake the Ph.D., which encouraged me to return to academia. My mother and sisters supported me with tenderness and encouraged me when I was up against the challenges. All I can say is that I would not have been able to complete this study without the understanding of my family.

April, 2018

Yuichiro YOKOUCHI

Table of Contents

Chapter 3 Study 1: Comparison of Retelling Tasks and Summarizing Tasks

Chapter 5 Study 3: The Effects of Pre-Task Planning and Reading Aloud During Pre-Task on Spoken Performance in Retelling Tasks

Chapter 7 General Discussion and Conclusion

List of Tables

Chapter 3

Chapter 4

Chapter 5

Chapter 6

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 6

List of Appendices

Chapter 1
Introduction

1.1 Overview of the Current Study

In this series of studies, the effects of task conditions on integrated-speaking tasks, particularly retelling tasks, are carefully observed. In the current English teaching and testing fields, interest in integrated tasks has attracted attention. As Cumming (2013) claimed, “the fundamental purpose of extended writing or speaking is usually to display one’s knowledge appropriately with references to the relevant source information” (p.216). In other words, learners need to collect information or data from reading or listening materials, and to integrate the information obtained to the products in academic courses.

The current situation in the Japanese education system also requires learners to integrate the receptive, i.e., reading and listening, skills with the productive, i.e., speaking and writing, skills. The university entrance examination system, the current uniform paper-based test, and the National Center Examination (hereafter, Center exam), will be officially terminated in 2020, and a new countrywide uniform examination will be used thenceforth. The Ministry of Education, Culture, Sports, Science and Technology in Japan (MEXT) has indicated that commercial English tests should be used as an alternative to the Center exam (MEXT, 2017). The new exam does not include a common English test as does the Center exam, and MEXT has announced that commercial English tests will be used instead of the English test in the Center exam. In fact, there is a three-year transition period; in other words, the Center exam English test will be partially used until 2023, but will certainly disappear in the near future. This reform of the entrance exam means that MEXT has decided to avoid mere cramming, and shifted to training students in authentic English skills; in other words, the four skills. Since the load in terms of assessment of performance skills is enormous, MEXT has decided to use commercial English tests as alternatives to the Center exam. The use of these commercial English tests means that the score users (i.e., the universities) may ask university entrance examinees to take tests that assess the four skills (reading, listening, writing, speaking). Thus far, high-school and cram-school teachers have tended to avoid teaching and assessing speaking skills, and instead focus on the knowledge and strategies required by the entrance exam; namely, vocabulary, grammar, reading and listening comprehension, and short English compositions. However, this change in the university entrance exam system will greatly impact not only universities but also secondary and elementary schools.

The reformation of the Center exam mainly impacts senior high schools, but another educational reform, the change of the course of study, will impact secondary and elementary schools enormously. The current course of study for junior high school in Japan (MEXT, 2008) aims to develop basic communications skills, in other words, the four skills of listening, speaking, reading, and writing (MEXT, 2008, p. 1). However, the next course of study for junior high school was announced in 2017, and it asserts that the integration of each skill is important (MEXT, 2017, p. 129). Furthermore, the next course of study presents five fields, reading, listening, writing, interaction, and presentation, so that speaking skills are divided in two. In fact, the CEFR (Common European Framework of Reference) uses the same categorization, and the next course of study has followed this framework. In the next course of study, teachers are required to train students in the skills required to relate ideas or impressions and to use the information obtained from the listening and reading materials, using easy phrases or sentences. Put differently, it is obvious that the skill of integrating receptive and speaking skills has become more important in secondary education. The next course of study points out the oral summary as an example of an integrated task, and a summary or similar task, for instance retelling, must be used as a pedagogical and testing task in secondary schools. Thus, it is pressing for methods of integrated speaking skills be implemented in practice.

MEXT asked 47 prefectural educational boards to conduct surveys on the implementation status of the speaking assessment in 2015, 2016, and 2017 (MEXT, n.d.). The results of these surveys are very poor, however, with most schools continuing to avoid performance tests as the term test. While two prefectures responded that they have conducted more than 50 performance tests, these responses are typically outliers. The average implementation status of the other 45 prefectures in the speaking test in senior high school was a mere 1.72%. As Fulcher (2003) mentioned, it is difficult to conduct speaking assessments due to issues of practicality. However, English teachers, and researchers of English education and language testing must create a practical method to assess Japanese EFL learners' speaking skills, and provide better feedback to students. The retelling tasks are notable for having the major advantage of controlling the stimuli (Hirai & Koizumi, 2009, p.154). Test takers can tell the story that the read or listened; therefore, their utterances can be related to the stimuli. It helps raters to listen to and understand the utterances and it produces high practicality. Hence, examining the elaboration of task features of the retelling task as a speaking test can have pedagogical implications for English teachers in Japan. Thus, in this dissertation, the effects of task conditions in retelling tasks are studied.

1.2 Organization of This Dissertation

Study 1 aims to compare retelling and summarizing tasks in terms of wording of task and task direction. Subsequently, Study 2, which consisted of two experiments, was conducted. The purpose of this part is to ascertain the effects of task conditions in retelling tasks. Experiment 2A seeks to clarify the effects of input mode, or types of stimuli, the text difficulty, and the length of the text. Due to the limitation in Experiment 2A, the effects of the input mode are ambiguous; therefore, a strict replication study was conducted in Experiment 2B. The purpose of Experiment 2B was to observe the effects of input mode, text length, and those interactions on spoken performances in reading-speaking and listening-speaking integrated tasks; in sum, on retelling tasks. Additionally, the effects of preparation on the performance in retelling tasks were also observed in Study 3. In Experiment 3A, the influence of pre-task planning on spoken performance in retelling tasks was observed, and the effects of a pre-task preparation task, in this case a reading aloud pre-task, were studied (Experiment 3B). Finally, combining and synthesizing those studies into one scale, the difficulties of each task condition were compared using the Many-Faceted Rasch Measurement (MFRM). These are the studies and experiments conducted in this dissertation, and this flow is shown in Figure 1.1.

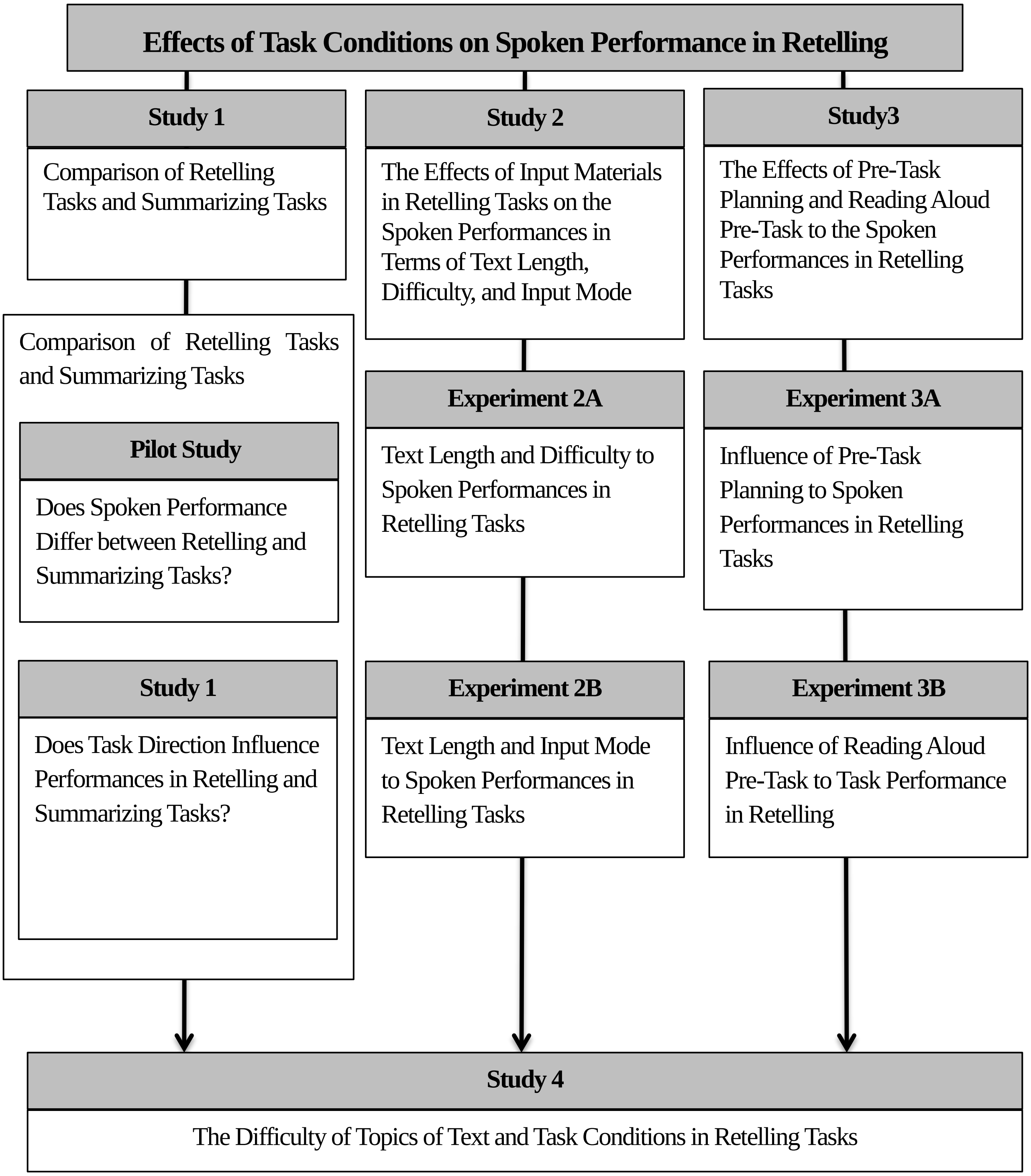

Effects of Task Conditions on Spoken Performance in Retelling
Study 1
Comparison of Retelling Tasks and Summarizing Tasks
Study 2
The Effects of Input Materials in Retelling Tasks on the Spoken Performances in Terms of Text Length, Difficulty, and Input Mode
Study3
The Effects of Pre-Task Planning and Reading Aloud Pre-Task to the Spoken Performances in Retelling Tasks
Comparison of Retelling Tasks and Summarizing Tasks
Pilot Study
Does Spoken Performance Differ between Retelling and Summarizing Tasks?
Study 1
Does Task Direction Influence Performances in Retelling and Summarizing Tasks?
Experiment 2A
Text Length and Difficulty to Spoken Performances in Retelling Tasks
Experiment 2B
Text Length and Input Mode to Spoken Performances in Retelling Tasks
Experiment 3A
Influence of Pre-Task Planning to Spoken Performances in Retelling Tasks
Experiment 3B
Influence of Reading Aloud Pre-Task to Task Performance in Retelling
Study 4
The Difficulty of Topics of Text and Task Conditions in Retelling Tasks

Figure 1.1. The organization of this series of studies.

Chapter 2

Literature Review

2.1 Introduction to Literature Review

This chapter introduces the background of this study. It first discusses the current situation of English education in Japan, which serves as the motivation for the study. Next, it covers the importance of task conditions, the quality of speaking tests, and the factors of speaking assessment. Subsequently, it examines the features of integrated tasks, including retelling tasks, and summarizes the effects of planning and pre-tasks on spoken performance. Finally, it covers use of complexity, accuracy, and fluency (CAF) indices in studies of English education and the ways of calculating these indices in practice.

2.2 Background of Speaking Assessment and Education in Japan

The current English course for Japanese senior high school students was first implemented in 2012. This course advocates that listening and reading skills be integrated with productive skills, such as speaking and writing (Ministry of Education, Culture, Sports, Science and Technology [MEXT], 2009). Integrating these receptive and productive tasks may deepen students' understanding of language and encourage their awareness of language learning. However, teaching and assessing integrated English skills is more difficult than focusing on one skill; furthermore, speaking skills are among the most difficult English language skills to teach and assess (Fulcher, 2003).

According to a survey conducted by the MEXT (2011), half of senior high school classrooms do not administer performance tests, including speaking and/or writing tests. Though this MEXT survey was conducted prior to the enforcement of the current curriculum, it can still be said that English teachers tend to avoid assessing English speaking skills. In addition to this survey, the MEXT and 47 prefectural boards of education open to the public reports on the implementation of speaking assessments in junior and senior high schools every year (MEXT, 2011). The author of this dissertation summarizes the report data and calculates the average implementation status of the speaking tests. The national average frequency of speaking tests carried per term is 3.41 times for junior high and 1.72 times for senior high. In two cases, prefectures reported conducting speaking performance assessments more than 50 times per year in junior high school; however, these results are atypical, and they may include students' peer evaluations or evaluations administered using computer assisted language learning (CALL). Therefore, these two cases were excluded to calculate the number mentioned above. The results of the MEXT survey reveal that teachers tend to avoid assessing speaking skills in their classrooms. When teachers assess students' speaking skills, they need to consider not only the validity, reliability, and washback of their tests, but also test practicality in terms of time costs, human resources (e.g., administrators, interviewers, and raters), and fairness. These practical matters may be one reason for the current lack of speaking assessments in the classroom.

The MEXT is reforming the current system of university entrance examinations, and reports on this reformation suggest that the preliminary government-administered university entrance examinations will be replaced by commercial English tests comprising four skills. As mentioned above, current opportunities to learn and teach speaking skills in schools are not sufficient; therefore, the reformation of university entrance exams is expected to have an enormous impact on not only universities, but also all secondary schools. For this reason, it is urgently necessary to develop and implement easier and more appropriate systems for teaching and assessing the speaking skills of Japanese English as a foreign language (EFL) learners.

The motivation to reform Japan's English education system is to educate students as global human resources. However, Japanese learners' English speaking skills are generally quite low. Ishikawa *et al.* (2009) examined Japanese college students' English speaking proficiency by collecting data using Versant™ and found that the participating Japanese college students typically had English speaking proficiencies at A1 or pre-A1 levels, according to the Common European Framework of

Reference for Languages (CEFR). In the CEFR, an A1 grade reflects a novice level of language learning; therefore, Japanese college students' English language proficiency is less than or equal to that of beginner language learners. Furthermore, Japanese examinees have some of the lowest scores worldwide on the speaking section of the Test of English as a Foreign Language Internet Based test (TOEFL iBT; ETS, 2016). Only 89% of third-year Japanese senior high school students have achieved an A1 level of English language speaking, and only approximately 10% of students exceed the A2 level (MEXT, 2015). Therefore, any speaking tests developed for the field of education should be sufficiently easy to elicit positive student performance.

Learning is closely connected to testing; therefore, testing students' speaking skills is an important part of English teaching. As Brown (2004) stated, "assessment is an integral part of the teaching-learning cycle" (p. 482). However, current research on speaking tests and their practice is not sufficient (Fulcher, 2003). In addition, Negishi (2013) noted that there is no speaking test in senior high school entrance examinations. Makino (2016) further argued that Japan's remedial speaking training courses are insufficient. These studies illustrate the clear lack of English-language speaking study and practice in Japan.

Studies of integrated speaking tasks also indicate that current test tasks are not sufficient. Most studies examine integrated listening–speaking tasks (e.g., Ferris & Tagg, 1996; Frost, Elder, & Wigglesworth, 2011; Murphy, 1991), and few studies focus on integrated reading–speaking tasks. Furthermore, most studies focus on the integrated tasks used in the speaking section of the TOEFL iBT (e.g., Wall & Horak, 2006, 2007, 2008). However, creating test materials, especially audio test materials, as inputs for such tests is challenging for junior and senior high schools, even if they procure help or technology assistance. The author of the present study focuses on the effects of input materials or stimuli on test-takers' spoken performance in integrated reading–speaking and listening–speaking tasks, with a particular focus on the effects of these materials on retelling tasks. The principal merit of integrated and retelling tasks is the ease with which they can be created. In particular, creating stimuli for reading–

speaking retelling tasks is easier than creating stimuli for other integrated speaking tasks, such as picture descriptions, role-playing, and so on. In addition, when creating stimuli for reading–speaking retelling tasks, it is possible to target specific expressions or grammar usages. In other words, test developers can focus on testing test-takers' understanding of particular expressions and rules (Hirai, 2015). On the other hand, other types of tasks are difficult to control. Furthermore, retelling tasks are closed tasks, meaning that the contents of utterances, expressions, and grammar use can all be controlled. Thus, there is potential to use retelling tasks as speaking tests in the field of English education in Japan in order to facilitate the creation of test materials and the assessment of performance. The features, benefits, and drawbacks of retelling tasks are discussed in greater detail later on.

In conclusion, there are several barriers impeding the practical development and implementation of speaking assessments in Japan. However, retelling tasks may help to overcome these impediments.

2.3 Process and Constructs of Speech

2.3.1 Models of Speech and Sub-Skills of Speaking Skills

Speaking is one of the most complex skills that humans learn, and speech and related skills are always developing (Levelt, 1989, p. 1). There are several models of speaking skills, such as Levelt's (1989) L1 speech production model (Figure 2.1) and Bygate's (1987, p. 50) summary of oral skills (Figure 2.2), which includes speaking sub-skills. This section introduces and examines these models and other sub-skills and factors.

Levelt's (1989) model illustrates the basis of the speaking process. It summarizes the four important elements of L1 speakers' spoken production: conceptualization, formulation, articulation, and self-monitoring. Conceptualizing refers to preparing what to say: that is, choosing or deciding on the contents of speech. During formulation, speakers encode their messages grammatically, lexically, and phonologically. Next, during articulation, speakers pronounce and make speech using their knowledge of sounds. Finally, while articulating, speakers monitor whether their spoken language is accurate or not, and this is called self-monitoring. L1 speakers can use the functions of formulation and articulation automatically, but must focus consciously to control their conceptualization. L1 speakers can use these functions simultaneously to create smooth and fast utterances. In addition, L1 speakers have far wider productive vocabularies, which contribute to their fluent and accurate utterances. de Bot (1992) suggested that L2 production operates in essentially the same way as L1 production; however, L2 speakers need to code-switch when saying something in their second language. In other words, L2 speakers have a far greater load of speech production than L1 speakers.

L2 speakers use some of the functions shown in Levelt's (1989) model, but they may also use some strategic functions when speaking in their second language. For example, speakers first need to understand the situation of communication or the presentation of their idea. Therefore, L2 speakers use the function of conceptualization, or a pre-verbal plan. However, L2 speakers also need to leverage their knowledge of their second language, including lexical and grammatical knowledge. Since their knowledge of the L2 is incomplete, they cannot formulate this language automatically. With respect to the articulator function, Roelof (2003) suggested that L2 phonological encoding mechanisms are similar to L1 encoding mechanisms. On the other hand, de Bot (1992) claimed that novice level L2 users often use L1 syllables. Thus, though only parts of Levelt's (1989) speech system are applicable for L2 learning, the function of conceptualization works for both the L1 and the L2, as long as the speech situation is understandable to the speaker.

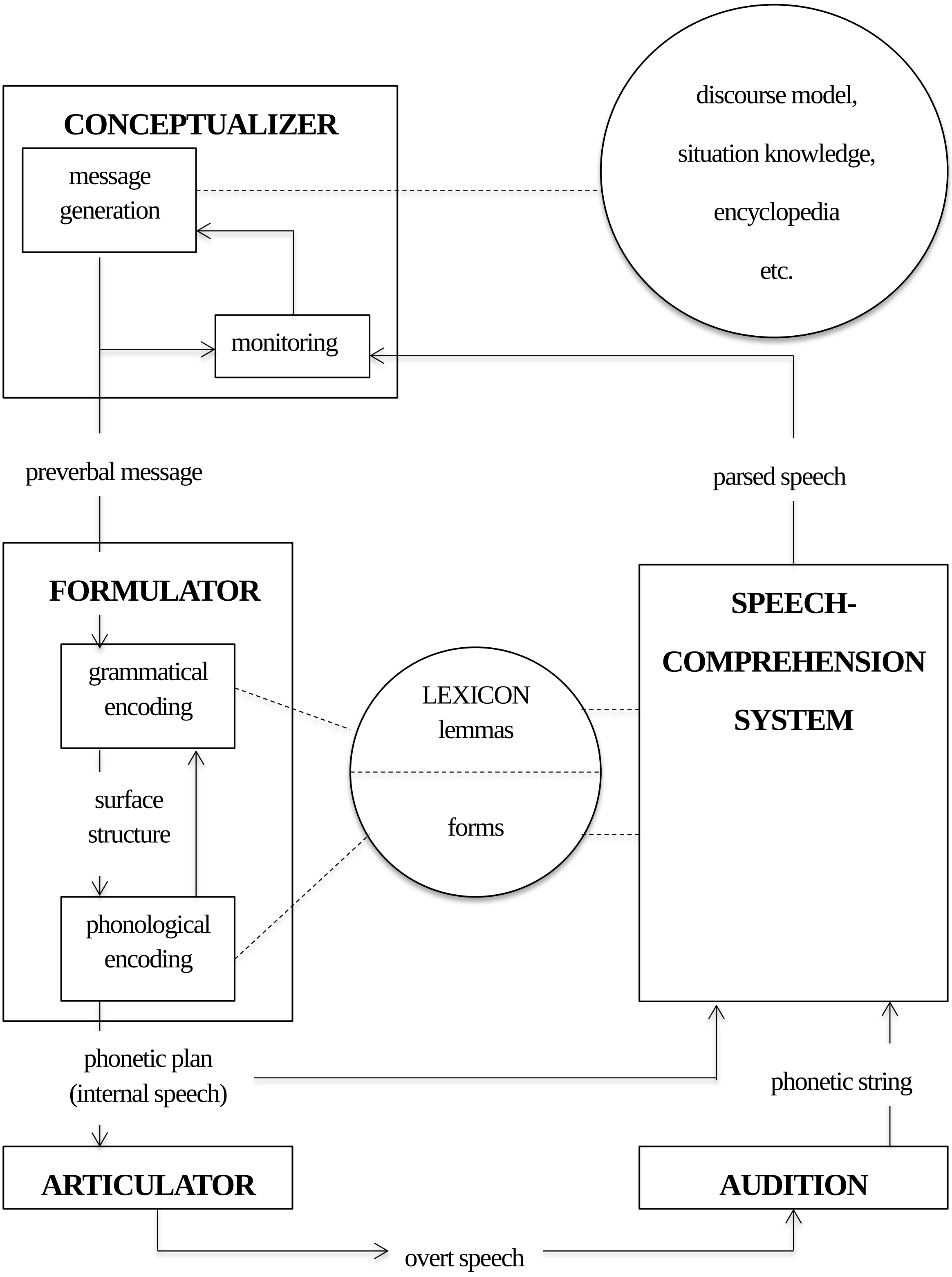


Figure 2.1. Model of L1 speech production (adapted from Levelt, 1989, p. 9).

The most significant difference between L1 and L2 communication is incomplete knowledge of the language. Dorneyi and Scott (1997) indicated that L2 communication suffers four primary issues: lack of information, time pressure for processing, own performance problems, and a lack of understanding of others' messages. These differences in communication skills are based on automatization for formulation and articulation (Kormos, 2006). Both L2 communication and L2 presentation are challenging. In particular, tests have a finite duration, meaning that test-takers must demonstrate their speaking ability in a limited period of time. This time pressure may lead to a one-way speech condition. In addition, as mentioned above, L2 learners lack the same language knowledge as L1 speakers and do not have sufficient performance skills.

Bygate (1987) presented another model of speech production that clearly illustrates the relations between skills and knowledge (Figure 2.2). This model groups speaking skills into two categories—knowledge and skills—and outlines three steps for generating language: planning, selection, and production. Bygate (1987) emphasized that "skills are dependent on some appropriate knowledge resource" (p. 49), meaning that knowledge is formed through the repeated exercise of skills. Therefore, skills and knowledge are closely related and typically developed in parallel, though not always at the same speed. Bygate (1987) also argued that skills are interdependent. When using skills to produce spoken words, speakers use each of the sub-skills shown in Figure 2.2. Using these skills repeatedly can lead to the development of higher-order skills and sub-skills.

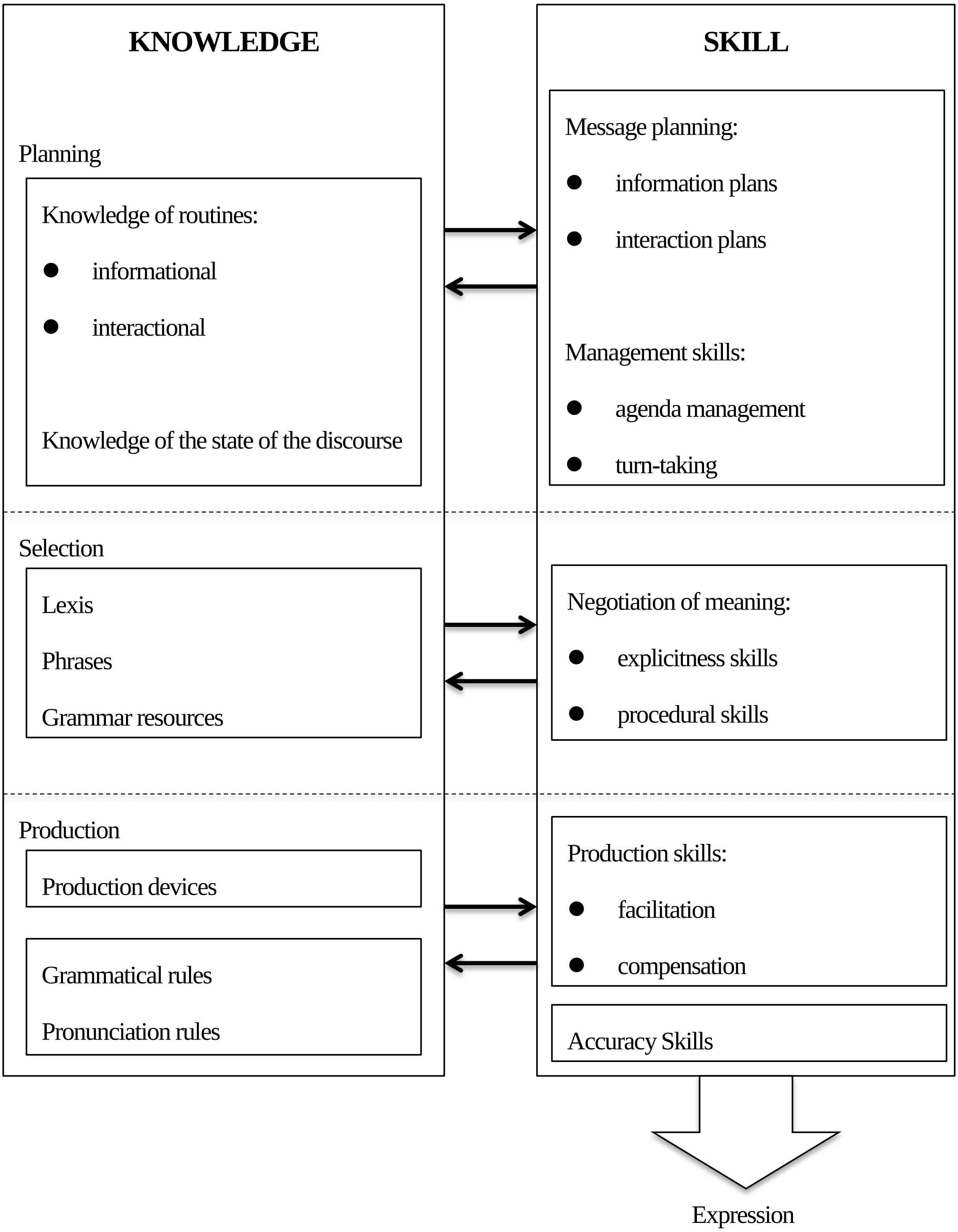


Figure 2.2. Summary of oral skills (adapted from Bygate, 1987, p. 50).

2.3.2 Constructs of Speaking

With respect to language sub-skills, speaking constructs are very important. Speaking skills consist of a number of factors or sub-skills (Higgs & Clifford, 1892), such as fluency (e.g., Brown, McNamara & O' Hagan, 2008), intelligibility, appropriateness, understanding of task (e.g., McNamara, 1990), and vocabulary (e.g., Brown et al., 2008; Douglas, 1994). de Jong et al. (2012) conducted structural equation modeling (SEM) to identify the elements of speaking skills and concluded that speaking skills include "knowledge of grammar, speed of lexical retrieval, speed of sentence building, and correct pronunciation of speech sounds and word stress" (p. 34). These sub-skills typically manifest in test criteria. For example, the criteria of the International English Language Testing System (IELTS) include fluency, coherence, lexical resources, grammatical range, accuracy, and pronunciation. Like the IELTS, other commercial speaking tests also use analytic scales or criteria, which typically include vocabulary, pronunciation, fluency, or other factors. However, these sub-skills are interrelated and none is capable of serving as an alternative to speaking skills. Therefore, all sub-skills are necessary to compose a speaking test.

2.4 Model of Speaking Assessment

As mentioned in the last paragraph, sub-skills are important for creating rating scales. However, task features are also important to consider when designing speaking tests. When test developers create a performance test, they must consider the test constructs. In performance tests, these construct should target test-takers' overall skills. In fact, Brown (2007) argued that "[P]erformance-based assessment implies productive, observable skills, such as speaking and writing" (p. 481). Thus, the main constructs of speaking tests should be *speaking skills*, even if the tests also measure several sub-skills, such as pronunciation, grammar, vocabulary, and/or the contents of the speech. Given this starting point, what kinds of factors should be considered when developing a speaking test? Many researchers have created models of speaking assessments (see Figures 2.3, 2.4, 2.5, 2.6, and 2.7) that focus on performance and the relationships between performance and the other factors that influence performance.

The first factors to consider concern the relationship between test-takers and examiners and/or raters. One of the simplest models of assessment of communication or speaking skills is Underhill's (1987, p. 34) model of recording oral tests (see Figure 2.3). In this model, the speaker or test-taker listens to a recorded stimulus and provides an answer that is recorded. Subsequently, the assessor or rater listens to the recorded answer and assigns a score. This is a typical tape-mediated assessment approach that is generally called a semi-direct speaking test model. Creating speaking tests, test developer must decide which style (e.g. direct, semi-direct, or indirect) to employ. A direct speaking test "include[s] any and all procedures in which the examinee is asked to engage in a face-to-face communicative exchange with one or more human interlocutors" (Clarke, 1979, p. 38). Thus, direct speaking tests involve real conversations with examiners. The examiner may be a rater or another test-taker, and evaluation is typically given immediately. This type of task condition is used in the Eiken second-stage examination, the Standard Speaking Test (SST), and the IELTS.

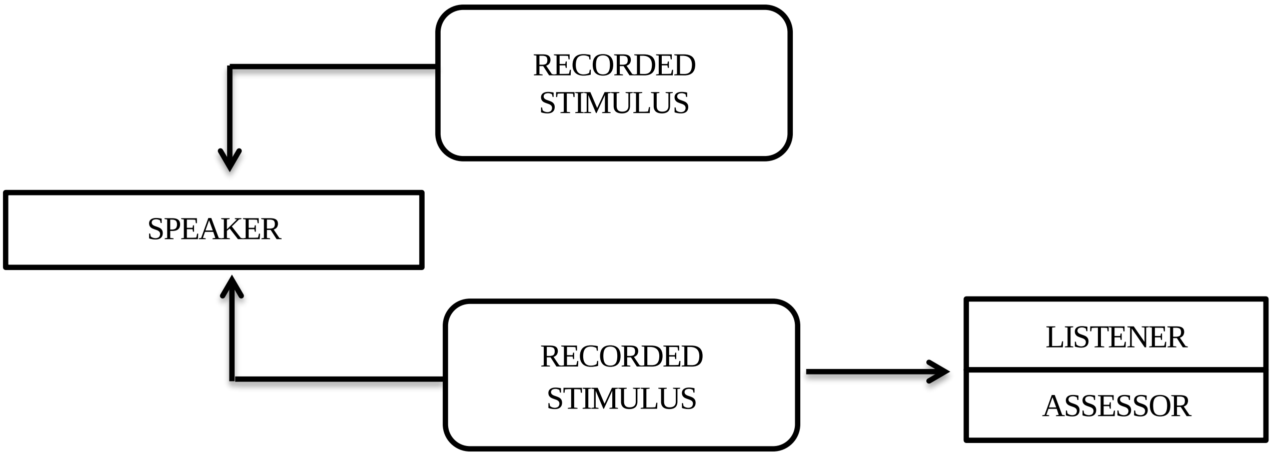


Figure 2.3. Underhill's (1987, p. 34) model of recording oral tests.

On the other hand, semi-direct speaking tests are used in the TOEFL iBT, the Test of English for International Communications (TOEIC) Speaking and Writing Test, Versant, the Telephone Standard

Speaking Test (TSST), the Oral Proficiency Interview-computer (OPIc), etc. Qian (2006) compared direct and semi-direct speaking tests from the perspective of test-takers' affective filters and found that test-takers prefer direct testing to semi-direct testing due to psychological barriers. However, semi-direct speaking tests have certain advantages in terms of practicality. Weir (1993, p. 22) claimed that "there is often great deal of pressure on teachers to make tests as short and hence as practical as possible." Though direct speaking tests are more authentic, may reduce test-takers' affective barriers, and have greater reliability and face validity, semi-direct speaking tests are more practical.

Finally, indirect speaking tests do not require real performances. In other words, they do not require real utterances in order to assess speaking skills. Typical examples of such tests include pronunciation tests, which measure accent, rhythm, or intonation. Shohamy (1994, p. 100) argued that indirect tests are "pre-communicative" and lack authenticity. In sum, indirect speaking test does not match the objective to assess real speaking skills. In this paragraph, the relationships between test-takers and examiners or raters, and the definitions of three types of speaking tests.

As noted above, more than half of commercial speaking tests are conducted in a semi-direct manner, and this style of speaking test is known to have high practicality. Underhill (1987) noted that semi-direct tests have five advantages. First, test-takers can take tests simultaneously, meaning that semi-direct tests can be conducted more efficiently than direct tests. Second and third, semi-direct tests are easy to score and rate. Since raters do not need to give their evaluations in real-time, they can rate test-takers' performances at their convenience. Furthermore, they can listen to recorded utterances repeatedly and stop, rewind, or fast-forward the audio as need; therefore, this style of test is highly practical. Fourth, semi-direct tests do not require real interviewers; therefore, they can be conducted without trained personnel. Finally, semi-direct tests are cost-effective. Thus, there are many practical advantages to using semi-direct tests.

The next step in creating a speaking test is considering the tasks to include. This step is one of the most important aspects of designing a speaking test. McNamara (1997) proposed a complex model of speaking performance assessment, which is shown below in Figure 2.4.

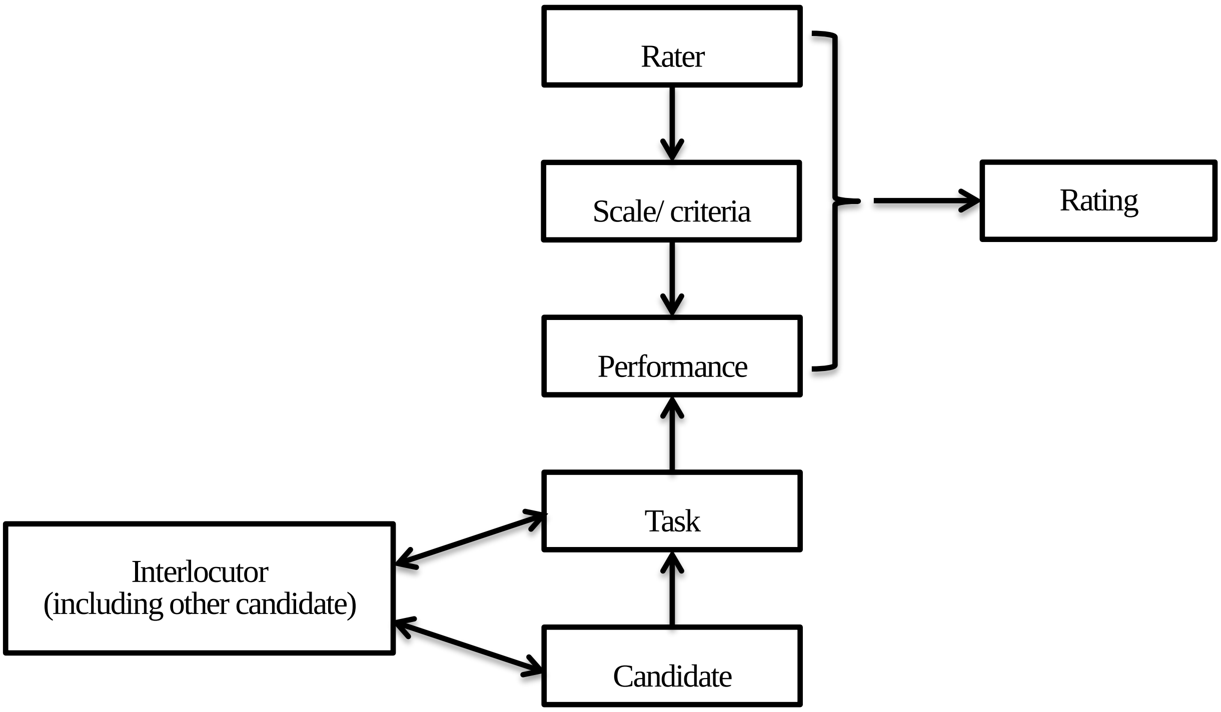


Figure 2.4. McNamara's (1997, p. 453) model of the relations among the various factors in speaking performance assessment.

McNamara's (1997) model is more complex than Underhill's (1987) model of semi-direct speaking tests, and it includes several factors: rater, criteria, task, candidate, and interlocutor. Many of these factors are included in other models (shown in Figure 2.3, 2.4, and 2.5), though some terms differ. McNamara's (1997) model suggests that the task factor is directly connected to performance and interacts with the interlocutor factor. In other words, spoken performance is not dependent on one's ability to measure and is affected by other factors (Van Moere, 2012). The factors of rater, scale or criterion, task, candidate, and interlocutor are all important in speaking performance assessment, and the task factor is affected by the candidate and interlocutor factors. Together, these models clearly show that tasks directly influence performance.

Skehan (1998) proposed an expanded model of speaking assessment that explores the factors of speaking assessment in more detail. In Skehan's (1998) model (Figure 2.5), the task factor is divided into task qualities and task conditions. The model also treats the ability to use dual-coding and underlying competences as factors of speaking assessment; however, these factors are typical factors that cannot be controlled by test developers, administrators, or teachers. Skehan's (1998) introduction of the task quality and conditions factors had a significant impact on later studies, influencing both Bachman's (2002) and Fulcher's (2003) models of speaking assessment. Figure 2.6 shows Bachman's (2002) model, and Figure 2.7 shows Fulcher's model.

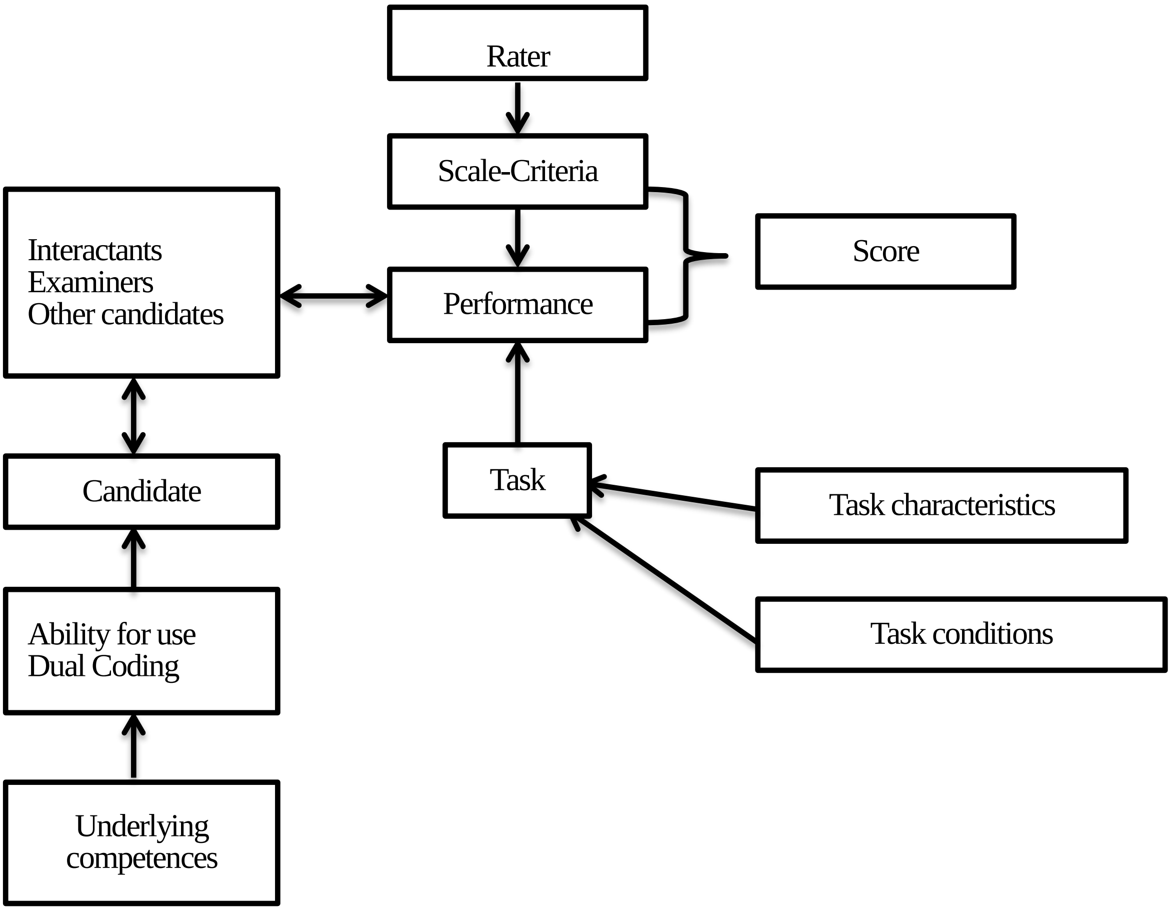


Figure 2.5. Skehan's model of speaking assessment (adapted from Skehan, 1998, p. 172).

Bachman's (2002) model is more complicated than the other models shown above. It includes factors for interactants, candidates, and performance, all of which interact with one another. Moreover, the factors of rater, scale, and performance also interact. There are several differences between Skehan's

(1998) model and Bachman's (2002) model, but the most significant is the interactivity of the rater factor. In Skehan's model, the effects of the rater are unidirectional and impact scale and criteria. Scale, in turn, impacts performance. However, it is natural for raters to be influenced by performance, which may lead to a biased score. On the other hand, the task factor is still unidirectional in relation to performance; thus, in Bachman's (2002) model, this factor is independent.

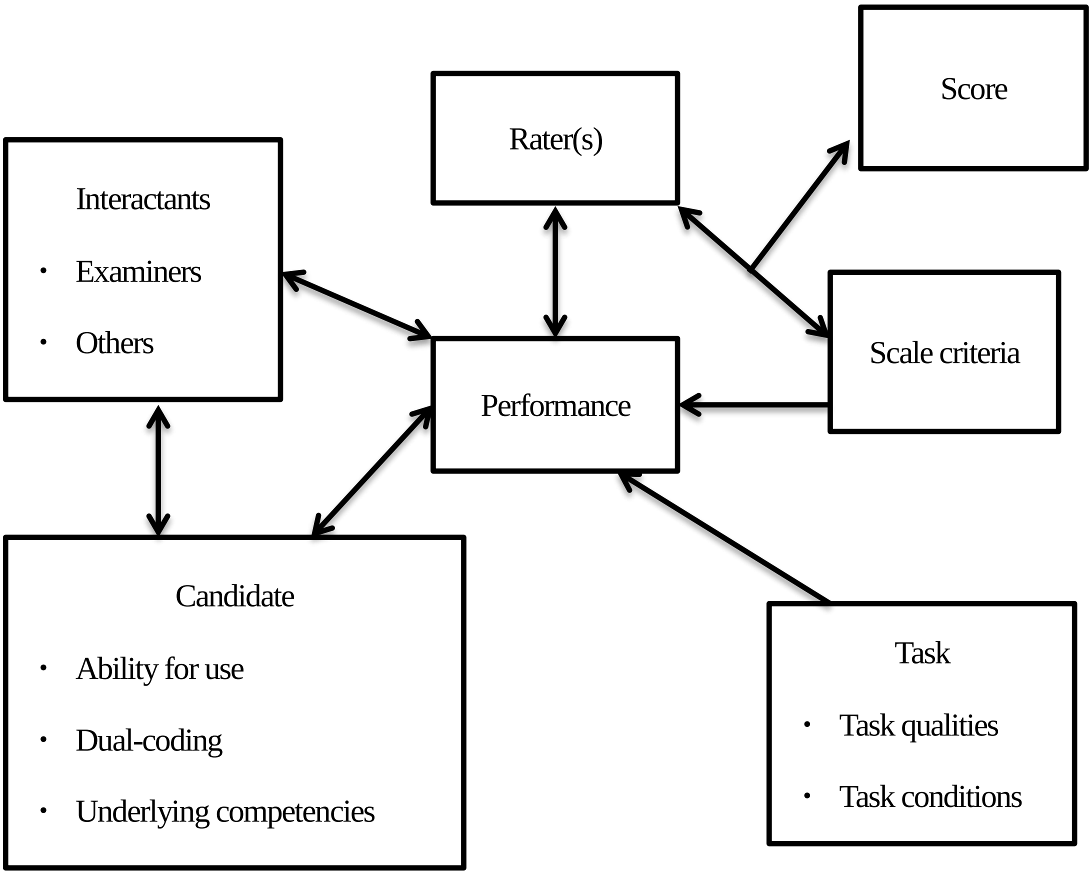


Figure 2.6. Bachman's model for speaking assessment (adapted from Bachman, 2002, p. 467).

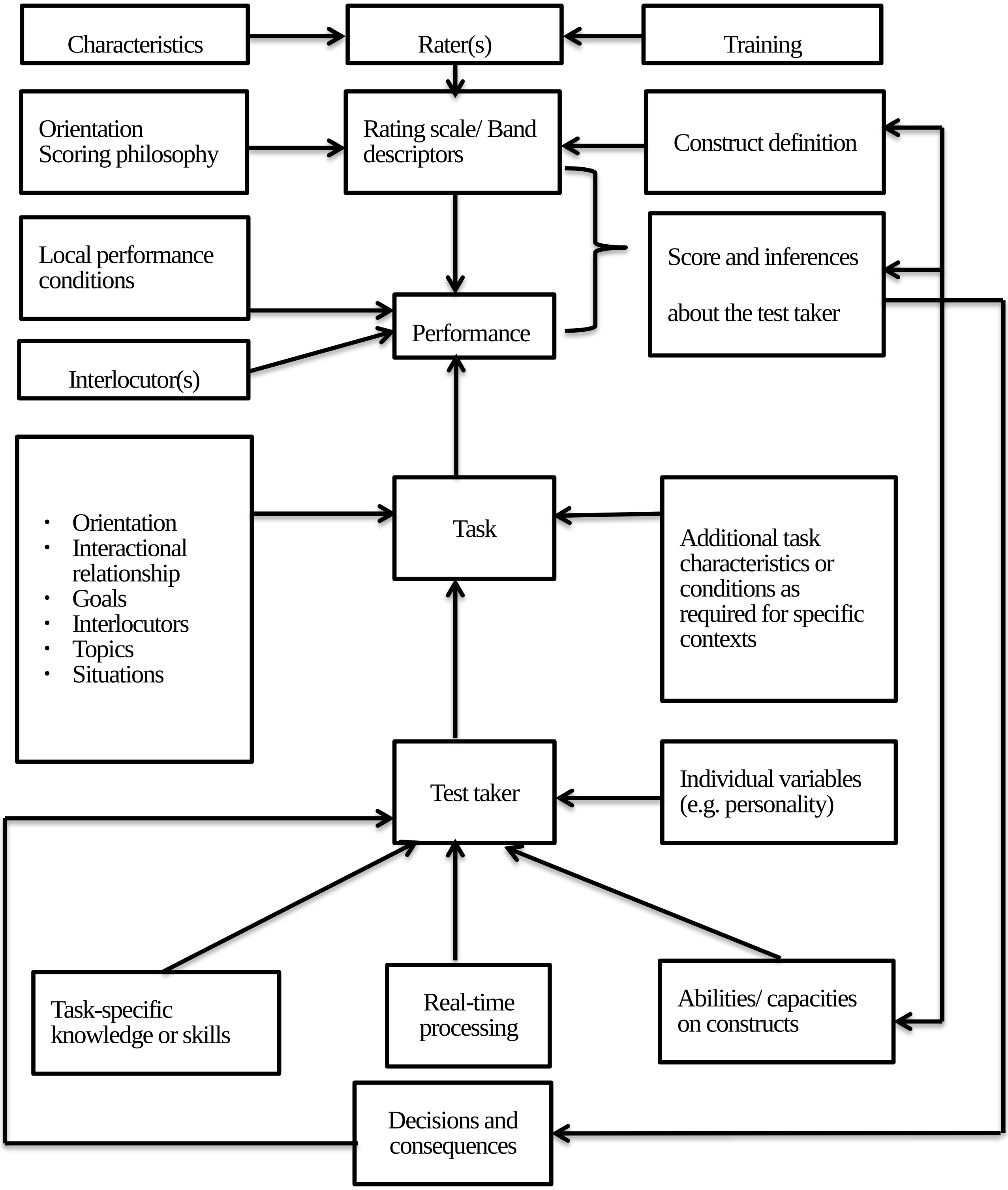

Characteristics
Rater(s)
Training
Orientation
Scoring philosophy
Rating scale/ Band descriptors
Construct definition
Local performance conditions
Score and inferences
about the test taker
Performance
Interlocutor(s)
Orientation
Interactional relationship
Goals
Interlocutors
Topics
Situations
Task
Additional task characteristics or conditions as required for specific contexts
Test taker
Individual variables (e.g. personality)
Task-specific knowledge or skills
Real-time processing
Abilities/ capacities on constructs
Decisions and consequences

Figure 2.7. An expanded model of speaking test performance (adapted from Fulcher, 2003, p. 115).

The newest model of speaking assessment is Fulcher's (2003) model, which Fulcher (2003) positioned as an expansion of Skehan's (1998) model. This model stresses the rating scale and band descriptors as the mainstays of construct definition. Scales and descriptors are important points of any test, not only for test users (i.e., language teachers), but also for test-takers. In addition, Fulcher (2003) claimed that "[t]ask characteristics and conditions still play a role in understanding the meaning of the score, but these are now only part of a larger system and interplay of variables at work in the model" (p. 114).

It is obvious that two-way and multi-way speaking tasks are more difficult than one-way speaking tasks (Fulcher, 2003). As shown in Figures 2.3 to 2.7, interlocutors affect test-takers' spoken performance, introducing variability even in cases of strict examiner training. In other words, though paired speaking tests are more authentic than one-way speaking tests, the interlocutor in a paired speaking test is essentially another test-taker; hence, utterances are not fully controllable. Figures 2.2 to 2.5 illustrate how the factor of the interlocutor affects performance. Test developers who wish to avoid these effects should use one-way speaking tasks.

In sum, there are many factors that must be considered when developing speaking tests, but all of the models introduced above agree on the importance of the task quality, rating scale, and rater factors. Though the factors of interlocutor and test-taker should also be considered, both of these are typically uncontrollable (except in cases in which the interlocutor is the examiner). By contrast, the factor of rater is easily controlled via rater training or rating scales. Given these considerations, it appears that the task factor is one of the most important factors in developing a speaking test.

2.5 Integrated Tasks

2.5.1 Definition, Advantages, and Disadvantages of Integrated Tasks

Recently, researchers and practitioners have emphasized the importance of integrating four English skills (reading, listening, writing, and speaking) in English education. In fact, several commercial tests use integrated tasks, such as summarizing, retelling, and other combined skill tasks. In addition, the next course of study for elementary and junior high school in Japan demands skill integration and emphasizes students’ performance skills. This chapter examines previous studies of integrated pedagogical and test tasks used in real commercial and classroom performance tests. It also explores the task features of retelling and summarizing.

Before examining the various types of integrated tasks, it is necessary to define an integrated task. This term can be defined as follows:

> Integrated tasks are related to “integrative” testing proposed by Oller (1979) to combine components of language ability through items like cloze passages (Lewkowicz, 1997; Plakans, 2012; Yu, 2013); however, integrated assessment focuses more on performance assessments that utilize source material that is read or listened to. (Plakans, 2015, p.1)

Plakans (2013) suggested that all tasks or questions that include two or more of the main four language skills (i.e. reading, listening, writing, and speaking) are integrated tasks. Other sources, such as the integrated speaking and writing tasks in TOEFL iBT, interpret an integrated task as one that integrates receptive and productive skills. Huang, Hung, and Hong (2016) argued that integrated test tasks are more authentic and fair and provide better positive washback to test-takers than independent tasks. From the perspective of authenticity, integrated tasks are far superior to other independent tasks.

The first advantage of integrated tasks is their authenticity. Cumming (2013) argued that integrated tasks are more authentic for academic purpose because university students need to integrate reading and listening skills with writing and speaking skills to fulfill university requirements. University students need to listen to lectures, discuss topics, collect information from many types of sources, and interpret information to present results. Therefore, integrated tasks are truly authentic in relation to English for academia.

The second advantage of integrated tasks is their fairness. Fairness is very important to consider when developing classroom speaking tests because some independent task topics may be unfair to certain test-takers. Plakans (2007) claimed that the performance on an integrated task depends on the stimulus and that integrated tasks may increase fairness because they allow test-takers to perform their language skills using given information. By contrast, in an independent speaking task, some test-takers may have more experience with the task topic than others. Therefore, integrated tasks are fairer than independent ones because they allow test developers or teachers to control performance via stimuli.

The third advantage of integrated tasks is their positive washback effects for test-takers. Learners or test-takers can take expressions from given stimuli and try to use these expressions in their utterances. In fact, Brown et al. (2005, cited in Barkaoui et al., 2013) claimed that "…integrated tasks are likely to diversify and improve the assessment of test-takers' speaking abilities and to lead to improvements in teaching and learning practices" (p. 305). Learners may also derive both content and expressions from input materials that they may use in the future. Therefore, integrated tasks provide positive washback to learners.

There are also several disadvantages to integrated tasks. The most common one has to do with the effects of input. Test takers cannot perform their speaking or writing skills if they cannot read or listen to the stimuli appropriately (Weir, 1990). Another disadvantage is the lack of reliability of integrated tasks. If the input material contains any information or background knowledge, the reliability of the test may decline (Barkaoui, Brooks, Swain, & Lapkin, 2013). However, as mentioned above, the effects of background knowledge are even greater in independent tasks; hence, this drawback represents only a minor demerit for integrated tasks. A final disadvantage is the difficulty of interpreting the results of task performance in integrated tasks (Barkaoui, Brooks, Swain, & Lapkin, 2013) due to the interaction of multiple constructs. Therefore, to use integrated tasks appropriately in the context of education, it is necessary to consider the best way to interpret the results of the test and give feedback.

In sum, there are several advantages and disadvantages to integrated tasks, but the advantages of authenticity, fairness, and positive washback tend to outweigh the disadvantages with respect to conducting speaking tests in the classroom. Moreover, as mentioned in the previous section, paired and multi-way speaking tests are often too difficult for novice learners, and interlocutors may affect spoken performance. Furthermore, as noted above, spoken performance in one-way independent tasks may be influenced by test-takers' prior experiences or knowledge; therefore, such tasks should not be used to assess young learners, whose experience and knowledge are limited.

2.5.2 Task Features of Retelling as an Example of an Integrated Task

Retelling tasks, in which learners or test-takers retell to a partner or examiner a story they have read or listened to (Kissner, 2006), can be used as pedagogical tasks or tests for English speaking skills. Retelling tasks are sometimes used to measure language learners' reading fluency (e.g., Reed & Vaughn, 2012), but they have also been used as test tasks for speaking assessments (e.g., Versant). Furthermore, Story Retelling Speaking Test (Hirai & Koizumi, 2008, 2009, 2013; Koizumi & Hirai, 2010, 2012) was developed as a highly practical tool for test construction and administration in low-stakes or classroom testing tasks. Thus, retelling tasks can be used for both high- and low-stakes speaking tests. Retelling tasks are also used in some junior and senior high school textbooks, and many practice reports published from prefectural education boards use retelling tasks to teach certain expressions or keywords (e.g., Takahashi & Matsumoto, 2015). As Kissner (2006) noted, in retelling tasks, learners' utterances or reactions are based on input materials, which can be provided visually, aurally, or both. Thus, retelling tasks are used as not only test tasks, but also pedagogical tasks in English classes.

Oral retelling tasks should be categorized as integrated tasks. It is impossible for test-takers to reproduce all of the words that they learn from input materials; therefore, they must paraphrase or summarize passages to demonstrate their speaking ability in a retelling task. If the length of the input material is short (i.e. one to several sentences), participants may be able to recall all of the words; however, in such cases, students are not demonstrating their real speaking skills. Retelling tasks are also sometimes used to measure reading ability, and Reed and Vaughn (2012) stated that "[d]epending upon the instrument or study, 'retell' and 'recall' could be used to elicit main ideas, summaries of the content, or a thorough restatement of the passage" (p. 188). Thus, the task features of retelling tasks differ. Klingner (2004) claimed that when learners retell information, recalling the information is the minimum requirement; therefore, the recall task can be categorized as part of the retelling skill.

Retelling tasks are similar to other types of integrated tasks. For example, Kissner (2006) stated that retelling is both different from and similar to summarizing and paraphrasing. One of the task features of retelling is the one-way speaking task, and another is the structural task. When testers ask test-takers to provide their opinions or impressions of input materials, the test features can resemble open-ended tasks. In other words, retelling tasks can be a hybrid of structural and open-ended tasks. Learners must express what they have read or listened to using the information in the input text; hence, their utterances should be similar. Accordingly, retelling tasks can be categorized as structural tasks.

Retelling tasks are used as sub-tests in the high-stakes speaking proficiency test Versant. In Versant, test-takers listen to an input passage via telephone or PC and must then retell the information they have heard within 30 seconds. Note taking is not allowed while taking the test, meaning that test-takers must memorize contents, keywords, and expressions. On the other hand, the Story Retelling Speaking Test (SRST), which was developed by Hirai and Koizumi (2009, 2013) and Koizumi and Hirai (2010, 2012), is a representative example of a practical speaking test that can be used in a classroom setting. The task procedure of the SRST is as follows: (1) test-takers read the text (approximately 100 to 150 words) within two minutes without reading aloud or taking notes; (2) test-takers answer three or four questions designed to enhance their comprehension of the given text and store the context in their mind; and, finally, (3) test-takers retell the story and provide ideas or opinions related to the given texts within two and a half minutes (Koizumi & Hirai, 2010). Thus, the factors of text length and input mode or material are of great importance when implementing this test in practice.

When providing input materials, teachers can change the input mode (e.g., reading condition, listening condition, or mixed condition), text length, and text difficulty and can also add pre-tasks or planning time. Knowing the influence of these factors on test-takers' performance is important for eliciting speaking performance; without them, test-takers cannot perform effectively.

Retelling tasks draw upon several language features, such as pronunciation, productive vocabulary, and communicative efficiency. These variables measure oral performance quality and are important for language learners' performance skills. Therefore, comparing typical tasks might yield useful information for implementing these tasks in educational settings. For example, Koizumi and Hirai (2012) reported that "the SRST tends to elicit relatively long utterances, even from lower proficiency students" (p. 48). Furthermore, the SRST distinguishes between beginner and intermediate L2 learners. However, Koizumi and Hirai (2012) also reported that the discriminatory power of the SRST might vary depending on the difficulty of the input text. Thus, considering the length and difficulty of the input and understanding how the information presentation mode may affect the utterance are important when designing speaking tests using retelling tasks.

Input materials typically differ in terms of text length, text difficulty, and input mode. These factors affect input comprehension, which can, in turn, affect test-takers' utterances. Koizumi and Hirai (2010) reported that, "a 50-word difference did not seem to consistently affect the output volume" (p. 216), even though they had expected that "longer texts may elicit more production than short texts, because longer texts include more information to retell" (p. 214). Hirai and Koizumi (2009) briefly discussed text difficulty, claiming that "the difficulty of the speaking task may not directly arise from the passage difficulty but rather from the nature of the speaking task itself" (p. 161). However, this claim was based on test-takers' answers to questionnaires, rather than indices of actual performance. Therefore, room for investigation remains.

Possible input modes for retelling tasks fall into three categories: visual input, aural input, and mixed input. Versant employs retelling tasks that use aural input, the SRST uses visual inputs, and the TOEFL uses mixed input tasks. However, though retelling tasks can use several different types of inputs, no inclusive study has yet addressed the presentation mode of the input.

Retelling tasks are practical for assessing learners' speaking skills because of their ease of material development, test administration, and rating. Bachman and Palmer (1996) mentioned that human, material, and time resources are all critical to test practicality. With respect to practicality, retelling tasks have a strong advantage, since they make it easy for teachers or test developers to develop and control the input material. In addition, retelling tasks are monologue and one-way tasks, meaning that learners can take tests based on retelling tasks without requiring an interviewer or interlocutor. Furthermore, test administrators can record test-takers' utterances using recording devices, allowing them listen to the speech data repeatedly. Finally, retelling tasks are authentic due to their mixture of reading comprehension and speaking skills (Underhill, 1987). Therefore, retelling tasks offer advantages related to both practicality and authenticity.

However, there is one controversial point regarding the use of retelling tasks in a classroom setting. As mentioned above, such tasks are somewhat difficult for novice learners, such as junior high and senior high school students. The *ELP Can Do Descriptor Database* (Tono et al., 2013) ranks the skill of telling a story or telling the plot of a book or movie at the B1 level, meaning that it is most appropriate for independent learners. Hence, teachers may need to adapt some features of retelling tasks to adjust their difficulty for use in the classroom.

2.6 Task Condition and Task Performance

2.6.1 Effects of Pre-task Planning on Spoken Performance

As mentioned above, a task condition and quality can control performance, and one of the most effective and easiest methods of changing a task difficulty is including preparation time before the main task. Many researchers studied the effects of preparation on performances related to not only speaking but also writing, and the effects of pre-tasks on the performances of main tasks were considered (e.g., Foster & Skehan, 1996: Mehnert, 1998; Yuan & Ellis, 2003). These studies indicate the positive effects of pre-task planning on language production, even though positive effects were found in only some perspectives. Rausch (2012) cited Skehan (2001, 2009), stating that pre-task planning has a significant impact on complexity and fluency, with a minor impact on the accuracy of production. In this section, the previously studied effects of pre-task planning on performance are reviewed.

Foster and Skehan (1996) found that pre-task planning affects L2 learners' spoken performance in terms of complexity, accuracy, and fluency when the participants were assigned ten minutes to prepare an answer to an oral task. Mehnert (1998) studied the relation between the length of preparation and the oral performances of EFL learners, as well as compared spoken performances after different lengths of preparation: in concrete; unplanned; and one-minute, five-minute, and ten minute preparation times. The results of this study indicated that the participants who allotted ten minutes of preparation time performed their spoken skills more fluently, and they used more complex structures than the other groups. On the other hand, a short preparation time did not work well; in other words, the performances of the groups that allotted one minute and five minutes of preparation time did not differ largely from those of the non-prepared group. However, ten minutes of pre-task preparation reduces authenticity, because people typically do not have this much time before they respond to another person; in sum, ten minutes of preparation leads is inauthentic. Yuan and Ellis (2003) studied the effects of pre-task planning and online planning on L2 spoken performances, and they revealed that pre-task planning elicits a more fluent performance; however, the accuracy of the oral performances was unchanged. In addition, participants who were assigned preparation time could use many words than the unplanned group of their study.

Therefore, it can be said that pre-task planning is effective in eliciting a much better performance in terms of fluency, and it increases complexity, both lexically and syntactically. However, the accuracy of the spoken performance was unchanged in some studies, even if a long (ten minutes) preparation time was given.

2.6.2 Effects of Reading Aloud on Speaking Performance

Reading aloud tasks are widely used in speaking assessment situations, such as the EIKEN second stage interview, the first part of the Versant English Test, etc. Generally, the reading aloud task is used to assess test takers' pronunciation and reading fluency (Shimizu, 2009). Furthermore, some researchers studied the effects of reading aloud tasks on improvements in reading, listening, and speaking skills. Iino and Yabuta (2013) summarized previous studies that focused on the effects of reading aloud on improving speaking skills, and they noted that few studies report the effects of reading aloud tasks on improving English learners' speaking proficiency, but the results of those studies are based on claims or advice from successful learners. Some studies have reported that reading aloud tasks contribute to improving receptive skills, such as reading or listening (e.g., Takahashi, 2013; Takayama, 2007), but the effects of reading aloud tasks on the performance of oral production skills have not been studied frequently.

The reading aloud test in the Versant English Test is independent of other sub-tests, and this task has no relevance to other tasks. On the other hand, the EIKEN second stage interview in the second grade asks test takers to read the input material silently within 40 seconds and to read the text aloud once. Test takers are required to respond to a number of comprehension questions after reading the text aloud. While test takers are reading the text, they can store the information and prepare for the subsequent questions. This reading task is also independent of other tasks within the same test set; however, the reading aloud task may aid in test takers' understanding of the text and in their production of utterances.

2.7 Measuring Speaking Skills Using Indices of Complexity, Accuracy, and Fluency (CAF)

In the current study, the indices of complexity, accuracy, and fluency (hereafter, CAF) are used to determine the effects of each task's condition on spoken performance. To elucidate the basic information of the CAF indices and the necessary indices for calculating CAF, the definition of each construct and index will be introduced in the following sections.

To observe EFL learners' speaking performances, the CAF variables can have significance. Van Lier (1989) emphasized the importance of analyzing actual speech to validate the speaking test. However, the consistency of the CAF indices used in studies in the fields of language or linguistics is insufficient. In fact, various indices have been used to observe CAF. For example, Housen, Kuiken, and Vedder (2012) summarized previous studies that use CAF measures, and they claimed there is no agreement among studies that use these indices. Deciding which CAF indices will be utilized in a study is quite important, and considering the consistency of indices with previous studies is necessary before starting a study. In the following section, the CAF indices used in previous studies are introduced.

2.7.1 Analytical Units

To analyze L2 learners' utterances in terms of CAF, researchers must first transcribe and tag their utterances, for example, with disfluency markers. To analyze accuracy and fluency, utterances should be divided into analysis units, such as the analysis of speech (AS)-units, T-units, C-units, or idea units. Foster, Tonkyn, and Wigglesworth (2000) categorized these analysis units into semantic, intonational, and syntactic units. However, they stated that some of those indices could not be applied to speech analysis, proposing that the AS-unit is one of the most effective units for analyzing speech. They defined an AS-unit as "a single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either" (Foster, Tonkyn, & Wigglesworth, 2000, p. 365). Thus, some incomplete sentences or clauses can be counted as one unit. An AS-unit has another restriction; if there is 0.5-second or longer silent pause between conjoint sentences, the AS-unit should be separated into different AS-units. However, this criterion is quite severe for L2 learners, because they tend to use many long, silent pauses in their utterances.

2.7.2 Complexity

Bluté and Housen (2012) gathered previous studies that use indices of complexity, and they claimed that the category of complexity could be divided into grammatical and lexical complexity. Grammatical complexity consists of the sub-categories of syntactic and morphological complexity, while lexical complexity can include lexical diversity, density, and sophistication. Syntactic complexity can be measured as words per clause, sub-clause, or analysis unit (e.g., words per AS-unit, words per clause); as clauses per analysis unit (e.g., clauses per AS-unit, subordinate clauses per clause); as subsentential (e.g., mean length of clause, s-nodes per clause); and so on. On the other hand, in terms of lexical complexity, lexical sophistication is used infrequently, but the indices of lexical diversity and density have been applied in many studies. Typical examples of diversity include the number of word types, type–token ratio, or Guiraud's Index. Meanwhile, density can be calculated as the number of lexical words per function word or total words. As stated, there are various types of complexity indices; therefore, choosing the appropriate index is quite important to allow researchers to extract useful implications.

2.7.3 Accuracy

Analyzing the accuracy of EFL learners' speaking skills is one of the most difficult and important aspects of CAF analysis to provide students with corrective feedback. Usually, the number of errors and analysis units are counted, and accuracy can be measured as follows: the number of errors per number of analysis units or clauses (e.g., errors per AS-unit, or errors per clause unit) or the number of error-free analysis units (e.g., the number of error-free AS-units, error-free clauses per all clauses). Creating a clear definition of error is the most important aspect of an accuracy analysis. For example, Kuiken and Vedder (2011) classified three types of errors: minor errors, including meaning and grammar, which did not interfere with the listener's understanding; errors of meaning and grammar that are more serious; and errors that are close to incomprehensible. However, categorizing the errors into the syntactic, semantic, phonologic, or mixed types is quite difficult. In particular, discriminating between error types in novice-to-intermediate-level student performances is significantly difficult, because some errors can be extended to different domains.

2.7.4 Fluency

According to Tavakoli and Skehan (2005), there are three types of fluency: breakdown fluency, speed fluency, and repair fluency. Housen, Kuiken, and Vedder (2012) provided a brief definition of those three sub-dimensions of fluency. They state that the number of words, syllables, or other kinds of countable indices represents the breakdown fluency; speed fluency is the rate of words per minute or syllables per second, the length, or the ratio of pauses; and repair fluency is the number of false starts, reformulations, and repetitions. To observe real fluency, all sub-dimensional fluency indices should be analyzed. However, the definition of a silent pause in terms of the duration of silence is still interrogation. Kormos (2006) summarized the methods of fluency measurement, as well as provided definitions of some indices of fluency measurement. He (Kormos, 2006) cited Riggenbach (1991), stating that the definition of speech rate includes information about the duration of a silent pause, where the cut-off of an unfilled pause is three seconds. On the other hand, the definition of the mean length of pauses is "[t]he total length of pauses above 0.2 seconds divided by the total number of pauses above 0.2 seconds" (Kormos, 2006, p. 163). As previously described, the durations of silent pauses are not consistent in each index of fluency measure index. Thus, the length of a silent pause that uses a measure fluency is inconsistent; in addition, 0.2 seconds is too short for EFL learners. Onoda (2013) collected spoken data from Japanese learners of English, and he defined the length of a silent pause as one second; therefore, it is most probable that the length of a silent pause can indicate the fluency of Japanese learners of English.

2.8 Summary of Previous Studies

In this chapter, the literature of the current English education situation in Japan; the processes of speech; the five models of speaking assessment and the importance of task condition and quality; and the characteristics of integrated tasks, mainly retelling, the effects of reading aloud, and the concepts of

CAF, are introduced.

The current Japanese Education system is reforming to focus on teaching productive skills, and the government now requires junior and senior high schools to integrate four skills, in addition to teaching productive skills. Moreover, the reformation of entrance examinations for universities is now changing to involve commercial tests that include a speaking test. However, most junior and senior high schools do not conduct speaking tests, and the speaking proficiency of secondary school students is quite low. To overcome this situation and to provide effective practices of speaking and assessment, a practical speaking test is needed.

When beginning studies of speaking assessment, it is quite important to consider the procedures of producing a spoken performance, and in this study, Levelt's (1989) and Bygate's (1987) models of speech production were introduced. The functions of speech processes are divided into conceptualizer, formulator, speech comprehension system, and articulation functions. L2 learners have a limited mental lexicon and knowledge of grammar or expressions; therefore, L2 English learners tend to produce their ideas slowly. In addition, oral models are largely divided into knowledge and skills, and speech processes are divided into planning, selection of words or grammar, and production (Bygate, 1987); all of these interact while the speaker produces their speech, and all are necessary for spoken performances.

In the next section, five models of speaking assessments are introduced, and the importance of task condition and quality are emphasized. The models of speaking assessment place task as central to the model, which means that considering task condition and quality, as well as rating scales, is quite important when developing a speaking test.

Following the above, the concepts of integrated tasks and the task features of retelling were introduced. Integrated tasks are currently developing in importance, because the new course of study for junior high schools in Japan highlights the importance of skill integration. The advantages of integrated tasks were covered in this section, and these include positive washback, fairness to test takers, and authenticity. Subsequently, the effectiveness of retelling tasks as speaking tasks is summarized, where retelling tasks are highly practical, and the possibility of changing the difficulty of retelling tasks is mentioned. Subsequently, the functions of two treatments for changing the difficulty of a task, pre-task planning and pre-task reading aloud, are discussed. Earlier studies that dealt with pre-task planning revealed that preparation is effective to elicit a greater performance, and a longer preparation time draws more fluent and complex utterances; on the other hand, the accuracy of the spoken performance is unchanged in some studies.

Finally, the concepts of CAF measurement and typical examples of each CAF index are introduced. In the next chapter, how those CAF indices are measured in the current study is summarized.

2.9 Relations Between Literatures and Following Studies

As mentioned in this chapter, the needs for research on task features and implications and for the development of a practical speaking test that can be conducted in classrooms are outlined. The author is interested in the effects of task conditions on retelling performances; if spoken performances were controlled by a change in task condition, it would lead to significant implications related to the design of speaking tests in the classroom. In the current study, the effects of different task conditions in retelling tasks on spoken performances are studied; in addition, the differences in task difficulty that are dealt with in this study series are examined. This study series shares some methods of analyzing CAF indices; indeed, some indices are not used or changed in different experiments. In the following section, the methods of this study are introduced.

2.10 Common Methods of Analysis in This Study

To collect the spoken performance data, IC recorders or computers with headsets were used in this study series. In all cases of data collection, other participants' voices were recorded as background noise, and some voice data were unusable because of equipment malfunctions. If other participants' voices were recorded loudly in other recordings, the automatic detection of silent pauses was not used, and the author and research collaborator aimed to calculate the length and number of pauses or other operations. If the background noise was not loud, the automatic silence detection tool Silence Finder, an expansion tool of Audacity® version 2.0.5 (Audacity Team, 2013[computer software]), was used.

Transcribers of each experiment listened to the original recorded data or noise-reduced data repeatedly, and they transcribed the utterances. In the second phase of data manipulation, transcribers created pruned versions of the transcriptions. Next, the AS-units were counted by the transcribers, and if there was a disagreement among the transcribers regarding the number of AS-units, the transcribers discussed and assimilated the number and locations of cut-lines of AS-units. After completing these treatments, a CAF analysis was conducted.

In the first phase of the CAF analysis, the numbers of word tokens, types words, and syllables were counted using the Syllable Counter & Word Count (wordcalc.com, n.d.), and those figures were used to calculate the indices of complexity and fluency.

2.10.1 Analysis of Complexity

The complexity of participants' utterances was analyzed during the first phase of the CAF analysis. To analyze lexical complexity, the numbers of tokens and types were counted using the JACET 8000 analysis program—*v8an* (Shimizu, 2004); subsequently, the indices of the type–token ratio (TTR) and Guiraud's Index were calculated. There is another viewpoint in terms of the analysis of complexity, and the number of words per AS-unit was used to ascertain syntactic complexity.

2.10.2 Analysis of Accuracy

Regarding the analysis of accuracy, syntactic accuracy was analyzed, and the index of the number of errors per AS-unit was used for this analysis. Errors were defined as global errors, the expressions that led to listeners' misunderstandings. In other words, incomplete sentences and sentences that did not have sufficient information to understand or guess what the test taker wanted to say were categorized as global errors. In this study, local errors, such as a missing "s" for plural words and third-person singular present, were not treated as errors. In this analysis, the AS-unit (Foster, Tonkyn, & Wigglesworth, 2000) was used as a unit of analysis, where an AS-unit is defined as "a single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either" (Foster, Tonkyn, & Wigglesworth, 2000, p. 365). If the unpruned transcriptions were used for the analysis, too many errors, including disfluency markers, would have been counted. The author wanted to focus on what the test takers wanted to say; therefore, the author followed previous studies (Iino & Yabuta, 2013; Iwashita, Brown, McNamara, & O'Hagan, 2008) by using the pruned transcription for the analysis.

2.10.3 Analysis of Fluency

Up to five indices were used to examine changes in the test takers' fluency, where fluency consists of three dimensions, as referred to in section 2.6. The words per minute (WPM) and syllables per second (SPS) variables were treated in this analysis as speed fluency, and the lengths of pauses and the number of one-second or longer pauses were treated as breakdown fluency. In addition, the number of disfluency markers was treated as repair fluency in this study.

2.10.4 Rating

The performances obtained from each experiment were rated by trained raters, all of whom have enough proficiency to rate the spoken performances of Japanese leaners of English. The details of each rater's background are shown in each methods section in all chapters of the studies. The criteria were adapted from Hirai and Koizumi's (2008) empirically derived, binary-choice, boundary definition (EBB) scale, as well as from Hirai and Koizumi's (2013) EBB scale. All criteria and documents used for rater training are attached in the Appendix N, O, P, and Q.

Chapter 3

Study 1: Comparison of Retelling Tasks and Summarizing Tasks

3.1 Introduction to Study 1

As noted by Kissner (2006), retelling and summarizing tasks are very similar, and even veteran language teachers cannot always clearly distinguish between them (Kissner, 2006, p.4). As integrated speaking test tasks are the main focus of this section, I will begin study 1 by providing a definition of retelling and summarizing tasks, respectively. The definition of *retell* in the Oxford Dictionary of English version 3.3 (Oxford University Press, & WordWeb Software, 2016 [computer software]) is "tell (a story) again or differently." It could be argued that learners or test takers could simply *reproduce* the original sentences, even if it is recommended that the original sentences be *rephrased*. The directions provided regarding how to respond to a task (henceforth, "task direction") can change the property of the response utterances to reproducing, rephrasing, or summarizing (Hirai, 2015). On the other hand, the definition of *summarize* is to "give a brief statement of the main points of (something) (Oxford University Press, & WordWeb Software, 2016 [computer software])"; that is to say, the summary should be shorter than the original text. However, these are merely differences in wording and, in reality, test takers and test developers cannot distinguish the differences between such tasks. What, then, are the differences between speaking tasks that require retelling and summarizing respectively? The author hypothesized that performances in oral retelling and summarizing tasks are homogeneous due to their similarity. To confirm this hypothesis, a pilot study was conducted to ascertain the differences in performance between the two task types. The effects of task direction on performances in retelling and summarizing tasks were then examined (Study 1).

3.2 Pilot Study: Does Spoken Performance Differ between Retelling and Summarizing Tasks?

3.2.1 Objective and Research Questions

The purpose of the pilot study was to compare spoken performances between retelling and summarizing tasks. Two research questions were designed to the observe the effects of naming of task.

RQ1A-1: Do spoken performances differ between retelling tasks and summarizing tasks?
RQ1A-2: How do the test takers and raters of this experiment interpret retelling and summarizing tasks?

3.2.2 Participants

Twenty-nine Japanese learners of English participated in this experiment. Of these, the data from 21 were used to compare the differences in spoken performance for retelling and summarizing tasks; the data of 8 participants were excluded due to recording failure. The study participants comprised electronic engineering majors, all of whom had completed liberal arts English courses in the first years of their undergraduate study. Nineteen of the 21 participants were sophomores, 1 was a junior undergraduate, and 1 was a graduate student. In the interests of time, only 10 participants engaged in the post-experiment interview, and they were asked to give their impressions of the retelling and summarizing tasks. In addition, two raters who had majored in English education participated in the rating phase of this study, and they were also interviewed.

3.2.3 Materials

Two descriptive texts were used for the pilot experiment. Material 1A-1 comprised a 149-word story on local production for local consumption. The text difficulty, in other words the Flesch-Kincaid Grade Level as the index of readability, was 7.8. Material 1A-2 comprised a 153-word story on life happiness with a Flesch-Kincaid Grade Level of 6.7. Detailed information concerning the input materials used in experiment 1A is shown in Table 3.1.

Table 3.1

Information on the Input Materials Used in Experiment 1A

	Text Length	FKGL	FRE	Level 1 words (%)
Material 1A-1	149	7.8	63.0	78.5
Material 1A-2	153	6.7	67.7	83.0

Note. JACET8000 Level 1 shows lexical diversity. FRE = Flesch Reading Ease. FKGLs and FREs were computed using Microsoft Word 2010; all materials are shown in the Appendices.

The empirically derived, binary-choice, boundary-definition (EBB) scale developed by Hirai and Koizumi (2008) was used as the criteria for assess the spoken performances in this study, in addition to the holistic scale based on the ACTFL Proficiency Guidelines 2012 for speaking (ACTFL, 2012). Although the ACTFL guidelines have ten levels, it was not expected that participants would exceed intermediate-high level; they were thus expected to span six levels from novice-low to intermediate-high. To ensure the intelligibility of the holistic scale, its descriptions were translated into Japanese by the author, and paraphrased. The resulting criteria are shown in Appendix Q.

3.2.4 Experiment Procedure

First, the author gave participants directions about the experiment and a tutorial in the use of the recording equipment. Subsequently, to trial the recording and as a warm-up, all participants stated their name, the date of the experiment, and the weather. All participants then had two minutes to read material 1A-1 (Table 3-1). Half of the participants then responded to a retelling task (group 1A-A), while the other half responded to a summarizing task (group 1A-B). Both groups had two-and-a-half minutes to complete the task. This procedure follows that of the Story Retelling Speaking Test (SRST) developed by Hirai and Koizumi (2008, 2009 & 2013) and Koizumi and Hirai (2010 & 2012). In the second part of the experiment, material 1A-2 was used as a stimulus, and the groups were reversed; that is, group

1A-A responded to the summarizing task, while group 1A-B responded to the retelling task.

Due to the time restriction, only ten participants participated in the post-experiment interviews, which were conducted by the author in Japanese. Participants were asked for their impressions of this experiment in terms of the differences between the two tasks that they had completed. If the interviewee was able to answer this question, the interviewer posed further questions: “How do you discriminate between the two task types?” and “Which task was easy to answer, for you and all learners of English?” All the data were recorded and subsequently transcribed.

All the data were transcribed by the author and one research collaborator, and two raters then scored the participants’ performances (details of the transcription and rating process are provided in the following sections). The raters were then interviewed with respect to the following: “How do you distinguish between retelling and summarizing tasks?” and “What points is it important to distinguish between summary-like performances and retelling-like performances?”

3.2.5 Analysis Procedure

There were two phases to this analysis. In the first phase, the utterances were transcribed and the pauses and errors were coded. This was conducted individually by the author and a contributor, who was a freshman computer science major, with a high English proficiency level. In concrete terms, he had achieved a pre-first grade on the Eiken test; thus, his proficiency was considered sufficiently high to participate in this phase of the analysis.

During the rating phase, a graduate student and undergraduate student majoring in English education and with an interest in the field of language testing participated in rating participants’ performances. Both had experience of marking spoken performance in course work, but this was almost their first time evaluating spoken performances in an experiment setting; therefore, the author created a document of written instruction for rating and provided rater training.

3.2.6 Data analysis

Having transcribed the data individually, the author and research contributor compared their respective transcriptions. As some inconsistencies were found, both parties listened to the recorded utterances once more, and engaged in discussion before finalizing the transcription. Next, they tagged the errors and pauses, again discussing these between them. Subsequently, the two analysts detected the unit of analysis, in this case, the AS-unit, and this information was added to the transcription. Finally, the author detected and counted silent pauses of longer than one second using the *Silence Finder* add-on of the sound editing software *Audacity,* as well as calculating the total duration of silent pauses. Next, the author calculated the figure of indices of complexity, accuracy, and fluency of the spoken performances. The indices used in this study were: the number of word tokens, number of word types, number of syllables, number of one-second or longer pauses, words per minute (WPM), syllables per second (SPS), number of disfluency markers, errors per AS-unit, Guiraud index, type token ratio (TTR), and words per AS-unit.

In the rating phase, the author instructed the raters on how to rate, and they then evaluated all speech data individually. The inter-rater reliability of their evaluations was calculated at $\alpha = .81$; therefore, their original evaluations were deemed acceptable and treated as the final version. The raters then listened to the speech data for a second time, and were asked to judge whether each spoken performance comprised a retelling-like performance or summary-like performance.

The figures for the CAF indices and raters' scores were compared using paired t-tests, and the participants' and raters' interviews were analyzed.

3.2.7 Results and Discussion

The paired *t*-test showed no significant difference between the spoken performances in response to the retelling and summarizing tasks. Table 3.2 shows the means and standard deviations of the performances in the retelling and summarizing tasks, while Table 3.3 shows the results of the *t*-test.

Table 3.2

Means and Standard Deviations of Performance in the Retelling and Summarizing Tasks

	Retelling		Summarizing	
	M (SD)	95% CI	*M (SD)*	95% CI
To	59.10 (32.84)	[26.26, 91.93]	56.57 (32.51)	[24.06, 89.08]
Ty	31.86 (14.78)	[17.08, 46.64]	30.95 (13.84)	[17.11, 44.79]
Sy	87.29 (43.54)	[43.75, 130.82]	84.38 (44.15)	[40.23, 128.53]
NP	27.76 (8.04)	[19.73, 35.80]	22.95 (8.97)	[13.98, 31.92]
PL	127.47 (23.19)	[104.27, 150.66]	131.69 (17.18)	[114.52, 148.87]
WPM	23.64 (13.13)	[10.50, 36.77]	22.63 (13.00)	[9.62, 35.63]
SPS	0.58 (0.29)	[0.29, 0.87]	0.56 (0.29)	[0.27, 0.86]
DM	17.86 (12.30)	[5.56, 30.16]	15.05 (10.28)	[4.77, 25.33]
ErAS	0.44 (0.29)	[0.14, 0.73]	0.43 (0.25)	[0.18, 0.67]
GI	4.11 (0.84)	[3.28, 4.95]	4.09 (0.82)	[3.28, 4.91]
TTR	0.59 (0.14)	[0.45, 0.73]	0.60 (0.12)	[0.48, 0.71]
WdAS	6.51 (1.73)	[4.78, 8.24]	5.97 (1.09)	[4.88, 7.06]

Note. To = number of word tokens, Ty = number of word types, Sy = number of syllables, NP = number of 1 second or longer silent pauses, WPM = words per minute, SPS = syllables per second, DM = number of disfluency markers, ErAS = errors per AS-unit, GI = Guiraud index, TTR = type token ratio, WdAS = words/AS-unit, and *N* = 21.

Table 3.3

Results of Paired T Test Comparing the Two Task Conditions

	t	*p*	*d*	95% CI
To	0.44	.659	0.07	[-9.21, 14.26]
Ty	0.35	.728	0.06	[-4.44, 6.25]
Sy	0.39	.697	0.06	[-12.43,18.24]
NP	1.93	.067	0.56	[-0.37, 9.99]
PL	-1.14	.266	0.20	[-11.95, 3.49]
WPM	0.44	.659	0.07	[-3.68, 5.70]
SPS	0.39	.697	0.06	[-0.08, 0.12]
DM	1.85	.079	0.24	[-0.35, 5.97]
ErAS	0.19	.851	0.03	[-0.09, 0.11]
GI	0.11	.911	0.02	[-0.34, 0.38]
TTR	-0.15	.876	0.04	[-0.07, 0.06]
WdAS	1.28	.215	0.37	[-0.34, 1.42]

Note. To = number of word tokens, Ty = number of word types, Sy = number of syllables, NP = number of 1 second or longer silent pauses, WPM = words per minute, SPS = syllables per second, DM = number of disfluency markers, ErAS = errors per AS-unit, GI = Guiraud index, TTR = type token ratio, WdAS = words/AS-unit, and *N* = 21.

As can be seen, the results show that the performances obtained in response to the retelling and summarizing tasks were homogeneous (RQ1A-1). A marginal significance was detected in the indices of number of pauses and number of disfluency markers, both of which were large in the retelling task condition. While this may have been because participants took more time to recall the contents of the stimuli in the retelling task, no significant difference was found in the index of pause length.

Hence, the name of task did not influence the spoken performance. In fact, no significant differences were found, although the mean number of word tokens was lower in summarizing than retelling, while there were fewer silent pauses in the summarizing task. On the other hand, the pause length was found to be longer in the summarizing tasks. Therefore, it may be the case that some participants took longer to paraphrase or rephrase than to reproduce. In other words, a larger cognitive load was applied in the retelling task.

To answer research question 1A-2, the interview data were summarized and participants' reactions were consolidated into three answers. Six of the interview participants stated, "It is hard for me (us) to discriminate between retelling and summarizing tasks. Both tasks ask test takers to tell the story that they have read." Another three participants stated, "Summarizing is easier than retelling because I (we) can reconstruct the contents if I (we) can understand the contents of the stimulus." The final two participants stated, "Retelling is easier than the summarizing tasks, since I (we) can reproduce or recall the sentence if we can memorize it." To sum up their reflections, half of the participants were unable to discriminate between the two tasks, while a quarter took a position regarding which task type is easy to answer. Thus, some of the participants were able to find a strategy to answer retelling and summarizing tasks. This suggests that test takers may be able to find shortcuts to score or perform better in those tasks.

In the interview, rater A stated, "Chronologically the contents of the utterance is proceed, the performance was categorized as retelling; in addition, if the same expressions frequently appear in the utterance, the performance is categorized as retelling." Rater B stated, "When the expressions used in the utterances are similar to the original text, the performances were judged as retelling; in other words, if there are many instances of recall or reproduction in the utterance, the performance can be categorized as retelling." Thus, both raters thought that similarity of expression between the spoken performances and original text was the most important determinant of whether the performance was retelling-like or summarizing-like. Although Kissner (2006) and Reed and Vaughn (2012) focused on the reproduction of the main topic of the original text, this point was not mentioned by the raters of this study who focused on language rather than topic. This indicates that the frequency of rephrasing and paraphrasing, or the similarity between the utterance and original text provides a clue to the definitions of retelling and summarizing, respectively. However, in this series of studies, both tasks are categorized as the same task since the performances of retelling and summarizing tasks are homogeneous.

3.3 Study 1: Does Task Direction Influence Performances in Retelling and Summarizing Tasks?

3.3.1 Objectives, and Research Questions of Study 1

In the pilot study, the participants were unable to discriminate between retelling and summarizing, although some participants outlined the advantages of each task type. Furthermore, the raters in the pilot study stated that similarity to the original text is one of the characteristics of retelling. Thus, if test takers and even raters cannot discriminate between the two tasks, how do performances change when clear direction is given to test takers?

The purpose of this study was to ascertain the effects of task direction on spoken performance in retelling and summarizing tasks. The pilot study revealed no significant difference between the oral performances in retelling and summarizing tasks; in other words, test takers performed almost the same in both task conditions, and were unable to distinguish between the two tasks in terms of their wording. However, there is still room to investigate the effects of task direction on the respective performances in the two task types. Furthermore, the length of input text may influence the appearance of reproduction; therefore, variations in spoken performance in terms of CAF and raters' evaluation were compared using two-way repeated analyses of variance (ANOVA). Furthermore, the effects of task instruction on the frequency of reproduction were also observed. The research questions of study 1 are as follows:

RQ1B-1: Do the performances in summarizing tasks differ from the performances in retelling tasks in terms of CAF and raters' evaluations?

RQ1B-2: Does text length affect oral performance in retelling and summarizing tasks?

RQ1B-3: Do participants use as the original expressions in retelling tasks more frequently than in summarizing tasks?

3.3.2 Participants

A total of 46 students participated in this research. The participants comprised freshmen or sophomores of junior college or university. The data from nine participants were excluded from the analysis due to the quality of the recording, so 37 sets of data were used for the final analysis. All participants had taken general English courses during both junior college and university, and all were science or engineering majors. Although they did not disclose their detailed scores, all of them had taken intermediate level English courses and had achieved scores of at least 350 in the TOEIC Reading and Listening Test including TOEIC IP, and had taken practice tests in college.

3.3.3 Materials

The materials used in this experiment are shown in Table 3.4. Material 1B-1 and 1B-2 were categorized as *short* texts, while 1B-3 and 1B-4 were *long* texts. However, the text length differed between all materials. Materials 1B-1 and 1B-3 were used as the input materials for retelling, while Material 1B-2 and 1B-4 were used as the materials for summarizing. Thus, one short and one long text was used for each task type. Different reading durations were allocated in accordance with the length of the input material. Thus, 90 seconds were allotted to read material 1B-1 (98 words), while 30 seconds were added for each increase in text length of 50 words. For the rating scale, Hirai and Koizumi’s (2013) EBB scale2 was implemented.

Table 3.4

Materials Used in Experiment 1B

	Text Length	FKGL	FRE	JACET 8000 Level 1(%)	Reading Duration (Secs)
Material 1B-1	98	7.6	59.3	80.81	90
Material 1B-2	149	7.8	63.0	78.52	120
Material 1B-3	303	8.4	58.9	78.11	240
Material 1B-4	482	8.3	58.6	74.33	385

Note. JACET8000 Level 1 shows lexical diversity. FRE = Flesch Reading Ease. FKGLs and FREs were computed using Microsoft Word 2010. All materials are shown in the Appendices.

3.3.4 Procedure

As with the pilot study, all participants stated their names, date of data collection, and their physical condition to check the recording instruments and warm up. Next, they read the material within the designated reading time, and were then given two-and-a-half minutes to either summarize or retell the story they had just read. In the first session, half of participants were assigned the retelling task, while the remainder were assigned the summarizing task; their task conditions were then reversed. The first group read material 1B-1 for the short retelling, 1B-3 for the long retelling, 1B-2 for the short summarizing, and 1B-4 for the long summarizing. The second group read material 1B-1 for the short summarizing, 1B-3 for the long summarizing, 1B-2 for the short retelling, and 1B-4 for the long retelling. In order to analyze the effects of task direction, the following detailed instructions were given. For the retelling task, the author told participants: “You can use the words or expressions directly if you can recall them, but if you cannot remember the exact expressions, phrases, or sentences, just use your own expressions.” For the summarizing task, the following instruction was given: “You should paraphrase the sentence or expressions that you have read as much as possible, but if you cannot find an alternative expression, you can use the expressions provided in the material." Participants were not permitted to reread the material while responding to either of the tasks; thus, they had to depend on their memory.

3.3.5 Analysis and Rating

As in the pilot study, the author first transcribed the responses, while two experienced raters conducted rating simultaneously. These were the same raters as had participated in the pilot study and thus no rater training was provided. However, the raters were given a rating manual.

The indices used in this study are the same as those used in the pilot study, although an index of frequency of reproduction was added in order to answer research question 1B-3. In this study, reproduction is defined as phrases consisting of three or more words used in the original texts.

3.4 Results and Discussion

To answer research questions 1B-1 and 1B-2, repeated two-way ANOVA was conducted. There were 15 indices in this analysis; therefore, a multivariate analysis of variance (MANOVA) could have been more suitable than an ANOVA. However, it was expected that strong multicollinearity would be found among some indices (e.g., the number of token, types, and syllables); therefore, a repeated two-way ANOVA was used. Generally, when ANOVAs were repeated, the Bonferroni correction should be applied, but each index had different features; therefore, the Bonferroni correction was not applied. Table 3.5 shows the descriptive statistics of this analysis, while Tables 3.6 to 3.11 show the results of the ANOVA. There was no significant interaction in all indices. Figures 3.1 to 3.16 show the interaction plots of each index. As post-hoc analyses, analyses of main effects were also conducted, the result of which are shown in Tables 3.12 and 3.13.

Table 3.5

Mean, Standard Deviations, and 95% Confidence Intervals of Each Index (Experiment 1B)

	Retelling						Summarizing					
	100 words (Short)			300 words (Long)			150 words (Short)			500 words (Long)		
Indices	*M* (*SD*)		95% CI	*M* (*SD*)		95% CI	*M* (*SD*)		95% CI	*M* (*SD*)		95% CI
To	52.84	(27.19)	[25.65, 80.02]	43.89	(22.94)	[20.95, 66.83]	48.30	(48.30)	[26.20, 70.40]	40.32	(21.61)	[18.72, 61.93]
Ty	31.30	(12.00)	[19.29, 43.30]	26.30	(10.64)	[15.66, 36.93]	33.03	(11.78)	[21.24, 44.81]	27.41	(12.27)	[15.14, 39.67]
Sy	80.76	(39.18)	[41.58, 119.94]	67.73	(35.13)	[32.60, 102.86]	75.08	(32.48)	[42.60, 107.56]	60.00	(30.94)	[29.06,90.94]
NP	25.30	(6.49)	[18.81, 31.79]	24.11	(7.38)	[16.73, 31.49]	26.35	(7.06)	[19.30, 33.41]	25.49	(6.78)	[18.70, 32.27]
PL	125.62	(16.40)	[109.22,142.02]	126.52	(18.67)	[107.85, 145.19]	128.16	(14.73)	[113.43, 142.88]	133.14	(15.02)	[118.13, 148.16]
WPM	21.14	(10.87)	[10.26, 32.01]	17.56	(9.18)	[8.38, 26.73]	19.32	(8.84)	[10.48,28.16]	16.13	(8.64)	[7.49, 24.77]
SPS	0.54	(0.26)	[0.28, 0.80]	0.45	(0.23)	[0.22, 0.69]	0.50	(0.22)	[0.28, 0.72]	0.40	(0.21)	[0.19, 0.61]
DM	5.41	(5.08)	[0.33, 10.49]	4.65	(5.25)	[-0.60, 9.89]	4.97	(3.57)	[1.40, 8.54]	4.11	(4.10)	[0.01, 8.21]
ErAS	0.13	(0.13)	[0.00, 0.25]	0.19	(0.21)	[-0.02, 0.41]	0.17	(0.19)	[-0.02, 0.36]	0.18	(0.23)	[-0.05, 0.40]
GI	5.48	(1.13)	[4.35, 6.61]	5.01	(1.10)	[3.91, 6.11]	5.67	(0.97)	[4.69, 6.64]	5.11	(1.17)	[3.93, 6.28]
TTR	0.66	(0.20)	[0.46, 0.85]	0.66	(0.16)	[0.50, 0.82]	0.71	(0.09)	[0.62, 0.80]	0.72	(0.12)	[0.60, 0.84]
WdAS	6.59	(1.20)	[5.39, 7.79]	6.76	(1.23)	[5.53, 7.99]	6.31	(1.56)	[4.75, 7.87]	6.15	(1.38)	[4.77, 7.53]
CE	2.08	(0.92)	[1.16, 3.01]	1.97	(0.99)	[0.99, 2.96]	1.95	(0.88)	[1.07, 2.83]	1.84	(0.83)	[1.00, 2.67]
GV	2.08	(1.09)	[0.99, 3.17]	2.11	(1.10)	[1.01, 3.21]	2.03	(0.96)	[1.07, 2.98]	2.00	(0.94)	[1.06, 2.94]
Pro	2.43	(0.90)	[1.53, 3.33]	2.62	(0.86)	[1.76, 3.48]	2.41	(0.86)	[1.54, 3.27]	2.54	(0.86)	[1.55, 3.53]
FR	1.38	(1.28)	[0.10, 2.66]	1.41	(1.52)	[-0.11, 2.92]	0.86	(0.95)	[-0.08, 1.81]	0.76	(0.89)	[-0.14, 1.65]

Note. To = number of word tokens, Ty = number of word types, Sy = number of syllables, NP = number of 1 second or longer silent pauses, PL = total length of silent pauses, WPM = words per minute, SPS = syllables per second, DM = number of disfluency markers, ErAS = errors per AS-unit, GI = Guiraud index, TTR = type token ratio, WdAS = words/AS-unit, CE = communicative efficiency, GV = grammar and vocabulary, Pro = pronunciation, FR = frequency of reproduction, and *N* =37.

The results revealed that there is no significant interaction in all indices; however, some significant main effects were detected in the factor of length in number of tokens, types, and syllables, pause length, WPM, SPS, number of disfluency markers, and Guiraud index. On the other hand, the significant main effects in the factor of task were found in pause length, TTR, words per AS-unit, and frequency of reproduction. There is no significant main effect in errors per AS-unit.

Table 3.6

Two-Way Repeated ANOVA for Each Index of the Number of Utterances

Sources	*SS*	*MS*	*F*	*p*	η^2
			Tokens		
Task	608.11	608.11	2.55	.119	0.01
Length	2647.81	2647.81	13.47	.001*	0.03
Task x Length	8.76	8.70	0.13	.722	0.00
Error	61831.61	1717.55			
Total	83213.11				
			Types		
Task	74.49	74.49	0.99	.327	0.00
Length	1043.57	1043.57	15.65	.000*	0.05
Task x Length	3.57	3.57	0.11	.747	0.00
Error	13343.74	370.66			
Total	20796.99				

(Table continues)

Table 3.6—Continued

		Syllables			
Task	1662.27	1662.27	2.90	.100	0.01
Length	7308.11	7308.11	17.37	.000*	0.04
Task x Length	39.03	39.03	0.27	.610	0.00
Error	131151.77	3643.11			
Total	181134.27				

Note. N =37, df = 1, 52, * p < .05

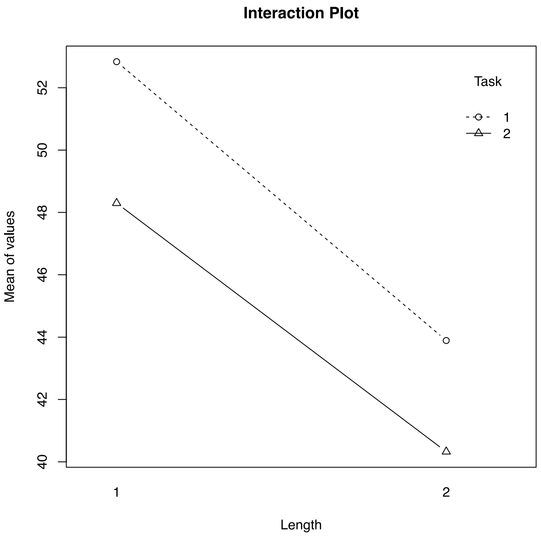


Figure 3.1. Interaction plot of number of word tokens.

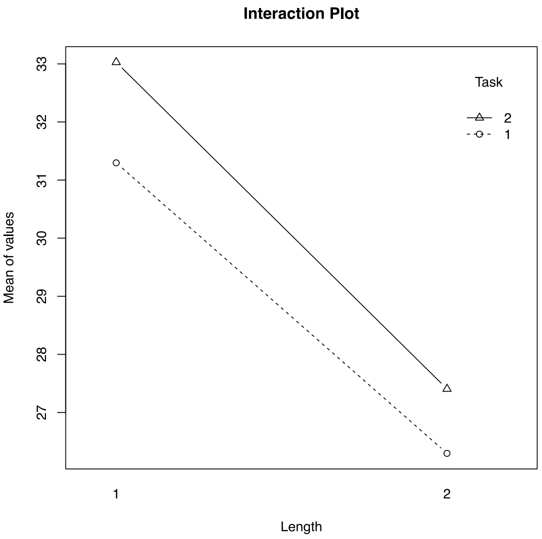


Figure 3.2. Interaction plot of number of word types.

Significant main effects were found in the indices for the tokens, types, and syllables of length factors, and the effect sizes were small for all the indices. Figures 3.1 to 3.3 show the interaction plots for tokens, types, and syllables, respectively. The results revealed that the participants performed shorter utterances when given a longer text.

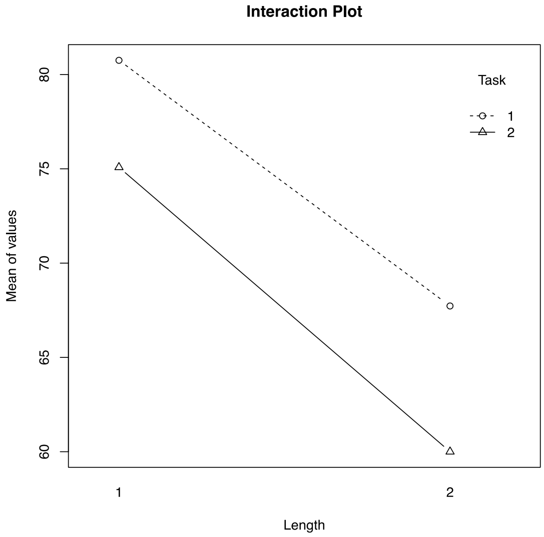


Figure 3.3. Interaction plot of number of syllables.

Table 3.7

Two-Way Repeated ANOVA for Each Index of Fluency

Sources	*SS*	*MS*	*F*	*p*	η^2
			No. of silent pauses		
Task	54.73	54.73	1.92	.175	0.01
Length	39.03	39.03	1.04	.314	0.01
Task x Length	0.97	0.97	0.05	.829	0.00
Error	3812.20	105.90			
Total	7021.70				

(Table continues)

Table3.7—Continued

		Pause length			
Task	775.81	775.81	9.25	.004*	0.02
Length	319.84	319.84	4.35	.044*	0.01
Task x Length	154.74	154.74	2.56	.117	0.01
Error	30329.66	842.49			
Total	39408.72				
		WPM			
Task	97.30	97.30	2.55	.120	0.01
Length	423.65	423.65	13.47	.001*	0.03
Task x Length	1.40	1.40	0.13	.722	0.00
Error	9893.06	274.81			
Total	13314.10				
		SPS			
Task	0.08	0.08	2.93	.096	0.01
Length	0.33	0.33	17.88	.000*	0.04
Task x Length	0.00	0.00	0.24	.627	0.00
Error	5.83	0.16			
Total	8.04				
		No. of disfluency markers			
Task	8.76	8.76	0.74	.400	0.00
Length	24.32	24.32	4.30	.050*	0.01
Task x Length	0.11	0.11	0.02	.890	0.00
Error	2170.08	2170.08			
Total	3017.0811				

Note. N =37, *df* = 1, 52, * *p* < .05.

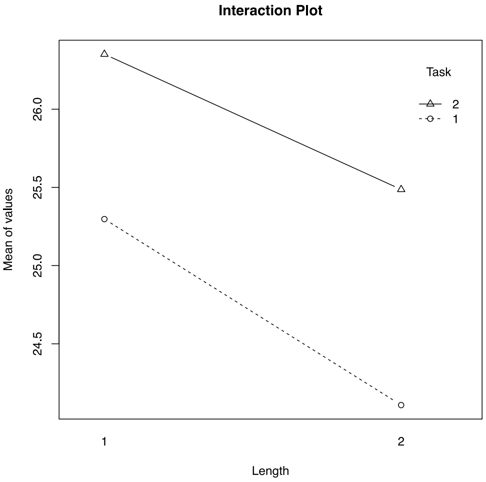


Figure 3.4. Interaction plot of number of 1 second or longer pauses.

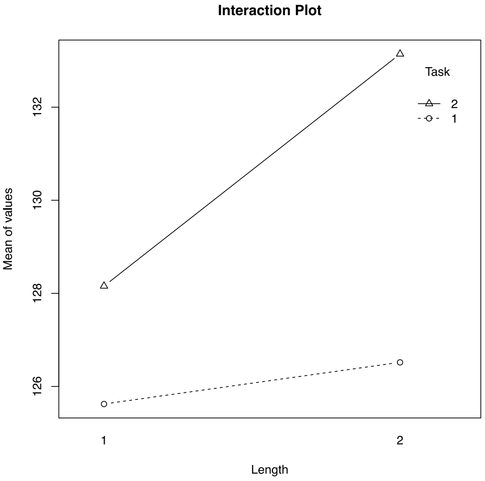


Figure 3.5. Interaction plot of pause length.

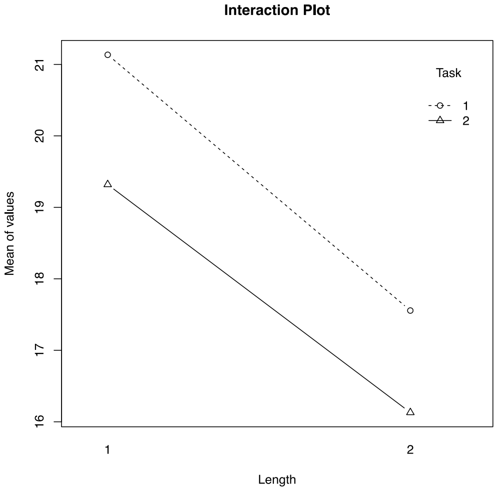


Figure 3.6. Interaction plot of WPM.

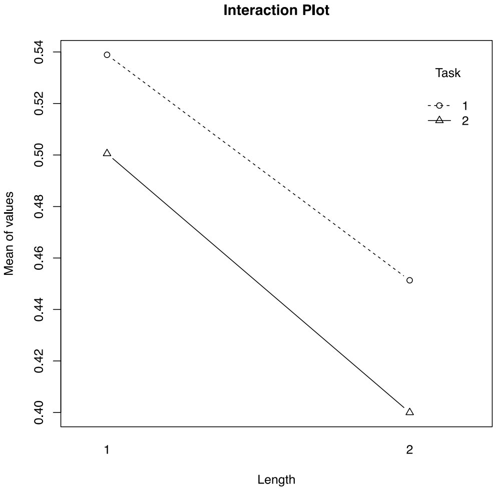


Figure 3.7. Interaction plot of SPS.

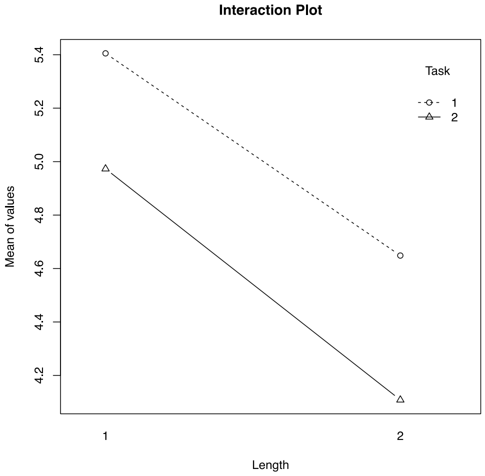


Figure 3.8. Interaction plot of number of disfluency markers.

Table 3.7 is the ANOVA table of each index of fluency, and Figures 3.4 to 3.8 show the interaction plots of each index of fluency. Significant main effects of length were found in the indices of pause length, WPM, SPS, and disfluency markers; in addition, a significant main effect of task in the index of pause length was found. These results indicate that the participants' fluency declined when given a longer text.

Table 3.8

Two-Way Repeated ANOVA for Index of Accuracy

Sources	*SS*	*MS*	*F*	*p*	η^2
			Errors per AS-unit		
Task	0.01	0.01	0.17	.683	0.00
Length	0.05	0.05	2.89	.098	0.01
Task x Length	0.03	0.03	1.05	.313	0.01
Error	2.03	0.06			
Total	5.41				

Note. N =37, *df* = 1, 37, * *p* < .05.

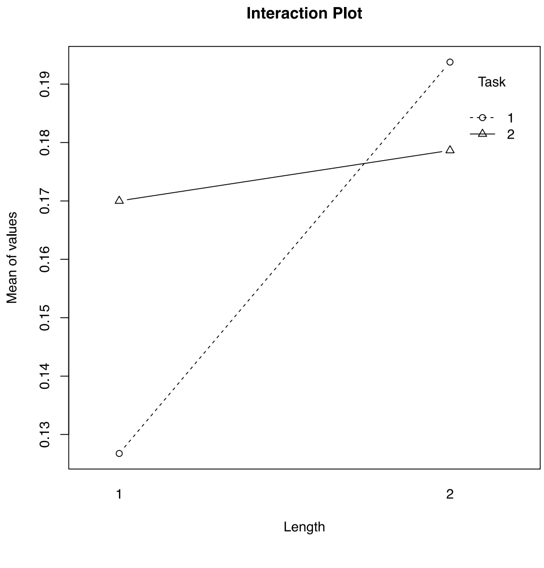


Figure 3.9. Interaction plot of errors per AS-unit.

Table 3.8 is the ANOVA table of accuracy, and Figure 3.9 is the interaction plot of errors per AS unit. There was no significant interaction or main effect for this index, meaning that the accuracy of utterance was not influenced by the length of the text or the task directions.

Table 3.9

Two-Way Repeated ANOVA for Each Index of Complexity

Sources	*SS*	*MS*	*F*	*p*	η^2
			Guiraud index		
Task	0.70	0.70	0.97	.333	0.00
Length	9.80	9.80	15.22	.000*	0.05
Task x Length	0.08	0.08	0.25	.620	0.00
Error	112.78	3.13			
Total	184.00				
			Type token ratio		
Task	0.12	0.12	6.34	.016*	0.04
Length	0.00	0.00	0.21	.649	0.00
Task x Length	0.00	0.00	0.00	.963	0.00
Error	1.79	0.05			
Total	3.25				
			Words per AS-unit		
Task	7.36	7.36	4.29	.046*	0.03
Length	0.00	0.00	0.00	.988	0.00
Task x Length	1.02	1.02	0.60	.445	0.00
Error	66.42	1.85			
Total	270.42				

Note. N =37, df = 1, 52, * p < .05.

Table 3.9 and Figures 3.10, 3.11, and 3.12 show the results of a two-way ANOVA of complexity. There were no significant interactions, but a significant main effect of length was found in the Guiraud index. In contrast, significant main effects of the task directions were found in the indices of type-to-token ratio and words per AS unit.

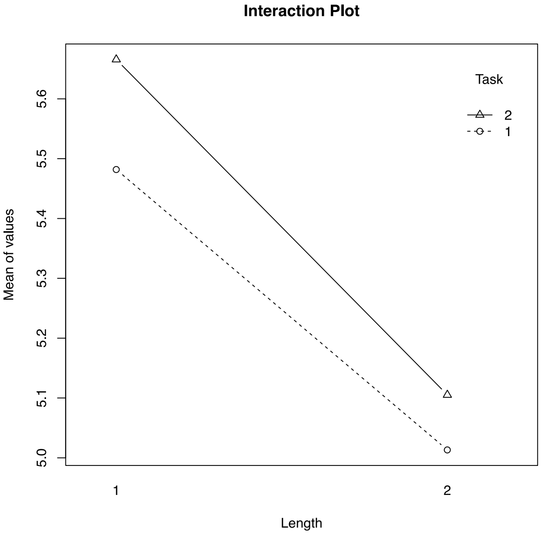


Figure 3.10. Interaction plot of Guiraud index.

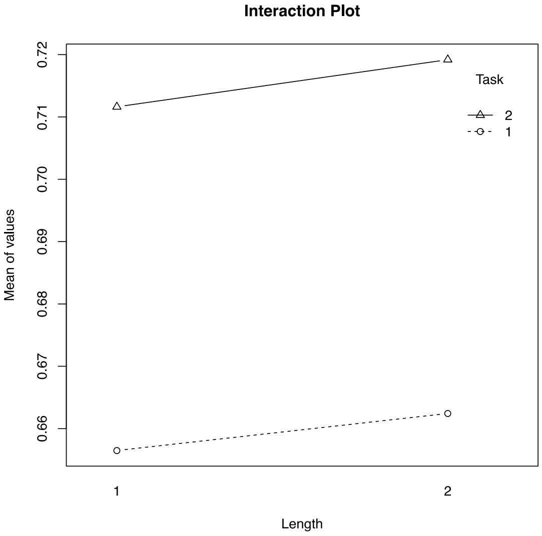


Figure 3.11. Interaction plot of type token ratio.

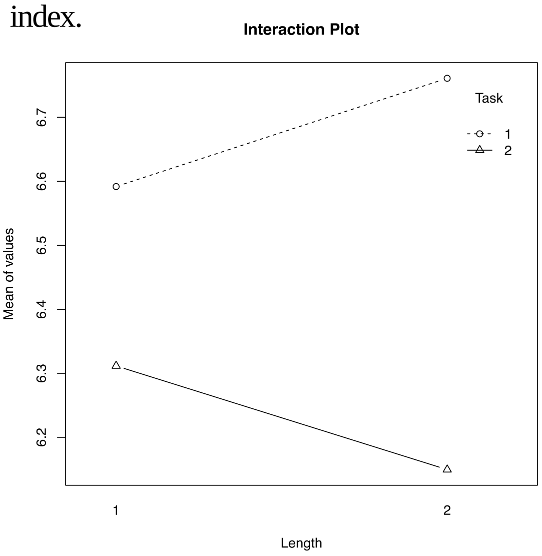


Figure 3.12. Interaction plot of words per AS-unit.

Table 3.10

Two-Way Repeated ANOVA for Each Result of Rater Evaluation

Sources	*SS*	*MS*	*F*	*p*	η^2
			Communicative efficiency		
Task	0.68	0.68	1.59	.216	.01
Length	0.43	0.43	1.63	.210	.00
Task x Length	0.00	0.00	0.00	1.00	.00
Error	82.76	2.30			
Total	110.76				
			Grammar & vocabulary		
Task	0.24	0.24	0.34	.564	.00
Length	0.00	0.00	0.00	1.00	.00
Task x Length	0.02	0.02	0.05	.820	.00
Error	89.57	2.48			
Total	151.57				
			Pronunciation		
Task	0.11	0.11	0.36	.554	.00
Length	0.97	0.97	2.69	.110	.01
Task x Length	0.02	0.02	0.10	.757	.00
Error	84.00	2.33			
Total	119.00				

Note. *N* =37, *df* = 1, 37, * *p* < .05.

Table 3.9 and Figures 3.10, 3.11, and 3.12 show the results of a two-way ANOVA of complexity. There were no significant interactions, but a significant main effect of length was found in the Guiraud index. In contrast, significant main effects of the task directions were found in the indices of type-to-token ratio and words per AS unit.

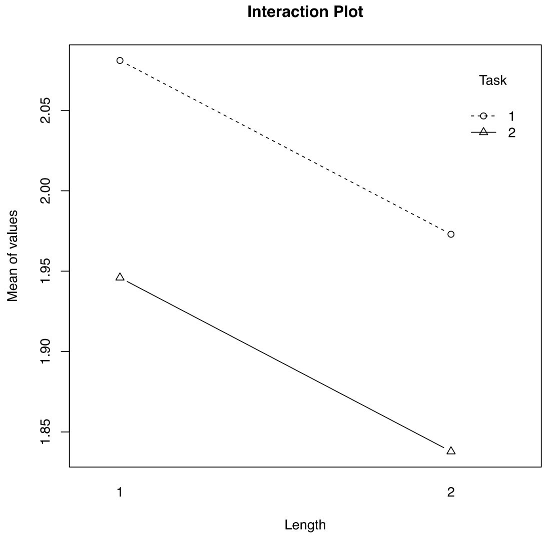


Figure 3.13. Interaction plot of the grade of communicative efficiency.

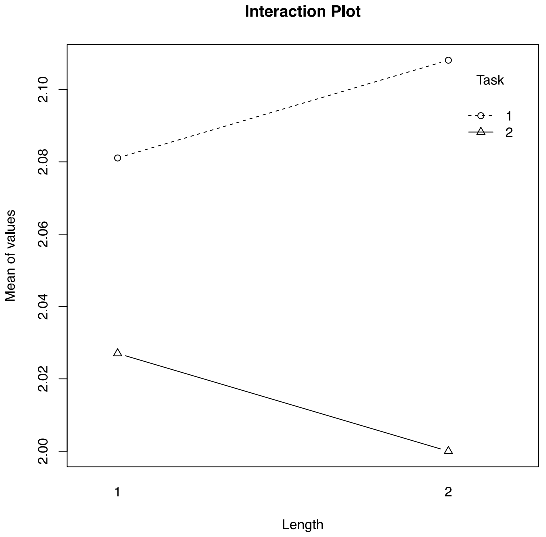


Figure 3.14. Interaction plot of the grade of grammar & vocabulary.

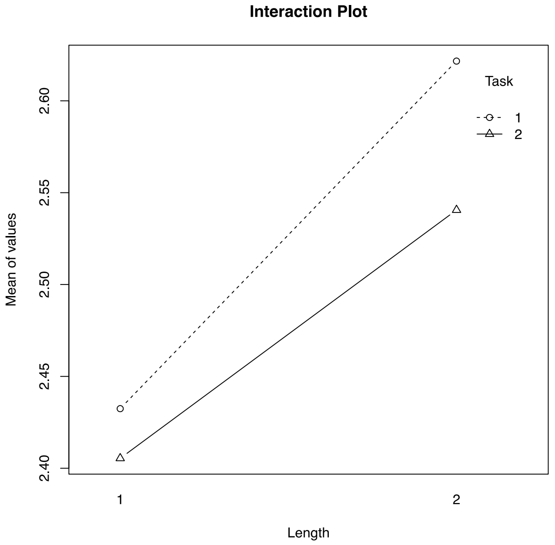


Figure 3.15. Interaction plot of the grade of pronunciation.

A significant main effect of task directions was found when the frequency of reproduction was compared, and Figure 3.16 indicates that participants tended to use the same expressions when given retelling tasks.

Table 3.11

Two-Way Repeated ANOVA for Frequency of Reproduction

Sources	*SS*	*MS*	*F*	*p*	η^2
		Frequency of Reproduction			
Task	12.49	12.49	8.95	.005*	0.58
Length	0.06	0.06	0.05	.820	0.00
Task x Length	0.17	0.17	0.16	.890	0.00
Error	37.58	1.04			
Total	215.48				

Note. N =37, *df* = 1, 36, * *p* < .05.

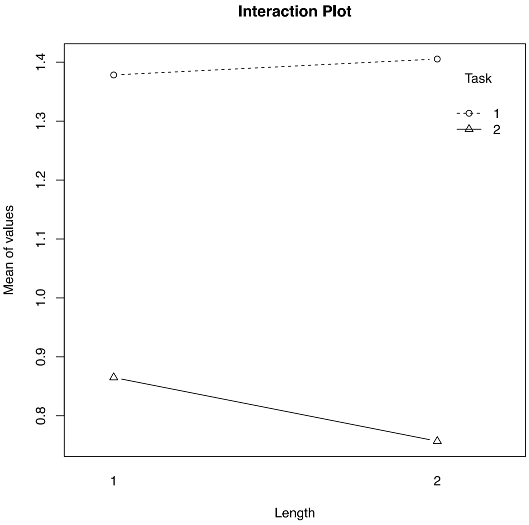


Figure 3.16. Interaction plot of frequency of reproduction.

To analyze the main effects, the significant main effects of text length on task directions were detected for the number of word tokens, pause length, TTR, and words per AS unit. Conversely, the significant main effects of task directions on text length were the number of word types, number of syllables, pause length, WPM, SPS, number of disfluency markers, Guiraud index, and frequency of reproduction.

Table 3.12

Main Effects for Each Index of Text Length on Task Direction

	Retelling		Summarizing					
	M	*SE*	*M*	*SE*	95%CI	*F*	*p*	η^2
To	50.57	3.76	42.11	3.43	[2.79, 13.13]	13.7	.001*	.01
Ty	28.78	1.68	30.22	1.79	[-4.31, 1.48]	0.99	.327	.00
Sy	74.24	5.74	67.54	4.90	[-1.28, 14.69]	2.90	.097	.01
NP	24.70	0.98	25.92	0.93	[-3.00, 0.57]	1.82	.175	.01
PL	127.07	2.67	139.65	2.33	[-7.63, -1.53]	9.25	.004*	.02
WPM	19.34	1.56	17.72	1.34	[-0.44, 3.68]	2.55	.119	.01
SPS	0.46	0.04	0.45	0.03	[-0.01, 0.10]	2.90	.097	.01
DM	5.03	0.80	4.54	0.58	[-0.66, 1.64]	0.736	.400	.00
ErAS	0.16	0.02	0.17	0.03	[-0.08, 0.06]	0.158	.693	.00
GI	5.25	0.16	5.39	0.16	[-0.42, 0.15]	0.97	.330	.00
TTR	0.66	0.03	0.72	0.01	[-0.10, -0.01]	6.26	.017*	.00
WdAS	6.68	0.16	6.23	0/16	[0.01, 0.88]	4.29	.046*	.03
CE	2.01	0.14	1.91	0.13	[-0.06, 0.23]	1.63	.210	.01
GV	2.05	0.15	2.05	0.13	[-0.24, 0.24]	0.00	1.00	.00
Pro	2.42	0.13	2.58	0.14	[-0.36, 0.04]	2.70	.110	.00
FR	1.39	0.18	0.81	0.12	[0.19, 0.98]	8.95	.005*	.06

Note. To = number of word tokens, Ty = number of word types, Sy = number of syllables, NP = number of 1 second or longer silent pauses, PL = total length of silent pauses, WPM = words per minute, SPS = syllables per second, DM = number of disfluency markers, ErAS = errors per AS-unit, GI = Guiraud index, TTR = type token ratio, WdAS = words/AS-unit, CE = communicative efficiency, GV = grammar and vocabulary, Pro = pronunciation, FR= frequency of reproduction, and *N* =37.

Table 3.13

Main Effects for Each Index of Task Direction to Text Length

	Short		Long					
	M	*SE*	*M*	*SE*	95%CI	*F*	*p*	η^2
To	48.37	3.91	44.31	3.34	[-1.01, 9.20]	2.55	.119	.03
Ty	32.16	1.73	26.85	1.71	[2.59, 8.03]	15.65	.000*	.05
Sy	77.92	5.47	63.87	5.00	[7.22, 20.89]	17.37	.000*	.04
NP	25.82	0.95	24.80	1.02	[-1.01, 3.07]	1.04	.314	.01
PL	126.89	2.39	129.83	2.59	[-0.08, -5.80]	4.35	.044	.01
WPM	20.23	1.50	16.84	1.37	[1.51, 5.24]	13.47	.001*	.03
SPS	0.52	0.04	0.43	0.03	[0.05, 0.14]	17.37	.000*	.04
DM	5.19	0.63	4.38	0.71	[0.02, 1.60]	4.30	.045*	.01
ErAS	0.15	0.02	0.18	0.02	[-0.08, 0.01]	2.913	.096	.01
GI	5.57	0.15	5.06	0.17	[0.15, 0.78]	15.23	.000*	.05
TTR	0.68	0.02	0.69	0.02	[-0.04, 0.02]	0.20	.658	.00
WdAS	6.45	0.17	6.46	0.15	[-0.48, 0.47]	0.00	.989	.00
CE	2.03	0.14	1.89	0.13	[-0.08, 0.35]	1.59	.216	.00
GV	2.10	0.16	2.01	0.13	[-0.20, 0.36]	0.34	.563	.00
Pro	2.53	0.13	2.47	0.13	[-0.13, 0.24]	0.36	.554	.01
FR	1.12	0.14	1.08	0.15	[-0.32, .040]	0.05	.820	.00

Note. To = number of word tokens, Ty = number of word types, Sy = number of syllables, NP = number of 1 second or longer silent pauses, PL = total length of silent pauses, WPM = words per minute, SPS = syllables per second, DM = number of disfluency markers, ErAS = errors per AS-unit, GI = Guiraud index, TTR = type token ratio, WdAS = words/AS-unit, CE = communicative efficiency, GV = grammar and vocabulary, Pro = pronunciation, FR = frequency of reproduction, and *N* =37.

Hence, if no directions are given, such as “Use the expressions written in the stimulus” or “Relate the main topic of the passage as similarly to the original as possible,” the performances in the retelling and summarizing tasks are basically the same, while the effects of text length and task direction partially affect the performances.

One interesting result was that a significant main effect of frequency of reproduction was found in the factor of task direction. The study participants could reproduce the same expressions as were used in the original text in the retelling task condition. Thus, it can be said that task direction partially affects performance in retelling and summarizing, but the effects of text length are greater than the effects of task direction.

3.5 Conclusion of Study 1

In both retelling and summarizing, the naming of the task does not affect the spoken performance, but if clear direction is given to the participants, they are likely to recycle the expressions used in the original text. However, the results of study 1 show that the effects of text length are greater than the effects of task direction. This point will be observed in study 2.

If teachers or test developers want to elicit specific expressions shown in input material from test takers and give directions such as “Use words or phrases written in the material as much as possible,” then retelling tasks can be practiced not only with speaking skills but also the learning of productive vocabulary. In any case, directing tasks is important to elicit expected performances.

Chapter 4

Study 2: The Effects of Input Materials in Retelling Tasks on Spoken Performances in Terms of Text Length, Difficulty, and Input Mode

4.1 Introduction to Study 2

Study 2 is composed of two experiments, the purpose of which were to observe the effects of input materials, in other words, stimuli, on spoken performances in retelling tasks. As mentioned in chapter 2, task condition and quality are amongst the most important aspects of creating speaking tests and conducting speaking assessments. This study focused on the type of task condition that can appropriately elicit a long spoken performance.

4.2 Experiment 2A: The Effect of Text Length and Difficulty on Spoken Performances in Retelling Tasks

4.2.1 Objective, Design, and Research Questions

The task conditions in retelling tasks are easily changed to accommodate the ability of the test takers; hence, by considering the effects of the task condition, controlling the task condition is easy to convey to all users of retelling tasks as a test task. However, some questions have arisen regarding controlling the input materials of retelling tasks. Koizumi and Hirai (2010) revealed that the number of utterances is not influenced by the text length of the input, a result that they did not anticipate. Therefore, there is room to investigate the effect of input text length. Additionally, no study has dealt with the effects of difficulty and input mode. Thus, the following three research questions were posed to observe the effects of input on retelling utterances:

RQ2A-1: How does the length of the input material affect the utterances in retelling tasks?

RQ2A-2: How does the difficulty of the input material affect the utterance in retelling tasks?

RQ2A-3: Does the input mode affect the utterance in retelling tasks?

4.2.2 Participants

Fifty-nine university students participated in experiment 2A, and the data for 56 were examined. Data for three students were excluded due to difficulties with the audio-recording equipment. All participants were university freshmen whose majors ranged widely. The participants' proficiency was intermediate level, according to the placement.

4.2.3 Materials

The input material information is shown in Table 4.1. Material 2A-1and 2A-2were used to observe the influence of text length to spoken performance, and Material 2A-3 and 2A-4 were used to answer to RQ2A-2. Material 2A-3 was treated as difficult material, and Material 2A-4 was treated as easy material. The length of the materials in the reading sentences differed, but their Flesch-Kincaid Grade Levels (FKGL) were almost identical. Contrariwise, the lengths of the listening materials were controlled, or the number of words in the materials was almost identical, but the text difficulty differed. To ascertain the effect of text difficulty, the listening speed (words per minute; WPM) was also controlled. To do so, a computer software, *Globalvoice English*, was used in this experiment. The sound quality was relatively good, and the study participants reported no drawback of the text-to-speech sounds. In fact, this result matched the study conducted by Hirai and O'ki (2011), and this software was profitable for controlling WPM.

Two materials in the reading input comprised quotations from *AFP World News Report: Looking at the World through AFP News* (Shishido, Allen, & Takahashi, 2012) but the author changed the expressions to control for text readability and length. Two other listening materials were written by a graduate student majoring in English education. The passages were checked for English in use by a native English speaker.

Table 4.1

Input Materials Information (Experiment 2A)

	Text length	FKGL	FRE	JACET8000 Level 1(%)	WPM
Material 2A-1	100	13.5	41.2	72.0	–
Material 2A-2	144	13.1	35.9	58.7	–
Material 2A-3	124	8.2	60.4	82.7	155
Material 2A-4	125	6.4	73.7	87.1	159

Note. JACET8000 Level 1 shows lexical diversity. FRE is the abbreviation for Flesch Reading Ease. FKGLs and FREs were computed using Microsoft Word 2010. All materials are shown in the Appendix A, B, C, D, E, and F.

4.2.4 Experiment Procedure

The experiment to observe the effects of text length, and the experiment using reading materials, took place in December 2012, and another experiment, to ascertain the effects of text difficulty, in other words using listening materials, was conducted in March 2013 with the same participants. The utterances were collected simultaneously in each experiment. All data were collected via the CALL system; *movie Teleco*. All participants had experience of answering this type of task, and had read the input materials used in the reading group as material in a lesson. However, none had listened to the listening material; thus, the input condition was not completely controlled.

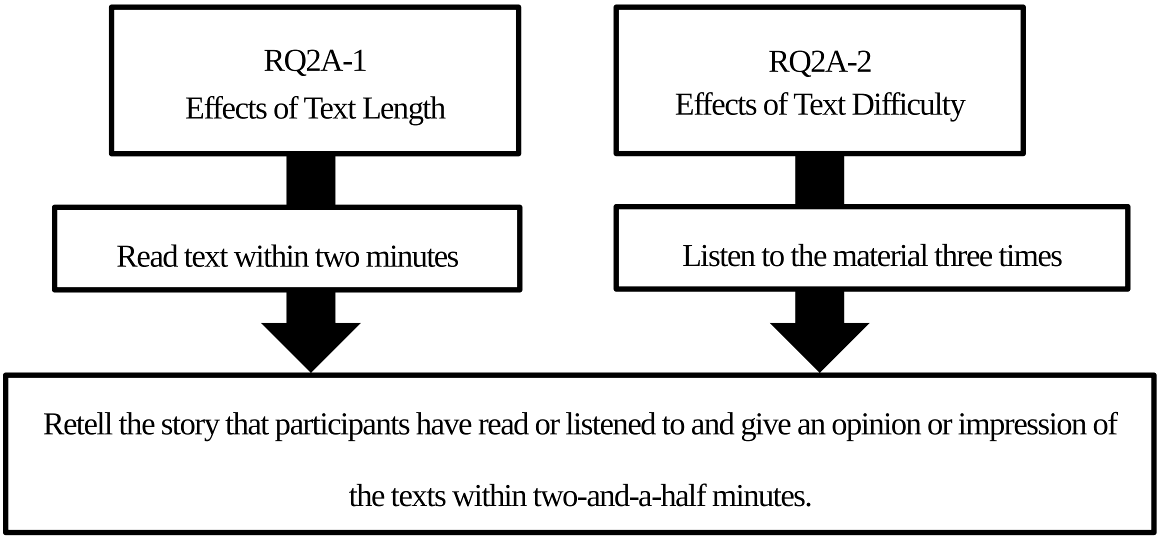

RQ2A-1
Effects of Text Length
RQ2A-2
Effects of Text Difficulty
Read text within two minutes
Listen to the material three times
Retell the story that participants have read or listened to and give an opinion or impression of the texts within two-and-a-half minutes.

Figure 4.1. Procedure of data collection.

Figure 4.1 outlines the procedure of data collection. In the experiment of text length, the text materials were presented on paper, and participants read the passage within two minutes. In the experiment of text difficulty, the listening information was given three times. If the participants read the story at the same pace as the listening material, or material 1A-3 or 1A-4, they would be able to read material 1A-1 three times within two minutes; therefore, the author decided to play the listening material three times in the experiments of text difficulty. After reading or listening to the material, the participants addressed to retelling the story and gave their opinion or impressions of the story they had read or listened to for two-and-a-half minutes. They were not permitted to use a dictionary, talk with others, or take notes while reading or listening to the materials. In addition, four keywords were shown on the test handout, which they could check as a clue to recall the contents.

4.2.5 Analysis Procedure

All utterances were transcribed and double-checked by graduate students and undergraduate students majoring in English education. Where disagreement occurred, the transcribers resolved it through discussion. The unpruned transcriptions included fillers, repetitions, and Japanese utterances. This unnecessary information was deleted, and pruned transcriptions were produced and then analyzed using the JACET8000 analysis program *v8an* (Shimizu, 2004). The number of words (tokens) and the variety of the vocabulary at the lemma level (indexes: these terms are often called *types*) were counted. The Guiraud index (= number of different word types / √Number of pruned tokens) was computed to find the productive vocabulary size. This indicator is widely used to observe learners' lexical richness, and Vermeer (2000) advocated the Guiraud index as an adequate measure for learners with a vocabulary size of fewer than 3,000 words. The WPM and the number of AS units (Foster, Tonkyn, & Wigglesworth, 2000) were also calculated to observe participants' fluency. Additionally, the test takers' performances were scored using the EBB (Empirically derived Binary-choice Boundary definition) scale, which was developed and used by Hirai and Koizumi (2008) to assess communicative efficiency, grammar and vocabulary, contents of utterance, and pronunciation. Performances were scored in terms of communicative efficiency, grammar and vocabulary, content, and pronunciation. The topics were scored from 1 (worst) to 5 (best), and the total scores ranged from 4 to 20.

4.2.6 Results and Discussion

The inter-rater reliability of the performance scoring was checked. Thirty-one performances were scored by a trained rater. The scores, which were rated by the author, were then used to calculate Cronbach's alpha. The inter-rater reliability of the two raters was high (see Table 4.2).

Table 4.2

Inter-rater Reliability of the Two Raters (N = 31)

	CE	GV	Contents	Pronunciation	Total score
α	.935	.955	.950	.841	.958

Note. N = 31. CE means "communicative efficiency" and GV means "grammar and vocabulary."

4.2.6.1 Effects of Text Length

Table 4.3 shows the standard deviation and means for each scale (number of word tokens, number of types words, Guiraud index, WPM, AS unit, performances). The paired *t*-test (Table 4.4) revealed significant differences between the indices for Reading 1 and Reading 2. Therefore, the effects of reading input length appear to have affected test takers' utterances and thus the test quality (RQ2A-1).

Table 4.3

Means and Standard Deviations for Each Scale: Text length

	Reading 1 (100 words)		Reading 2 (144 words)	
	M	*SD*	*M*	*SD*
number of tokens words	94.90	25.62	108.95	33.45
number of types words	53.66	11.58	60.76	12.84
Guiraud index	12.85	2.30	13.80	3.01
WPM	37.96	10.25	43.58	13.38
number of AS units	11.02	3.48	11.71	3.49
total EBB scale scores	13.71	2.15	12.91	2.05

Note. N = 56.

Table 4.4

Differences Between the Number and Quality of the Utterances: Case of Text Length

		t	*p*	*d*	
number of utterances	tokens	–5.70	.000	0.55	1 < 2
	indexes	–6.36	.000	0.71	1 < 2
lexical diversity	Guiraud index	–2.42	.019	0.42	1 < 2
fluency	WPM	–5.70	.000	0.55	1 < 2
	number of AS units	–1.71	.093	0.20	1 = 2
evaluation	total EBB scale scores	0.61	.544	0.08	1 = 2

Note. *N* = 56. Cohen’s d was calculated with the *Effect size calculation sheet* (Mizumoto, n.d.).

This result reveals that test takers speak more words more fluently when they input relatively long texts, which is contrary to the result of Koizumi and Hirai (2010). However, the larger number of utterances produced in longer materials seems natural because the test takers use a larger vocabulary, as shown in the stimuli. Although the result was contrary to the earlier study, Koizumi and Hirai (2010) expected that longer sentences would imply more production than shorter sentences. Therefore, the result for RQ2A-1 in this study met their expectations.

4.2.6.2 Effects of Text Difficulty

Table 4.5 shows the means and standard deviations for listening input, while Table 4.6 shows the result of the paired *t*-test of different input difficulty. Due to time constraints, participants could only answer one question; therefore, the sample size of this experiment was half that of the others. Except for the Guiraud index, the other indices were statistically nonsignificant; in fact, the difficulty of the input might not have affected the utterance in retelling tasks in the listening condition (RQ2A-2). Compared to the means of the indices in the reading case and listening case, only the means of the Guiraud index differed largely from each other. This point should be examined in future research.

Table 4.5

Means and Standard Deviations for Each Scale: Case of Length of Text Difficulty

	Listening 1 (FKGL =8.2)		Listening 2 (FKLG = 6.4)	
	M	*SD*	*M*	*SD*
tokens	92.79	49.91	90.79	33.00
indexes	52.25	13.80	47.49	14.75
Guiraud index	5.51	0.56	5.00	0.76
WPM	37.11	19.91	36.19	13.20
number of AS units	11.96	4.69	12.25	4.07
total EBB scale scores	10.92	2.64	10.46	2.86

Note. *N* = 28.

Table 4.6

Differences Between the Number and Quality of Utterances: Case of Different Text Difficulty

		t	*p*	*d*	
number of utterances	tokens	0.21	.838	0.06	1 = 2
	indexes	1.17	.247	0.31	1 = 2
lexical diversity	Guiraud index	2.82	.007	0.76	1 < 2
fluency	WPM	0.21	.838	0.06	1 = 2
	Number of AS units	–0.22	.810	0.07	1 = 2
evaluation	total EBB scale scores	0.56	.580	0.15	1 = 2

Note. *N* = 28.

4.2.6.3 Effects of Text Mode

Finally, the effect of the input mode was compared. The results showed that the reading input might draw upon more features of spoken English (Table 4.7). To see the effects of the input mode, the averages of the indices were compared with a *t*-test. In terms of fluency, the WPM differed significantly, but the number of AS units did not. In fact, the effect size was small in WPM (*d* = 0.26); therefore, the difference in input presentation mode might not have affected fluency. Although fluency was not affected by the input mode, other terms (number of utterances, variety of vocabulary, and analytic performances) differed significantly. The effect size of the analytic performances was small (*d* = 0.36), but the number of utterances and lexical diversity were large (*d* > .80). Thus, the input presentation mode affected test takers' number of utterances (RQ2A-3). A limitation of this experiment was the lack of consideration for other factors such as syntactic or morphological complexity. This point should be considered and controlled in future research.

Table 4.7

Differences Between the Number and Quality of Utterances: Case of Different Input Mode

		t	*p*	*d*	
number of utterances	tokens	10.98	.007	2.93	R > L
	indexes	18.64	.000	2.49	R > L
lexical diversity	Guiraud index	7.05	.000	3.49	R > L
fluency	WPM	2.78	.007	0.26	R > L
	number of AS units	-0.98	.334	0.12	R = L
analytic scales	total EBB scale score	2.70	.009	0.36	R > L

Note. *N* = 56.

The results reveal that the factor of text length in the reading condition and the input presentation mode might influence oral performance in retelling tasks; however, the text difficulty in the listening condition might not have affected the utterances. Text length and input mode should be considered when teachers create oral retelling tests, but difficulty might not be eligible for consideration in test development. Of course, a specific or difficult topic might affect the test taker's utterance. However, the effect of the input difficulty might be small.

RQ2A-1 reveals interesting results. The effect of text length is opposite the results of an earlier study by Koizumi and Hirai (2010). This disagreement with previous research shows room for follow-up research. This disagreement might have been caused by differences in text difficulty. The FKGL of the input text, which is shown in Koizumi and Hirai's (2010) work, was 2.8, while the FKGLs of this experiment were 13.5 and 13.1. To ascertain the effect of text length and difficulty, these points should be completely controlled. Thus, these points will be revised and retried in future research.

The effect of different input modes clearly affected the number of utterances (RQ2A-3). Only the fluency indices showed different aspects, but this tendency was evident in the different input lengths. In fact, the effect size of WPM was small (d = 0.26); therefore, the input mode might not have affected fluency in the retelling task. However, the number of uttered words and lexical diversity in the reading mode were clearly larger than in the listening input mode. Thus, when teachers or test developers want to evaluate learners' speaking fluency through oral retelling tasks, they need not consider the input mode. However, as the results show, the reading input implies that test takers' produce more words and richer vocabulary.

In relation to RQ2A-2, in contrast, no significant differences between the difficult input and easy input were observed for the scale and number of utterances, vocabulary diversity, fluency, and performance evaluation. Koizumi and Hirai (2009) argued that the difficulty of the input material does not affect the speaking task; therefore, this result supports their claim based on their questionnaire study and this factor need not be considered when teachers or test developers design a test. Furthermore, they might not need to give much thought to selecting input materials.

In conclusion, when teachers or test developers design a speaking test using oral retelling tasks, they do not need to consider small differences in text difficulty, but should respect the factors of text length and information presentation mode. This implication might affect test development in the classroom setting.

There is a relevant description of retelling tasks in the B1.2 level of the CEFR-J: “I can give an outline or list the main points of a short story or a short newspaper article with some fluency, adding my own feelings and ideas” (Tono, 2013). This descriptor is an intermediate-level description of spoken production; therefore, if test takers can perform sufficiently, they might demonstrate an intermediate level of spoken performance. This point is also noteworthy when test developers design a speaking test. In CEFR-J, B1.2 level is upper middle; therefore, retelling tasks might be assumed to be difficult for beginner or elementary-level learners. However, by setting the goal or objectives to appropriately challenge speaking presentation skills, the retelling tasks may have a positive impact on improving learners’ speaking skills. To observe the general difficulty of retelling tasks, their effects examined in this study should be analyzed in the following study using many-faceted Rasch measurement (MFRM).

This study has several limitations, the largest being the unregulated materials. The input materials for RQ2A-3 (Does the input mode affect the utterance in retelling tasks?) should be the same. This limitation may have affected the results. Thus, this factor will be revised, and deeper analysis undertaken. Another limitation is that the errors in utterances were not analyzed. To evaluate the actual oral production performance, this should be checked and considered. Therefore, this point will also be reformulated in the next experiment.

4.2.7 Conclusion of Experiment 2A

This study examined the utterances of Japanese learners of English to observe the effects of input length, input difficulty, and input mode. The effects of text length and input mode might influence the number of utterances and the subsequent evaluation. These results might have educational implications, especially for test design. Test developers should consider the text length and input mode; they can then draw test takers' utterances more effectively. Using retelling tasks as a testing task or classroom activity might help integrate speaking skills with reading or listening skills. Then, using retelling tasks of an appropriate text length and input information might benefit language teaching and testing.

4.3 Experiment 2B: The Effect of Text Length and Input Mode on Spoken Performances in Retelling Tasks

4.3.1 Objective and Research Questions

In experiment 2A, I investigated the effects of input mode, text length, and text difficulty on speaking performance. Experiment 2A revealed that test takers can produce more words under the reading input condition or when longer input material is provided. However, experiment 2A was limited, amongst other factors, by its small sample size and uncontrolled experimental design. Retelling tasks have certain benefits such as the ease of material development and ability to control difficulty levels or target items (e.g., expressions, grammar, or content), although the effects of input mode or text length have not been studied in detail. Therefore, the current study was conducted to find an effective combination to elicit a speaking performance from learners.

To overcome the limitations, especially the uncontrolled experimental design, experiment 2B was designed. Figure 4.2 shows the flow of data collection of experiment 2B. Thus, it was planned to analyze the effects of text length and input mode on the spoken performances in retelling tasks. In the current study, two research questions were set to observe these effects.

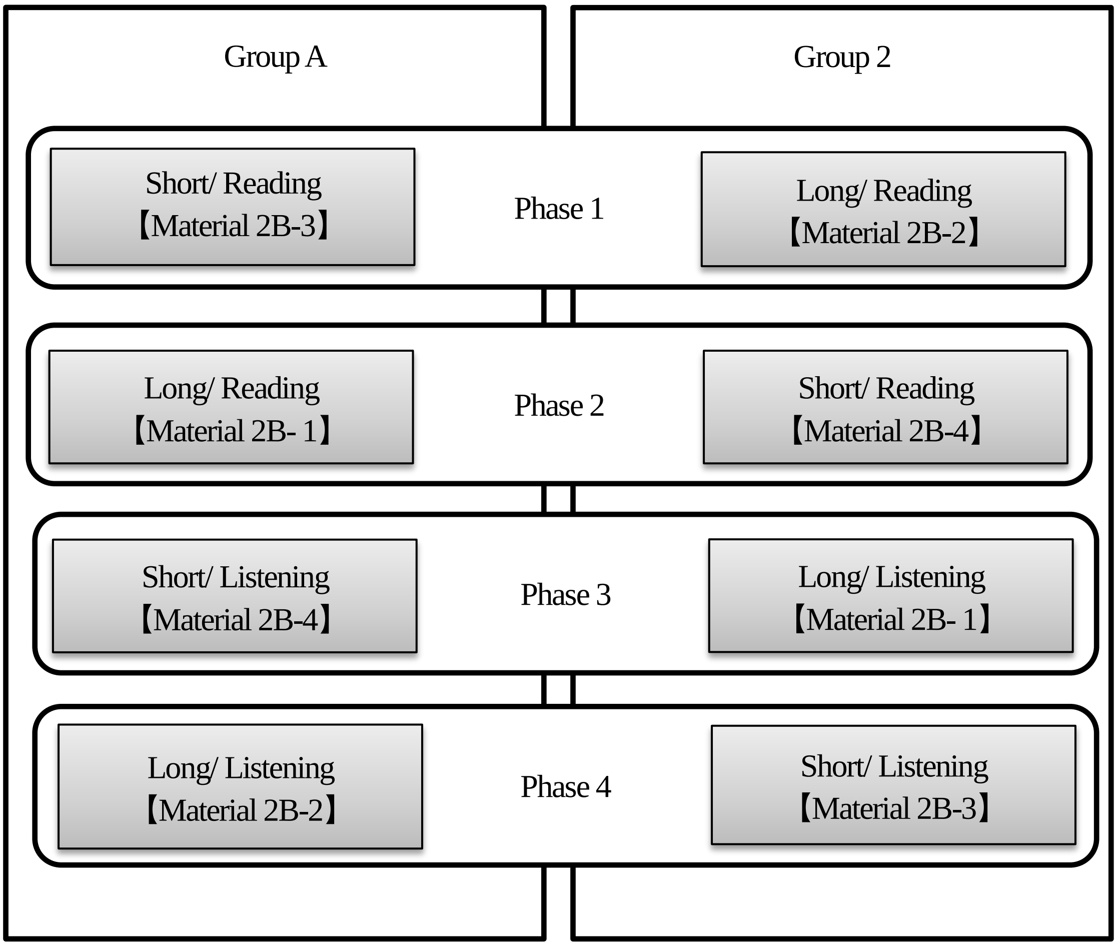


Figure 4.2. Flow, Task Conditions, and Materials Used in Experiment 2B.

RQ2B-1: Does the text length of input materials affect performance in retelling tasks?

RQ2B-2: Does the information presentation mode (reading or listening) affect performance in retelling tasks?

4.3.2 Participants

A total of 63 university students participated in this research. Data were collected during their English classes over the course of three weeks. Some students were absent for one or more of the classes; therefore, 10 data sets were excluded from the data and 53 data sets were used. All students were university freshmen, and had studied English for at least six years. Some had visited

English-speaking countries, but none had experience as a student in those countries.

4.3.3 Materials

Four passages were created to observe the influence of information input mode and text length. Table 4.8 shows the text information in the input materials. The audio files for the listening condition retelling tasks were developed using Globalvoice English® to control the speech rate. The topics of each material are (2B-1) using bikes to make a good city, (2B-2) buying locally produced food, (2B-3) care robots instead of care workers, and (2B-4) studying in a coffee shop, respectively.

Table 4.8

Input material data

	Text length	FKGL	FRE	Level 1 words (%)	Reproduction time (secs)	WPM
Material 2B- 1	149	7.8	65.0	79.9	62.1	146.1
Material 2B-2	149	7.8	63.0	78.5	60.9	146.9
Material 2B-3	98	7.6	64.8	88.0	40.5	146.8
Material 2B-4	98	7.6	59.3	80.8	40.9	143.7

Note. Level 1 words show lexical diversity; FRE refers to Flesch Reading Ease; FKGL refers to Flesch-Kincaid Grade Level; FKGLs and FREs were computed using Microsoft Word® 2010.

4.3.4 Experiment Procedure

Participants were divided into two groups. To avoid the effect of passage topic, the task conditions and topics were nested (Figure 4.2). All utterances were recorded using PCs, and all participants used headsets to record their utterances clearly and to exclude other participants' utterances.

4.3.5 Analysis Procedure

Transcription and coding were conducted by the author and four research collaborators, all of whom were studying English education and two of whom had master's degrees in English education and related study fields. First, the audio files were transcribed. The author transcribed all data (53 participants × 4 tasks), and the collaborators transcribed 53 audio files of the same topic. Having finished transcribing the assigned data, they added tags to the disfluency markers. The author collected the transcriptions and checked for inter-rater agreement. When disagreement occurred, the author and collaborator engaged in discussion to reach agreement. In case agreement could not be reached, a third collaborator gave their opinion and the case was decided by majority.

4.3.6 Data Analysis

4.3.6.1 Analysis of fluency

First, the indices of fluency were calculated. For fluency analysis, Tavakoli and Skehan (2005) have identified three categories of fluency: breakdown fluency, speed fluency, and repair fluency. In this study, words per minute (WPM) and syllables per minute (SPS) were used as indices of speed fluency. The numbers of one-second or longer silent pauses, pause length, and pause ratio were used to calculate breakdown fluency. In addition, the number of disfluency markers was used to analyze repair fluency. To count the number of tokens and types, the JACET 8000 analysis program—*v8an* (Shimizu, 2004) was used, and to count the syllables, the *Syllable Counter & Word Count–WordCalc.com* was used.

To automatically detect silent pauses, a noise reduction operation was first conducted. To exclude background noise, a noise reduction tool was used and a noise reduced version of sound files was created. To detect silent pauses, the *silence finder* of *audacity* was used. To detect the noise effectively, the threshold of the noise reduction cut-off point was set at 30db, and the threshold of silent pauses was set at less than 26dB. In this study, a silent pause was defined as a pause of one second or longer in accordance with Onoda (2014). Finally, the *silence finder* was used to find silent pauses, and the number of pauses was detected and counted.

4.3.6.2 Analysis of accuracy

In the current study, errors per AS-unit (Er/AS) were used as an accuracy index. To analyze the accuracy of utterances, the global errors were first counted. In this study, global errors were defined as syntactic or lexical errors that influence the listeners' comprehension. Minor errors such as lack of plural "s" or comprehensible word order errors were not counted as errors. Disfluency markers such as filled pauses, repetitions, reformulations, and false starts were not treated as errors because pruned transcription was used to count the errors. The number of AS-units was also counted using pruned transcription, and finally, the Er/AS was calculated. The author and research collaborators counted the number of errors and AS-units, and when disagreement occurred, consensus was reached through discussion.

4.3.6.3 Analysis of complexity

The complexity indices are categorized into lexical, morphological, syntactic, and phonological complexity. In the current study, lexical and syntactic complexity were analyzed. To analyze lexical complexity, the Guiraud index and type-token ratio (TTR) were selected. Furthermore, words per AS-unit was used as an index of syntactic complexity.

4.3.6.4 Rater evaluation

In addition to the CAF indices, it was necessary to conduct objective assessment. In the current study, the author acted as a rater and evaluated the participants' utterances using Empirically-derived, binary-choice boundary definition (EBB) scales for SRST (Hirai & Koizumi, 2013). Three criteria were analyzed: communicative efficiency, grammar and vocabulary, and pronunciation. One research collaborator evaluated part of the data (*N* = 53) in terms of communicative efficiency, grammar and vocabulary, and pronunciation. The inter-rater reliability between first and second raters was calculated, and the Cohen's kappa revealed substantial inter-rater reliability for communicative efficiency and pronunciation at κ = .804 and .821, respectively. The inter-rater reliability of grammar and vocabulary was not sufficient (κ = .638), but this task was not used as a high-stakes test; therefore, this result can be acceptable.

4.3.7 Results and Discussion

Two-way (Length x Mode) repeated analysis of variance (ANOVA) was conducted to ascertain the effect of text length and input mode on the EFL learners' oral performances in the retelling tasks. There were four independent variables: short text, long text, reading condition, and listening condition. Fifteen indices (number of word tokens, word types, number of syllables, number of one-second or longer silent pauses, total length of silent pauses, words per minute, syllables per second, number of disfluency markers, errors per AS-unit, Guiraud index, type token ratio, words per AS-unit, communicative efficiency, grammar and vocabulary, and pronunciation) were used in this analysis. The descriptive statistics of each factor and level are shown in Table 4.9, and the results of ANOVA in Tables 4.10 to 4.14. The results of subsequent analyses of simple main effect are shown in Table 4.15, and the results of main effect analysis in Tables 4.16 and 4.18.

Significant interactions were observed between the following indices: number of tokens, types, syllables, disfluency markers, pause length, pause ratio, WPS, SPS, Guiraud index, and pronunciation. To observe the detail of each factor, an analysis of simple main effects was conducted.

Significant interactions were not observed in the indices of number of pauses, errors per AS-unit, type-token ratio, words per AS-unit, Communicative Efficiency, or Grammar and Vocabulary. In the later part of this chapter, a detailed breakdown of the analysis will be provided.

Regarding utterance length, in other words the number of tokens, types, and syllables, these were found to have significant interactions respectively (F (1, 52) = 5.54, p = .022, η^2 = 0.01; F (1, 52) =

27.05, p = .000, η^2 =0.04; F (1, 52) = 492.87, p < .001, η^2 = 0.02). Significant simple main effects for Length under the Reading condition were found for Types and Syllables (F (1, 52) = 10.46, p = .002, η^2 =0.04, and F (1, 52) = 5.55, p = .022, η^2 = 0.02) and significant simple main effects for Length under the Listening condition were found for tokens, types, and syllables (F (1, 52) = 7.42, p = .009, η^2 = 0.03; F (1, 52) = 7.54, p = .008, η^2 = 0.04; F (1, 52) =6,77, p = .012, η^2 = 002) and under Mode were found for the Long condition (F (1, 52) = 8.9, p = .004, η^2 = 0.05; F (1, 52) = 19.05, p = .000, η^2 = 0.11; F (1, 52) = 13.13, p = .001, η^2 = 0.06). These results mean that test takers can produce more words when they receive information under the reading condition, which is consistent with the result of Experiment 2A. In reality, there was no significant simple main effect on Mode under the Short condition; it is certain there were large differences in the number and length of utterances in Mode under the Long condition. Thus, it is concluded that test takers can produce more words when given information under the reading condition. On the other hand, text length partially affected the number and length of utterances. The participants performed almost the same number and length of utterances under both the short and long reading conditions, but their performances drastically decreased under the latter. Koizumi and Hirai (2010) studied the effects of length of the story on the volume of utterances on the SRST. They expected test takers to produce more words when they received longer information, but the result was contrary to their expectations, and they concluded that a 50-word difference does not affect the number and length of utterances. Integrating the result of Koizumi and Hirai (2010) and experiment 2A with the current study, the volume of utterances can be affected by the text input mode or other factors, such as text topics. To clarify this result, an investigating survey should be conducted.

Table 4.9

The Mean Standard Deviations and 95% Confidence Intervals of Each Index

	Short				Long			
	Reading		Listening		Reading		Listening	
Indices	*M (SD)*	95% CI	*M (SD)*	95% CI	*M (SD)*	95% CI	*M (SD)*	95% CI
To	91.02 (36.87)	[54.15, 127.89]	89.26 (37.56)	[51.71, 126.82]	90.62 (29.20)	[61.42, 119.83]	78.13 (28.23)	[49.90, 106.36]
Ty	46.19 (13.13)	[33.06, 59.31]	49.06 (18.27)	[30.79, 67.32]	51.75 (13.61)	[38.14, 65.37]	42.96 (11.39)	[31.57, 54.36]
Sy	120.04 (46.96)	[73.08, 167.00]	127.45 (48.31)	[79.14, 175.77,]	132.87 (38.78)	[94.09, 171.65]	113.21 (38.89)	[74.32, 152.10]
NP	36.74 (8.88)	[27.86, 45.62]	33.68 (6.99)	[26.69, 40.67,]	32.64 (8.53)	[24.12, 41.17]	30.72 (5.94)	[24.77, 36.66]
PL	98.07 (32.18)	[65.89, 130.26]	67.04 (22.16)	[44.89, 89.20]	74.13 (27.92)	[46.21, 102.04]	77.59 (22.28)	[55.31, 99.87]
WPM	36.41 (14.75)	[21.66, 51.15]	35.71 (15.02)	[20.68, 50.73]	36.25 (11.68)	[24.57, 47.93]	31.25 (11.29)	[19.96, 42.54]
SPS	0.80 (0.31)	[0.49, 1.11]	0.85 (0.32)	[0.53, 1.17]	0.89 (0.26)	[0.63, 1.14,]	0.75 (0.26)	[0.50, 1.01]
DM	22.77 (18.15)	[4.62, 40.93]	24.57 (19.74)	[4.83, 44.30]	15.81 (10.29)	[5.52, 26.10]	9.53 (6.01)	[3.52, 15.54]
ErAS	0.20 (0.15)	[0.05, 0.34]	0.23 (0.21)	[0.02, 0.44]	0.26 (0.18)	[0.09, 0.44]	0.26 (0.18)	[0.08, 0.44]
GI	4.62 (0.62)	[4.00, 5.24]	4.77 (0.93)	[3.85, 5.70]	4.50 (0.79)	[3.71, 5.28]	4.05 (0.49)	[3.55, 4.54]
TTR	0.48 (0.10)	[0.38, 0.58]	0.51 (0.15)	[0.36, 0.65]	0.40 (0.08)	[0.32, 0.48]	0.39 (0.06)	[0.34, 0.45]
Wd/AS	7.08 (1.15)	[5.93, 8.23]	7.11 (1.64)	[5.47, 8.75]	7.79 (1.28)	[6.51, 9.08]	7.33 (1.23)	[6.10, 8.57]
CE	2.40 (0.86)	[1.55, 3.28]	2.42 (1.12)	[1.30, 3.53]	2.09 (0.95)	[1.15, 3.04]	2.06 (0.95)	[1.11, 3.01]
GV	2.23 (0.85)	[1.38, 3.07]	2.43 (1.05)	[1.39, 3.48]	2.17 (0.89)	[1.28, 3.06]	2.06 (0.91)	[1.15, 2.96]
Pro	2.74 (0.79)	[1.95, 3.52]	2.49 (0.97)	[1.52, 3.46]	2.45 (0.82)	[1.63, 3.27]	2.57 (0.84)	[1.72, 3.41]

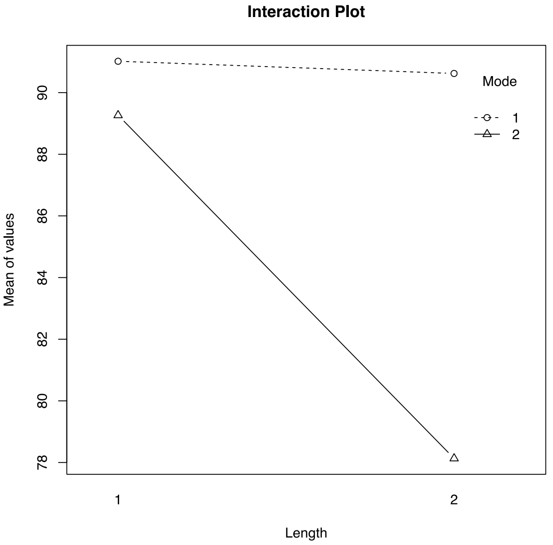
Note. To = number of word tokens, Ty = number of word types, Sy = number of syllables, NP = number of 1 second or longer silent pauses, PL = total length of silent pauses, WPM = words per minute, SPS = syllables per second, DM = number of disfluency markers, ErAS = errors per AS-unit, GI = Guiraud index, TTR = type token ratio, Wd/AS = words/AS-unit, CE = communicative efficiency, GV = grammar and vocabulary, Pro = pronunciation, and *N* =53.

Table 4.10

Two-Way Repeated ANOVA for Each Index of the Number of Utterances

Sources	*SS*	*MS*	*F*	*p*	η^2
			Tokens		
Length	1760.95	1760.95	3.34	.073	.01
Mode	2688.80	2688.80	3.29	.076	.01
Length x Mode	1527.17	1527.17	5.54	.022*	.01
Error	14327.58	275.53			
Total	235782.73	1117.45			
			Types		
Length	3.70	3.70	0.03	.873	.00
Mode	465.08	465.08	2.39	.128	.01
Length x Mode	1810.53	1810.53	27.05	.000*	.04
Error	3463.47	66.61			
Total	44964.98	213.1			
			Syllables		
Length	26.53	26.53	0.02	.876	.00
Mode	1986.80	1986.80	2.12	.151	.01
Length x Mode	9713.33	9713.33	19.67	.000*	.02
Error	25681.42	492.87			
Total	404630.5	1917.68			

Note. *N* =53, *df* = 1, 52, * *p* < .05.



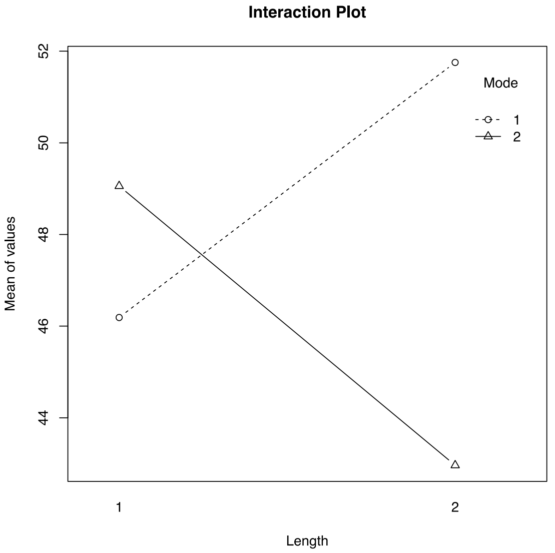


Figure 4.3. Interaction plot of number of word tokens.

Figure 4.4. Interaction plot of number of word types.

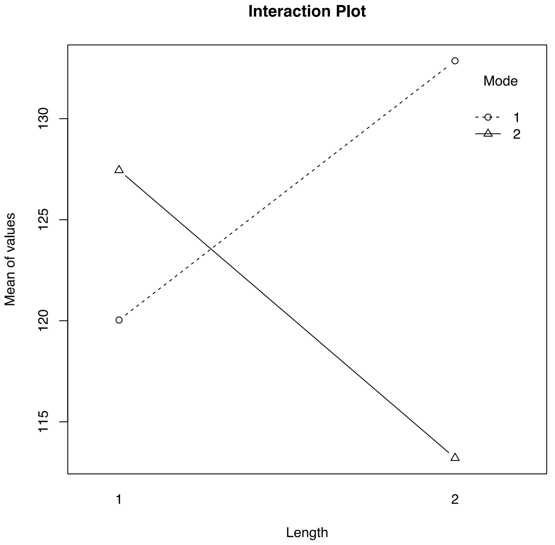


Figure 4.5. Interaction plot of number of syllables.

The next viewpoint is fluency, and detailed results are shown in Table 4.10. Significant interactions were found in pause length (F (1, 52) = 55.75, p = .000, η^2 = 0.09), WPM (F (1, 52) = 5.54, p = .022, η^2 = 0.01), SPS (F (1, 52) = 19.55, p = .000, η^2 = 0.02), and number of disfluency markers (*F* (1, 52) = 13.68, *p* = .001, η^2 = 0.02). Only the number of silent pauses did not have significant interaction (*F* (1, 52) =16.98, *p* = .535, η^2 = 0.00). A simple main effect analysis was conducted to observe the effects of each task condition, revealing that all fluency-related variables have significant simple main effects on Length under the Listening condition. There were also significant simple main effects on Length under the Reading condition for pause length (*F* (1, 52) = 33.51, *p* = .000, η^2 = 0.14), SPS (*F* (1, 52) = 5.56, *p* = .022, η^2 = 0.02), and number of disfluency markers (*F* (1, 52) = 15.61, *p* = .000, η^2 = 0.21). Pause length had a significant main effect and learners tended to produce more fluent speech under the short passage conditions. The results revealed that high speed fluency can be observed under the long listening condition, and pause length and number of disfluency markers decrease under the listening condition. In other words, the participants did not spend much time on silent and filled pauses under the long listening condition. The participants may be able to store serial words or sounds effectively under the listening condition, and hence they performed well under this task condition. Another reason for high fluency under the listening condition may be the task sequence. In the current study, the data collection for the listening condition occurred in the latter half of the experiment. Therefore, the participants' fluency may be improved under the listening condition. This should be treated as a limitation of the present study.

Table 4.11

Two-Way Repeated ANOVA for Each Index of Fluency

Sources	*SS*	*MS*	*F*	*p*	η^2
			No. of silent pauses		
Length	659.32	659.79	17.16	.000*	.05
Mode	328.75	328.75	8.12	.006*	.03
Length x Mode	16.98	16.98	0.39	.535	.00
Error	2262.02	43.50			
Total	13262.32	62.85			
			Pause length		
Length	2378.92	2378.92	4.35	.042*	.01
Mode	10068.25	10068.25	19.41	.000*	.06
Length x Mode	15764.91	15764.91	55.75	.000*	.09
Error	14704.66	282.78			
Total	172931.19	834.32			
			WPM		
Length	281.75	281.75	3.34	.073	.01
Mode	430.21	430.21	3.29	.076	.01
Length x Mode	224.35	224.35	5.54	.022*	.01
Error	2292.41	44.08			
Total	37725.24	178.79			

(Table continues)

Table 4.11—Continued

			SPS		
Length	0.001	0.001	0.02	.005*	.00
Mode	0.09	0.09	2.17	.146	.01
Length x Mode	0.43	0.43	19.55	.000*	.02
Error	1.14	0.02			
Total	17.96	0.09			
			No. of disfluency markers		
Length	6413.00	6413.00	35.65	.000*	.12
Mode	267.19	267.19	6.03	.017*	.01
Length x Mode	864.08	864.08	13.68	.001*	.02
Error	3283.92	63.15			
Total	52319.89	247.96			

Note. N =53, *df* = 1, 52, * *p* < .05.

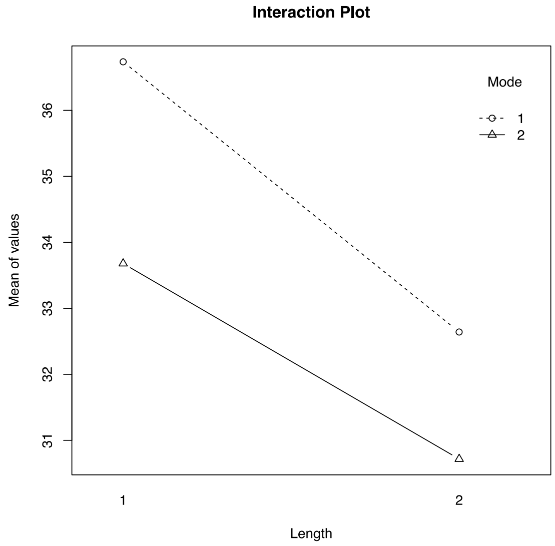


Figure 4.6. Interaction plot of number of 1 second or longer pauses.

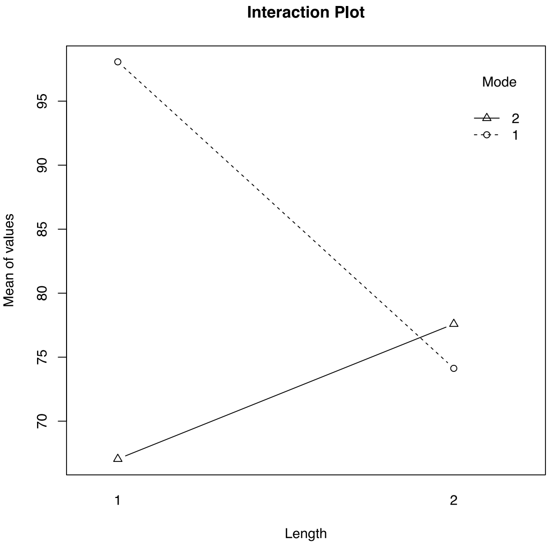


Figure 4.7. Interaction plot of pause length.

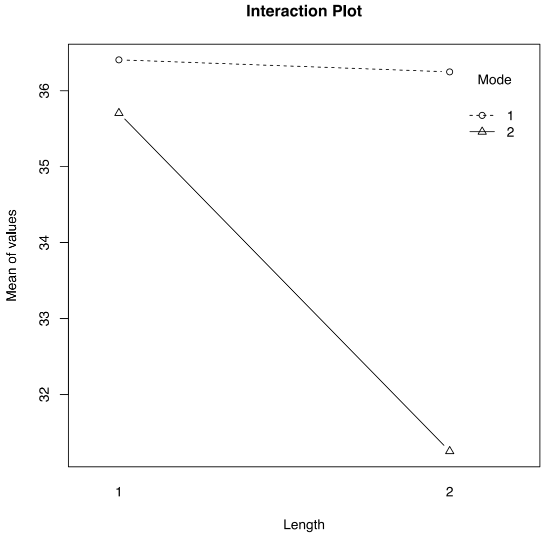


Figure 4.8. Interaction plot of WPM.

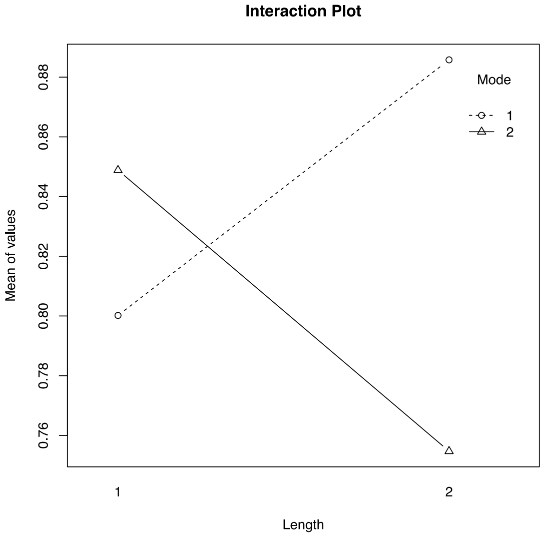


Figure 4.9. Interaction plot of SPS.

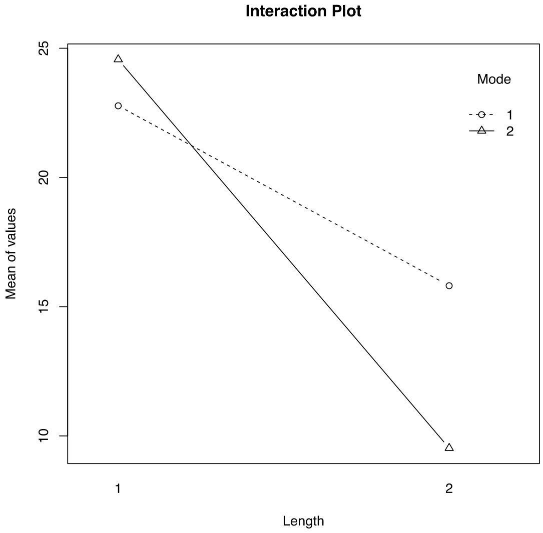


Figure 4.10. Interaction plot of number of disfluency markers.

Regarding accuracy, the Errors per AS-unit were compared (Table 4.12), and no significant interaction was found (F (1, 52) = 5.02, p = .029, η^2 = 0.02). The main effect analysis revealed that there is no main effect on each independent variable (F (1, 52) = 5.02, p = .290, η^2 = 0.02; F (1, 52) = 0.26, p = .614, η^2= 0.00). This means that participant accuracy was not affected by task condition.

Table 4.12

Two-Way Repeated ANOVA for Index of Accuracy

Sources	*SS*	*MS*	*F*	*p*	η^2
			Errors per AS-unit		
Length	0.13	0.13	5.02	.029*	.02
Mode	0.10	0.10	0.26	.614	.00
Length x Mode	0.02	0.02	1.04	.312	.00
Error	1.08	1.08			
Total	6.94	0.03			

Note. N =53, *df* = 1, 52, * *p* < .05

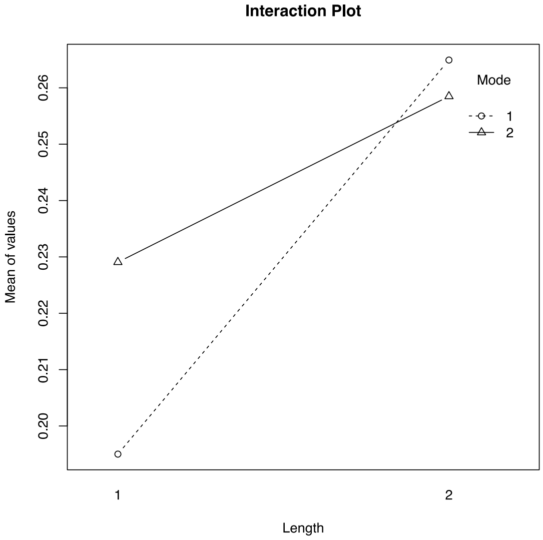


Figure 4.11. Interaction plot of errors per AS-unit.

In the analysis of complexity, the Guiraud index and type token ratio were used to observe the effects of task condition on lexical complexity, and words per AS-unit was used to observe the effect of syntactic diversity. The Guiraud index had significant interaction (F (1, 52) = 12.82, p = .001, η^2= 0.04), and has the simple main effect on Mode under the long condition (F (1, 52) = 13.71, p = .000, η^2 = 0.11) and listening condition (F (1, 52) = 32.93, p = .000, η^2 = 0.02). On the other hand, there was no significant interaction on the type token ratio (F (1, 52) = 0.23, p = .274, η^2 = 0.01, see details in

Table 4.13). These two results indicated different conditions, but may have been caused by the participants' productive vocabulary size. Vermeer (2000) claimed that the Guiraud index could be an effective measurement method for lexical richness, and this index targets learners who have a vocabulary of less than 3,000 words. Thus, this result cannot be generalized as the task condition may have affected participants' lexical diversity. Regarding syntactic complexity, there was also no significant difference (*F* (1, 52) = 2.27, *p* = .138, η^2 = 0.01). Subsequent analysis revealed a significant main effect on Length, and the participants could use a more complex structure under the Long condition.

Table 4.13

Two-Way Repeated ANOVA for Each Index of Complexity

Sources	*SS*	*MS*	*F*	*p*	η^2
			Guiraud index		
Length	9.59	9.59	19.66	.000*	0.08
Mode	1.19	1.19	2.39	.128	0.04
Length x Mode	4.70	4.70	12.82	.001*	0.04
Error	19.06	0.37			
Total	124.68	0.59			
			Type token ratio		
Length	0.51	0.51	77.1	.000*	0.19
Mode	0.03	0.03	0.23	.633	0.00
Length x Mode	0.01	0.01	0.23	.274	0.01
Error	0.56	0.01			
Total	2.67				

(Table continues)

Table 4.13—Continued

			Words per AS-unit		
Length	11.73	11.73	8.12	.006*	0.03
Mode	2.42	2.42	1.51	.225	0.01
Length x Mode	3.21	3.21	2.27	.138	0.01
Error	73.52	1.414			
Total	11387.23	4201.20			

Note. *N* =53, *df* = 1, 52, * *p* < .05.

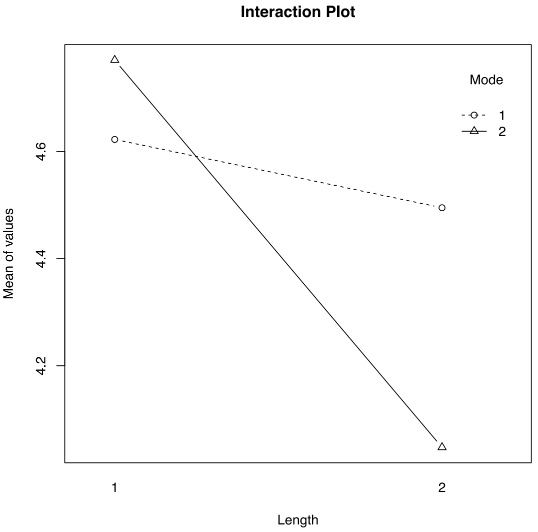


Figure 4.12. Interaction plot of Guiraud index.

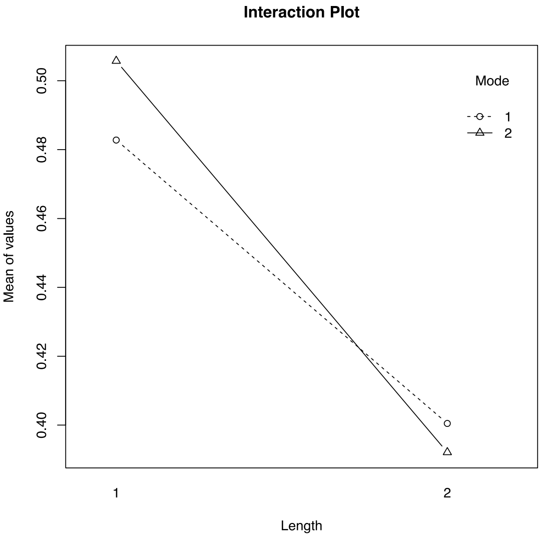


Figure 4.13. Interaction plot of type token ratio.

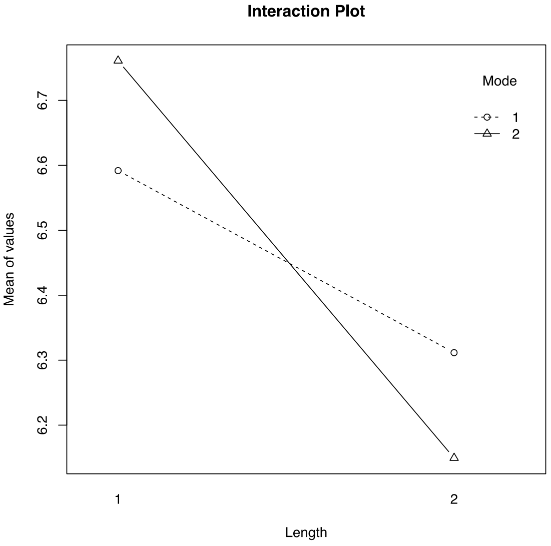


Figure 4.14. Interaction plot of words per AS-unit.

Finally, the results of rater evaluation were compared, and a significant interaction was found for pronunciation (F (1, 52) = 8.81, p = .005, η^2 = 0.01), but none for communicative efficiency (F (1, 52) = 0.06, p = .805, η^2 = 0.00) or grammar and vocabulary (F (1, 52) = 3.31, p = .074, η^2 = 0.01, Table 4.14). Since pronunciation could not change drastically within the three-week data collection period, this result may comprise noise. It is natural that the index of grammar and vocabulary does not have significant interaction because at least the index of errors per AS-unit and type token ratio does not have significant interaction, nor does the index of communicative efficiency. The criterion of communicative efficiency includes the construct of three types of fluency and the number and length of utterance; thus, non-significant interaction on communicative efficiency is also natural.

Table 4.14

Two-Way Repeated ANOVA for Each Result of Rater Evaluation

Sources	*SS*	*MS*	*F*	*p*	η^2
			Communicative efficiency		
Length	6.11	6.11	11.01	.002*	0.03
Mode	0.02	0.02	0.03	.855	0.00
Length x Mode	0.02	0.02	0.06	.805	0.00
Error	15.98	0.31			
Total	203.25	0.96			
			Grammar & vocabulary		
Length	2.50	2.50	6.41	.014*	0.01
Mode	0.12	0.12	0.19	.662	0.00
Length x Mode	1.36	1.36	3.31	.074	0.01
Error	21.39	21.39			
Total	182.58	0.87			
			Pronunciation		
Length	0.57	0.57	2.65	.109	0.00
Mode	0.23	0.23	0.83	.367	0.00
Length x Mode	1.70	1.70	8.81	.005*	0.01
Error	10.05	0.19			
Total	156.2	0.74			

Note. N =53, *df* = 1, 52, * *p* < .05.

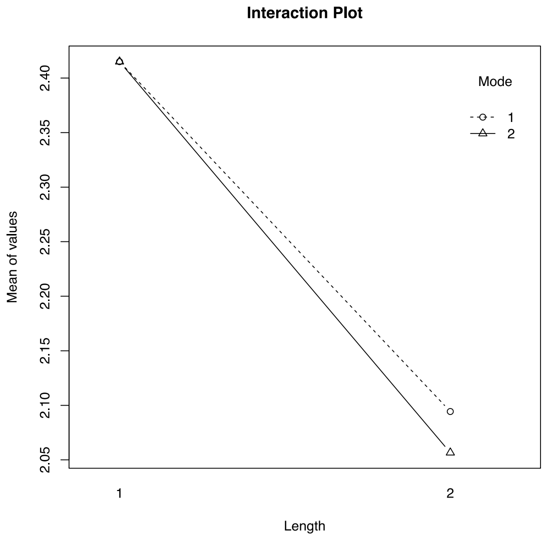


Figure 4.15. Interaction plot of the grade of communicative efficiency.

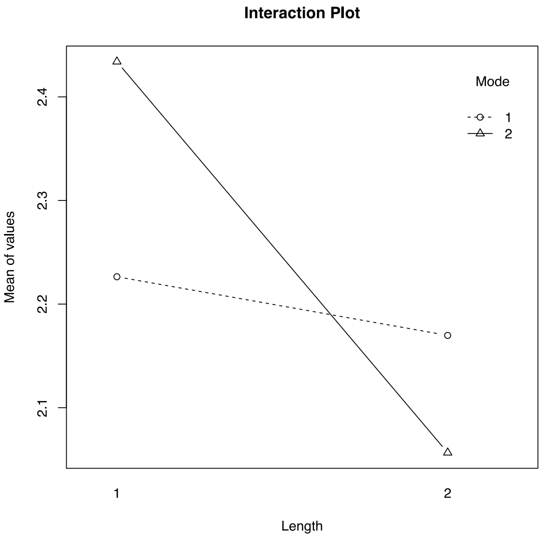


Figure 4.16. Interaction plot of the grade of grammar & vocabulary.

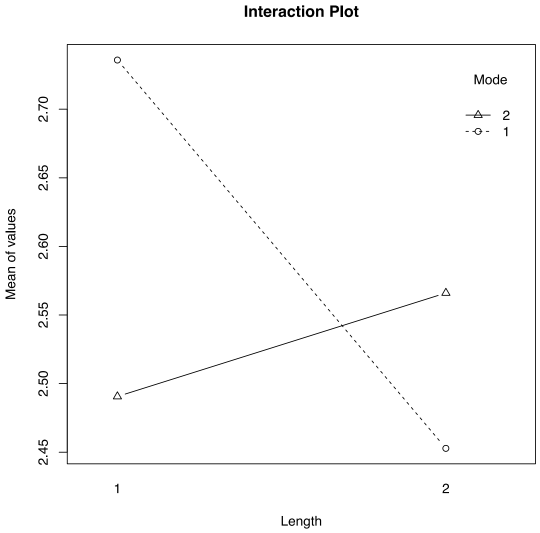


Figure 4.17. Interaction plot of the grade of pronunciation.

Thus, length of input text and input mode affects the volume of utterances and degree of fluency. The participants performed more words under the reading condition, but performed their speaking skills more fluently under the listening condition. The reason they could not perform more words under the listening condition might be based on the immediacy of the listening input. Since they could not store the contents by listening twice, the number and length of utterances decreased under the listening condition. On the other hand, their fluency improved under this condition, perhaps because they could listen to and learn the serial sounds and produce fluent sounds based on the listening materials.

Table 4.15

Simple Main Effects: Length × Mode

Simple Main Effects	*SS*	*MS*	*F*	*p*	η^2
			Tokens		
Mode under short condition	81.59	81.59	0.13	.720	0.00
Mode under long condition	4134.38	4134.38	8.91	.004*	0.05
Length under Reading condition	4.16	4.16	0.01	.920	0.00
Length under Listening condition	3283.96	3283.96	7.42	.008*	0.03
			Types		
Mode under short condition	217.96	217.96	1.42	.239	0.01
Mode under long condition	2048.64	2048.64	19.05	.001*	0.11
Length under Reading condition	820.99	820.99	10.46	.002*	0.04
Length under Listening condition	984.24	984.24	7.54	.008*	0.04

(Table contines)

Table 4.15—Continued

Syllables					
Mode under short condition	1457.07	1457.07	2.25	.140	0.01
Mode under long condition	10243.06	10243.06	13.13	.001*	0.06
Length under Reading condition	4362.26	4362.26	5.55	.022*	0.02
Length under Listening condition	5377.59	5377.59	6.77	.012*	0.03
Pause length					
Mode under short condition	25515.19	25515.19	54.84	.000*	0.24
Mode under long condition	317.97	317.97	0.95	.335	0.00
Length under Reading condition	15195.92	15195.92	33.51	.000*	0.14
Length under Listening condition	2947.91	2947.91	7.83	.007*	0.05
WPM					
Mode under short condition	13.06	13.06	0.13	.720	0.00
Mode under long condition	661.50	661.50	8.90	.004*	0.05
Length under Reading condition	0.67	0.67	0.91	.920	0.00
Length under Listening condition	525.43	525.43	7.42	.009*	0.03
SPS					
Mode under short condition	0.06	0.06	2.19	.145	0.01
Mode under long condition	0.46	0.46	13.13	.001*	0.06

(Table continues)

Table 4.15—Continued

Length under Reading condition	0.19	0.19	5.56	.022*	0.02
Length under Listening condition	0.23	0.23	6.67	.013*	0.03
		No. of disfluency markers			
Mode under short condition	85.14	85.14	1.31	.257	0.00
Mode under long condition	1046.12	1046.12	24.49	.000*	0.12
Length under Reading condition	1284.54	1284.54	15.61	.000*	0.05
Length under Listening condition	5992.54	5992.54	37.28	.000*	0.21
		Guiraud index			
Mode under short condition	0.58	0.58	1.22	.274	0.01
Mode under long condition	5.30	5.30	13.71	.001*	0.11
Length under Reading condition	0.43	0.43	1.00	.323	0.01
Length under Listening condition	13.85	13.85	32.93	.000*	0.20
		Pronunciation			
Mode under short condition	1.59	1.59	8.37	.006*	0.02
Mode under long condition	0.34	0.34	1.20	.278	0.00
Length under Reading condition	2.12	2.12	11.77	.001*	0.03
Length under Listening condition	0.15	0.15	0.66	.419	0.00

Note. N =53, *df* = 1, 52, * *p* < .05.

Table 4.16

Main Effects for Each Index of Mode to Length

	Short		Long					
	M	*SE*	*M*	*SE*	95% CI	*F*	*p*	η^2
No. of silent pauses	35.21	0.87	31.68	0.82	[1.82, 5.24]	17.16	.000*	0.05
Errors per AS-unit	0.21	0.02	0.26	0.18	[-0.09, -0.01]	5.02	.290	0.02
Type token ratio	0.49	0.11	0.40	0.08	[0.09, 0.12]	77.11	.000*	0.19
Words per AS-unit	7.09	0.15	7.56	0.13	[-0.80, -0.14]	8.12	.006*	0.00
Communicative efficiency	2.42	0.12	2.08	0.10	[0.13, 0.55]	11.01	.002*	0.03
Grammar & vocabulary	2.33	0.10	2.11	0.11	[0.05, 0.39]	6.41	.014*	0.00

Note. N =53, *df* = 1, 52, * *p* < .05.

Table 4.17

Main Effects for Each Index of Length to Mode

	Reading		Listening					
	M	*SE*	*M*	*SE*	95% CI	*F*	*p*	η^2
No. of silent pauses	34.69	1.01	32.20	0.66	[0.74, 4.25]	8.12	.006*	0.03
Errors per AS-unit	0.23	0.01	0.24	0.02	[-0.07, 0.04]	0.26	.614	0.00
Type token ratio	0.44	0.01	0.45	0.01	[-0.04, 0.02]	0.23	.633	0.00
Words per AS-unit	7.44	0.13	7.22	0.16	[-0.14, 0.56]	1.51	.225	0.00
Communicative efficiency	2.26	0.11	2.24	0.10	[-0.19, 0.23]	0.04	.034*	0.00
Grammar & vocabulary	2.20	0.11	2.25	0.12	[-0.26, 0.17]	0.19	.662	0.00

Note. N =53, *df* = 1, 52, * *p* < .05.

The purpose of using a retelling task as a speaking assessment tool is to elicit participants' speaking performance and give corrective feedback to students in terms of language use. Therefore, it is recommended to use the reading input as a stimulus for the retelling task to elicit more utterances. Accuracy, part of the lexical complexity (type token ratio), and syntactic complexity were not affected by the task condition; these factors were affected by the test taker's own performance.

4.3.8 Conclusion of Experiment 2B

This study was conducted to investigate the effects of text length and input mode on performance in retelling tasks. To ascertain the effects of those factors, in total 15 indices of complexity, accuracy, fluency, and rater evaluations were analyzed. The analysis revealed that participants could produce more utterances when they received the information under the reading condition, but performed more fluently under the listening condition. On the other hand, the accuracy, complexity, and rater evaluation were not affected by the task condition; thus, those factors cannot be controlled by the task condition. The results indicate that teachers should avoid using listening-based retelling tasks when teaching novice or low-intermediate learners. Because learners cannot perform well under the listening condition, teachers or test developers should use reading-based retelling tasks for low-proficiency learners if they want to elicit longer utterances.

There are some limitations to this study, including task sequence and the lack of information about participants' proficiency. If detailed scores had been available for their reading and listening proficiency, more detailed interpretations could have been included. The results of the current study did not examine the exact difficulty of each factor; therefore, the difficulty of each task condition should be examined in future studies.

4.4 Summary of Study 2

Study 2 aimed to ascertain the effects of text length, difficulty, and input mode on spoken performances, while experiment 2A revealed how the length of input text and information mode influenced spoken performances in terms of the length of utterances, and grades given by raters. This result matched with the expectation of Koizumi and Hirai (2010), and thus, longer texts might comprise an easy task for participants. However, there are serious limitations to this experiment, including the unbalanced material design. To solve this limitation and ascertain more appropriate effects of the factors of text length of material and input mode, experiment 2B was conducted. The results of experiment 2B show that the study participants could produce many words under the condition of retelling, while listening input elicited fluent utterances.

Both experiments produced interesting results, and the WPM indices were almost similar in all analyses. Namely, most participants uttered almost the same information; additionally, experiment 2B revealed that the average pause length detected in all conditions was 79.20 seconds, meaning that the participants could not say anything during the answering time (150 seconds) of this experiment.

Other findings of this study are that performances of grammatical accuracy and lexical and syntactic complexity do not change in any task condition. Thus, accuracy and complexity do not change even if test developers change the length, difficulty, and input mode. These results indicated that if test developers or teachers want to elicit more sophisticated performances, clues of grammar or expressions should be provided in the retelling tasks.

Chapter 5

Study 3: The Effects of Pre-Task Planning and Reading Aloud During Pre-Task on Spoken Performance in Retelling Tasks

5.1 Introduction to Study 3

This chapter identifies solutions to alter the difficulty of retelling tasks other than changing the text itself or input mode. Previous research shows that adding participants' preparation time can reduce the difficulty of retelling tasks as well as elicit a more sophisticated performance. In Study 3, two treatments are used as pre-task task conditions: (1) providing increased preparation time and imposing that input material must be read aloud. This study seeks to show that the former condition may activate test takers' function of conceptualization, thereby improving their performance; subsequently, the latter condition can aid in practice of pronunciation, which could help activate the function of articulation. In Experiment 3A, the influence of pre-task preparation to the spoken performance is observed, and in Experiment 3B, the effects of pre-task reading aloud to the spoken performance in retelling are observed.

5.2 Experiment 3A: Influence of Pre-Task Planning to Spoken Performance in Retelling Tasks

5.2.1 Objective and Research Questions

Previous research has explored the effectiveness of planning (such as pre-task preparation) on language performance in many types of tasks (Ahangari & Abdi, 2011; Li, Chen, & Sun, 2014; Rouhi & Marefat, 2006; Yuan & Ellis, 2000). These studies show that pre-task planning can be one of the easiest modifications to adjust task difficulty. Experiment 3A explores the effects of preparation time to spoken performance in retelling tasks and observes the effects of pre-task preparation. Following are the research questions of Experiment 3A:

RQ3A-1: Can test-takers produce more words when they have one minute of preparation time than when they do not have any preparation time?

RQ3A-2: Are test-takers' spoken performances—in terms of complexity, accuracy, and fluency—improved when they have one minute of preparation time?

5.2.2 Participants

All participants in this study are English-language learners of Japanese origin between the ages of 19 and 20 years old who have taken one or more TOEIC preparation course. All participants study technology and engineering and have taken at least seven years of English-language education across junior and senior high school and college. All had experience in answering spoken (oral) examinations as preparation for the TOEIC Speaking and Writing Test®. Twenty-nine students participated in this experiment, but eight participants' data were excluded from the analysis due to either significant background noise or that the participant's voice was indistinct, even after amplification; in brief, 21 participants' data were used for analysis.

5.2.3 Materials

In this study, two reading passages were used (Table 5.1). Excluding the influence of text length and difficulty, two materials were controlled as being the same level.

Table 5.1

Materials Used in Experiment 3A

Material	Wordage	FKGL	FRE	Level 1 words (%)
Material 3A-1	149	7.8	65.0	79.9
Material 3A-2	149	7.8	63.0	78.5

5.2.4 Procedure

In order to counterbalance the practice effects of answering to the retelling tasks, participants were divided into two groups: 3Aa and 3Ab. Group 3Aa read material 3-A silently within two minutes. Following this silent reading session, one minute of preparation time was given prior to the first task. During this preparation time, however, they were not allowed to write down the contents and were prohibited from talking with other participants, to avoid introducing factors that would negatively affect participants' speaking skills being assessed correctly. After the preparation time, they retold the story and also provided their impressions of the story within two minutes and thirty seconds. A three-minute break was given, and then Group 3Aa read material 3-B within a timeframe of two minutes and then answered retelling tasks. Group 3Ab engaged in silent reading prior to the retelling tasks with no preparation time provided, and then took a three-minute break. After the break, they engaged in retelling tasks following one minute of preparation time.

5.2.5 Analysis of the Procedure

To analyze participants' performances in this experiment, first, their utterances were transcribed by the author and research collaborator (who have collaborated previously on experiments within this series of study). Tags of disfluency markers were also added into the script and then counted. Next, the utterances were divided into AS-units to analyze accuracy (errors per AS-unit) and syntactic complexity (words per AS-unit), wherein pauses one second or longer were counted and calculated for length. Other indices such as number of word tokens, number of word types, and number of syllables were detected using a syllable counter and word count program (www.wordcalc.com). Lastly, WPM, SPS, Guiraud index, and TTR were calculated.

5.2.6 Results and Discussion

Table 5.2 shows the means, standard deviations and 95% confidence intervals of two different tasks conditions. Most indices may initially appear to be improved in preparation condition; however, the *t*-test shows significance differences across the indices of number of syllables, number of pauses one second or longer, and syllables per second. Results from this experiment reveal that participants performed fluently under the time condition that included preparation time. Other indices of fluency, such as pause length, WPM, and number of disfluency markers were not significantly different. Similar to the results of studies 2A and 2B, the indices of lexical complexity and lexical errors were not significantly different between the task conditions.

Table 5.2

Means, Standard Deviations, and 95% CIs in Each Task Condition (Experiment 3A)

	No preparation time condition		Preparation time condition	
	M (SD)	95% CI	*M (SD)*	95% CI
To	50.50 (30.43)	[20.07, 80.93]	60.20 (32.76)	[27.44, 92.96]
Ty	28.95 (14.15)	[14.80, 43.10]	32.65 (14.41)	[18.23, 47.07]
Sy	71.00 (44.00)	[27.00, 115.00]	88.70 (43.13)	[45.57, 131.83]
NP	26.45 (7.07)	[19.38, 33.52]	22.35 (8.83)	[13.52, 31.18]
PL	132.33 (16.94)	[115.39, 149.28]	128.73 (13.39)	[115.34, 142.13]
WPM	20.20 (12.17)	[8.03, 32.37]	24.08 (13.10)	[10.98, 37.18]
SPS	0.47 (0.29)	[0.18, 0.77]	0.59 (0.29)	[0.30, 0.88]
DM	16.70 (9.30)	[7.40, 26.00]	18.35 (12.36)	[5.99, 30.71]
ErAS	1.21 (1.63)	[-0.42, 2.85]	1.23 (0.65)	[0.58, 1.89]
GI	4.05 (0.81)	[3.24, 4.87]	4.20 (0.78)	[3.42, 4.98]
TTR	0.62 (0.13)	[0.49, 0.75]	0.60 (0.15)	[0.45, 0.74]
WdAS	6.51 (1.72)	[4.80, 8.23]	6.70 (1.56)	[5.15, 8.26]

Note. To = number of word tokens, Ty = number of word types, Sy = number of syllables, NP =

number of silent pauses one second or longer, PL = pause length, WPM = words per minute, SPS = syllables per second, DM = number of disfluency markers, ErAS = errors per AS-unit, GI = Guiraud index, TTR = type token ratio, WdAS = words/AS-unit, and *N* = 21.

With regard to answering this experiment's research questions, these results mean that the effects of preparation in retelling tasks are not significant, as participants of this study were able to perform similarly across two different task conditions (RQ3A-1) using the same text. In addition, the quality of performance—at least in terms of indices of speed fluency—partially improved in the case where preparation time was given to participants (RQ3A-2).

Table 5.3

Results of Paired T-Test Comparing Two Different Task Conditions (Experiment 3A)

Indices	*t*	*p*	*d*	95% CI
To	-1.64	.118	0.31	[-22.08, 2.68]
Ty	-1.30	.208	0.26	[-9.64, 2.24]
Sy	-2.33	.031*	0.41	[-33.59, -1.80]
NP	2.08	.050*	0.51	[-0.00, 8.20]
PL	0.90	.374	0.23	[-4.65, 11.85]
WPM	-1.63	.118	0.30	[-8.83, 1.07]
SPS	-2.33	.031*	0.41	[-0.22, -0.01]
DM	-0.89	.382	0.15	[-5.51, 2.21]
ErAS	-0.05	.959	0.02	[-0.75, 0.72]
GI	-0.73	.473	0.18	[-0.56, 0.27]

(Table continues)

Table 5.3—Continued

TTR	0.68	.499	0.15	[-0.04, 0.08]
WdAS	-0.45	.654	0.11	[-1.05, 0.68]

Note. To = number of word tokens, Ty = number of word types, Sy = number of syllables, NP = number of silent pauses one second or longer, PL = pause length, WPM = words per minute, SPS = syllables per second, DM = number of disfluency markers, ErAS = errors per AS-unit, GI = Guiraud index, TTR = type token ratio, WdAS = words/AS-unit, and *N* = 21.

The results of this study are similar to those found by Yuan and Ellis (2003), who demonstrated that pre-task planning enhances grammatical complexity of oral performance, where participants performed more words fluently and used a variety of vocabulary terms in their speech. The performance of participants in this current experiment corroborate these previous findings, as participants performed more words fluently under the pre-task preparation condition. Li, Chen and Sun (2014) claimed that such planning has a positive impact on performance in terms of fluency, accuracy and complexity, and they emphasized that accuracy was the attribute most significantly improved in sessions that included one minute of planning time. Planning or preparation times that are too long or too short, however, are not as effective in eliciting better test-taking performance. The results of this study indicated that only fluency and the length of utterances were improved under the preparation condition, which was restricted to one minute of preparation time. This particular duration of pre-task preparation was effective for participants, however, the effects are limited and contrary to Li, Chen and Sun's study, as accuracy and complexity were not improved. Ahangari and Abdi (2011) also observed the effects of pre-task planning on spoken performance, finding that the complexity of oral performance was improved following preparation time; conversely, their study revealed no effect on accuracy. Rouhi and Marefat (2006) also studied the effects of planning on spoken performance, finding that planning did have an effect on fluency and accuracy; however, their experiment provided ten minutes of preparation time, a duration that may have strongly affected participants' spoken performance. Due to the differences among previous study results, the results of this current experiment should not be generalized.

Participants in this experiment performed more words under the pre-task preparation condition, which can enable teachers to provide more detailed feedback following the tasks; as such, even if there is no effect on accuracy and complexity of spoken performance, this task condition can be useful to elicit a higher number and quality of utterance from learners.

5.2.7 Summary of Experiment 3A

The purpose of this experiment was to see the effects of pre-task planning (or preparation time) on the spoken performance during retelling tasks. The results of this experiment reveal that pre-task planning elicited more words, or utterances; in addition, fluency in terms of SPS was significantly improved over the condition with no preparation time. Other indices of fluency, however, did not improve under this task condition or were invariant, such as accuracy and complexity.

5.3 Experiment 3B: Influence of Reading Aloud During Pre-Task on Task Performance

5.3.1 Objectives and Research Questions

The purpose of the current study is to investigate the effects of reading aloud during the pre-task period on retelling task performance. Specifically, if the test-taker reads the text aloud before answering the retelling tasks, will more sophisticated utterances, in terms of pronunciation and fluency, be produced? Following are the research questions of Experiment 3B:

RQ3B-1: Can test-takers produce more words when they read the input text aloud?

RQ3B-2: Does test-takers' speech production improve in terms of complexity, accuracy, fluency, and pronunciation when they read the input text aloud?

5.3.2 Participants

Twenty students participated in this experiment, though two data sets were excluded because of recording failures; therefore, the remaining 18 participants' speech data were used for the analysis. All participants were freshman undergraduate students that had been learning English for at least six years across both junior and senior high school. None of the participants had experience participating in an overseas study program. The students' English proficiency level was considered intermediate, based on placement test allocation.

5.3.3 Materials

Three materials were prepared for this experiment: Material 3B-1 was used as a practice task, while Materials 3-B2 and 3B-3 were used for analysis. Materials 3B-2 and 3B-3 each contained 149 words, and their readability was set to approximately the same level (see Table 5.4). Input materials and directions were included on a worksheet that was created with reference to the SRST (Hirai & Koizumi, 2008, 2009, 2013; Koizumi & Hirai, 2010, 2012). Materials were shown on the front of the worksheet while four keywords were included on the back of the worksheet to remind participants of the content. Participants read the text on the worksheet silently within a timeframe of two minutes; following, they retold the story and provided their own opinions or impressions within two and a half minutes. In the read-aloud condition, they read the text silently within two minutes and then read it aloud once. Soon after the read-aloud session, they retold the story and provided their opinions and impressions about the topic discussed in the materials text (see Appendix J, I, and H, respectively).

Table 5.4

Text Information of Materials 3B-1 and 3B-2

	Words	FKGL	FRE	Level 1 words (%)
Material 3B-1	98	7.6	64.8	88.0
Material 3B-2	149	7.8	65.0	79.87
Material 3B-3	149	7.8	63.0	78.52

Note. FKGL = Flesh Kincaid Grade Level, FKE = Flesh Reading Ease, and Level 1 words shows the proportion of level 1 words based on the JACET 8000 word list.

5.3.4 Procedure

The author initially provided instructions about the task in English along with the relevant materials. To avoid any effects of practice on the results, the students were divided into two groups, as in Experiment 3A. In the first session, Group 3Ba was assigned Material 3B-2, which they read silently for three minutes and then retold the story. Group 3Bb was assigned to read the same material as Group 3Ba within two minutes, and after the silent reading, they read aloud the material and retold the story that they had read. After the first session, the materials and task conditions were reversed between the groups and the experiment was repeated. Actually, the reading time for silent group differ from other experiments conducted in this series of study, under the condition of reading aloud, the participants have more time to read; therefore, three minutes of reading duration was given to the control group: no reading aloud group.

5.3.5 Analysis of the Procedure

In this experiment, similar indices were used in both Experiment 3A and 3B, but the analysis of the rater's assessment was different in Experiment 3B. In addition to the CAF analysis, evaluation by two raters was employed for this study who each gathered their impression of test-takers' speech. The empirically derived, binary-choice, boundary definition scale (EBB) was used for raters' evaluations. The characteristics of the retelling task were identical to the SRST; consequently, the items on the EBB2 scale, developed by Hirai and Koizumi (2013), were used as the criteria. Utterances were scored on a scale of 1 (worst) to 5 (best), and the author and one research collaborator rated all performances. The inter-rater reliability was high and sufficient based on Cronbach's α of communicative efficiency (0.830), grammar and vocabulary (0.851) and pronunciation (0.906).

The speech data were divided into two groups: no read-aloud condition and read-aloud condition. The two conditions were compared using a paired *t*-test. The significance level was set to $p < 0.05$. Before the author conducted the *t*-test, the *a priori* power analysis was conducted. The expected effect size was set at $d = 0.5$ and the power analysis revealed that at least 34 data were necessary to secure the statistical power. There were 36 data sets in this study (18 participants and two task conditions); therefore, the number of participants was sufficient to conduct the experiment.

5.3.6 Results and Discussion

To answer Experiment 3B's research questions, a paired *t*-test was conducted, the results of which showed significant improvements in some indices. In Table 5.5, the means, standard deviations, and 95% confidence intervals are shown, and Table 5.6 shows the result of the paired *t*-test.

The result of the current experiment revealed that there are significant differences in score of scale in communicative efficiency, types, and SPS. This result means that the test takers' fluency was improved when they have read aloud the input text. The indices of tokens, WPM, and errors per AS-unit are marginally significant. The indices of SPS and WPM are categorized as fluency index; however, the different result has appeared. The different feature of those indices causes the reason of this result. There is a tendency that the amounts of uttered words increased when test takers read aloud the input text. Despite the fact that the index of tokens is marginally significant, the index of types is significantly improved. This result means that test takers can produce more words under the reading aloud condition.

Table 5.5

The Means, Standard Deviations, and 95% CIs of Each Index of the Read Aloud and No Read Aloud Condition

	Read aloud condition		No read aloud condition	
	M (SD)	95% CI	*M (SD)*	95% CI
CE	1.94 (0.64)	[1.62, 2.26]	1.56 (0.70)	[1.21, 1.91]
GV	2.44 (0.92)	[1.99, 2.90]	2.28 (0.67)	[1.95, 2.61]
Pronunciation	2.83 (0.62)	[2.53, 3.14]	2.56 (0.78)	[2.17, 2.95]
Tokens	69.39 (24.33)	[57.29, 81.49]	59.72 (25.61)	[46.99, 72.46]
Types	39.61 (12.12)	[35.59, 45.64]	35.28 (11.68)	[29.47, 41.08]
WPM	27.76 (9.73)	[22.92, 32.60]	23.89 (10.24)	[18.80, 28.98]
SPS	0.67 (0.22)	[0.56, 0.78]	0.55 (0.22)	[0.44, 0.66]
TTR	0.59 (0.10)	[0.54, 0.64]	0.62 (0.12)	[0.56, 0.68]
GI	4.74 (0.71)	[4.39, 5.10]	4.57 (0.82)	[4.17, 5.00]
Wds/AS	7.79 (2.07)	[6.76, 8.82]	6.72 (1.30)	[6.08, 7.37]
Ers/AS	2.43 (1.37)	[1.75, 3.11]	2.13 (1.64)	[1.31, 1.94]

Note. CE = communicative efficiency, GV = grammar and vocabulary, WPM = words per minute, SPS = syllables per second, TTR = type-token ratio, GI = Guiraud index, Wds/AS = words per AS-unit, Ers/AS = errors per AS-unit, CI = confidence interval, and *N* = 18.

Table 5.6

The Result of Paired T-Test comparing of Two Different Task Conditions

	t	*p*	*d*	95% CI
CE	2.36	.03*	0.58	[0.04, 0.74]
GV	0.82	.42	0.21	[-0.26, 0.59]
Pronunciation	1.43	.17	0.39	[-0.13, 0.69]
Tokens	1.98	.06	0.39	[-0.61, 19.95]
Types	2.42	.03*	0.37	[0.55, 8.11]
WPM	1.98	.06	0.39	[-0.25, 7.98]
SPS	2.64	.02*	0.56	[0.02, 0.22]
TTR	-1.05	.31	0.28	[-0.09, 0.03]
Guiraud index	1.39	.18	0.22	[-0.09, 0.43]
Wds/AS	2.02	.06	0.62	[-0.05, 2.18]
Ers/AS	0.53	.61	0.20	[-0.92, 1.53]

Note. CE = communicative efficiency, GV = grammar and vocabulary, WPM = words per minute, SPS = syllables per second, TTR = type-token ratio, Wds/AS = words per AS-unit, Ers/AS = errors per AS-unit, and *N* = 18.

Results of the current experiment reveal that a read-aloud pre-task elicited higher quality performances in terms of fluency during a retelling task (RQ3B-2). The most important implication of this is that students can produce more words under the condition of reading aloud (RQ3B-1). The index of tokens is marginally significant, but the reading aloud task did contribute to more production. This result is logical because students had more time to practice before they conducted their speech performances; in addition, they may store information differently due to the read-aloud task.

The cognitive demand of retelling tasks is rather high for novice learners, although the practicality of these tasks, particularly the ease of material development and controlling material, is an advantage for implementing them as an assessment tool for learners' speaking skills. By changing the task condition, test developers or teachers can elicit more utterances as well as more sophisticated utterances for evaluation. In this experiment, the effects of a read-aloud pre-task prior to the performance of a retelling task were established; therefore, if learners are having difficulty in producing sufficient speech, teachers could add this type of pre-task as an aid. The natural or more common task condition of retelling tasks is not to include a read-aloud condition; hence, if learners can produce sufficient speech to assess and provide feedback, then the read aloud task should be excluded. When students can perform well in the no read aloud condition, teachers can provide the input aurally.

Counting errors within participants' pronunciation was too severe of a criterion for the present experiment, so the author decided instead to compare pronunciation using the results of scores given by the human rater. In this experiment, Hirai and Koizumi's (2013) EBB2 scale was used and the differences in scores were also compared. The results of the *t*-test revealed no significant differences between the two task conditions; therefore, it can be said that pronunciation performance did not change when the read-aloud pre-task was assigned to. In fact, the author considered that pronunciation would most likely improve if the read-aloud pre-task was given because this task can be good practice for pronunciation; however, the results contradicted this expectation. In future research, such expectations should be carefully considered, as they can serve as limitations of similar types of studies.

5.3.7 Summary of Experiment 3B

Experiment 3B observed the effects of reading aloud as a pre-task for retelling tasks, and results revealed that the read-aloud pre-task reading activated fluency and tended to elicit more words. The results of Experiment 3B are similar to Experiment 3A; therefore, it can be said that pre-task treatment affects fluency as well as the length and/or number of utterances.

5.4 Conclusion of Study 3

This study demonstrated that pre-task preparation and reading aloud is effective in eliciting more fluent oral performances during retelling tasks. This result can encourage the implementation of practical speaking tests; in addition, engaging in read-aloud pre-tasks could provide positive effects for students because when more speech is elicited, teachers can provide more detailed feedback. Thus, when teachers use retelling tasks for lower-proficiency students, read-aloud re-task should be added to elicit speech.

Chapter 6

Study 4: The Difficulty of the Text Topics and Task Conditions in the Retelling Tasks

6.1 Introduction to Study 4

The purpose of study 4 is to synthesize the results of the previous studies conducted in this series and compare the difficulty of each task condition used, as well as to observing the difficulties related to the text topics as stimuli for the retelling tasks. In this study, the term *task condition* signifies the input mode × text length. To compress the results of the first three studies into one scale, many-faceted Rasch measurement (MFRM) was conducted, and the difficulty of each task condition was compared with the others. However, in the presentation of the results here, the task condition of summarizing is excluded, as it was concluded in study 1 that oral retelling and summarizing are close to being the same in terms of performance results. Therefore, summarizing performances were dropped, and reading tasks of 100, 150, 300, or 500 words were substituted for them.

In study 1, the differences in spoken performance for the retelling and summarizing tasks were compared, and the results revealed no significant difference between those task types; subsequently, consideration of the effects of task instruction in retelling and summarizing, as well as the factor of text length, reflected no significant interaction between task instruction and the length of input texts. However, the results of the analysis of the main effect indicated that some indices of complexity, accuracy, and fluency (CAF) showed significant differences on the length factor, while the task instruction factor showed that the index of reproduction frequency reflected significant differences between task types, and the participants reproduced more words in the retelling tasks.

In study 2, the effects of text length and difficulty and input mode on the retelling tasks in spoken performances were investigated. The results of experiment 2A showed that text difficulty, based on the Flesch-Kincaid Grade Level, did not affect performance; in contrast, spoken performance did improve when participants were given longer input material. The effects of the input mode on performance were also observed, despite some problems with the data collection design, and the participants performed well when they obtained information visually: Reading input elicited longer performances than listening input did. To address this limitation of experiment 2A, the task design was adjusted, and experiment 2B was conducted. This experiment revealed that participants could produce more words when they received information under the reading condition than under the listening condition, but their fluency was better under the listening condition. In contrast, the indices of accuracy, difficulty, and rating score were not affected by the task condition. To summarize study 2, the task conditions considered in this study affected the amount (length) of utterances and their fluency, but accuracy and complexity were not altered. Similar results were found in study 3, which aimed to determine the effects of pre-task planning and pre-task reading aloud on performance in retelling. It was found that the participants tended to produce more words if additional treatment was assigned. In addition, neither the accuracy nor difficulty of the spoken performance changed in either condition.

The studies did not use the same criteria or raters; therefore, it was difficult to synthesize the results. Hence, new raters were hired, and a modified version of Hirai and Koizumi's (2013) empirically derived binary-choice boundary (EBB) definition scale was employed to evaluate participants' speaking abilities. The facets treated in this study were participants, rater, task (task × length of stimuli), topic of stimuli (hereafter, topic), and criteria; in sum, there were five facets in this analysis, and a five-faceted Rasch analysis was initially conducted. Subsequently, two 3-faceted Rasch measurements (participants × rater × task: participants × rater × topic) were conducted to compare the participants' English speaking ability in each design. The research questions developed for the study were as follows:

RQ 4-1: How are topic of texts distributed by difficulty?

RQ 4-2: How are task types distributed by difficulty?

6.2 Methods

6.2.1 Participants

The participants in this study were the aggregated participants from studies 1, 2, and 3. In all, data on 128 participants' performances were analyzed. All the participants were undergraduate or graduate students, and they had all completed at least 6 years of English education in junior and senior high school. The age of the participants ranged from 18 to 30 years, and they were students in national or public universities in Japan. Their levels of English-speaking ability were heterogeneous, as the experiments in studies 1 to 3 were conducted in different student groups, and their experiences in other countries were diverse. This could be an important limitation, but MFRM can calibrate the ability or difficulty of a task or other facets without population influences.

6.2.2 Scoring and Raters

The scoring instruments used in study 1 and experiment 2A were Hirai and Koizumi's (2008) original EBB definition scales, and in experiments 2B, 3A, and 3B, Hirai and Koizumi's (2013) revised version was employed. The original EBB had four subscales, as follows: communicative efficiency (fluency, coherency, elaboration, and sufficiency of the story and opinion), grammar and vocabulary (accuracy of grammar and vocabulary use), content, and pronunciation (accuracy of pronunciation, stress, intonation, and rhythm; Hirai & Koizumi, 2013, p. 401); the revised version dropped the content subscale, as "the correlation between [the] Communicative Efficiency and Content criteria was high at over .80, suggesting that the two shared a large proportion of variance" (Hirai & Koizumi, 2013, p. 401). Reflecting this change, the new version of the EBB scale was used in experiments 2B, 3A, and 3B; however, as the purpose of study 1 was comparing participants' performance between the retelling and summarizing tasks, a content criterion was important.

To synthesize the experiments, evaluations were carried out using the revised criteria. The revised scale simply adds a "0" option to the original version. Test takers may say something using nearly appropriate grammar, vocabulary, and pronunciation that nevertheless does not satisfy the task demands; in that case, raters give the performance a “1.” In contrast, participants may pronounce some words appropriately, but not successfully produce sentences or complete phrases; in such a case, the rater can assess pronunciation, but should give a score of “1” for communicative efficiency, grammar and vocabulary, and content. The validity of these criteria is considered in the analysis below. In addition to these scales, a descriptor of ACTFL-OPI was adapted to create a holistic rating scale (Appendix Q). The author translated the ACTFL-OPI speaking descriptor listed in the guidelines into Japanese, and revised it to assess spoken performance in retelling more easily by retaining the descriptions of language performance and deleting the descriptions that were not assessable from retelling tasks.

The author acted as chief rater; three other raters were trained using practice data and the criteria. One of these three raters had a master’s degree in English education and two years of English teaching experience in junior high school; another was a graduate student majoring in English education; and the third was an undergraduate student majoring in English education. All three had high English proficiency, with scores of at least 800 on the TOEIC Listening and Reading Test. All the raters were allowed to listen to the data repeatedly and change their marks after listening to other speech data, refer to the criteria while rating the performances, and listen to model performances. The chief rater shared the typical performance of each score range.

6.2.3 Analysis

To compare the difficulty across tasks, MFRM analyses were conducted using FACETS 3.71.4 (Linacre, 2014). A five-faceted Rasch measurement was conducted first, to cover all the facets in the analysis: participants, rater, task, topic, and criteria. Subsequently, five- and three-faceted Rasch measurements were conducted to observe the difficulty of the task and topic facets. The thresholds of the fit statistics were set at 0.7 to 1.3 (Bond & Fox, 2007, p. 243).

6.3 Validation of two types of EBB Scale

6.3.1 Validation of the Six-level EBB Scale

First, the validity of the new criteria, which were added to “0” in the original EBB scale (Hirai & Koizumi, 2008), was checked. The six-level EBB scale consists of four subcategories, as follows: communicative efficiency, grammar and vocabulary, contents, and pronunciation; in addition to these criteria, a holistic rating based on the ACTFL OPI guidelines was used. Therefore, five items were employed in this study. To validate these scales, a five-faceted Rasch measurement was conducted, and the facets of this analysis consisted of participants, rater, tasks, topics, and items.

The aim of this analysis is validating the new EBB scale. The five-faceted Rasch measurement revealed that some items of the new criteria do not work well. Focusing on the validity of the new criteria, the outputs of the category statistics of the five-faceted Rasch measurement were first checked.

Table 6.1 shows the category statistics of communicative efficiency, and the outfit mean square (MNSQ) values are fit to the Rasch model. However, the margin of the Rasch–Andrich threshold between scores of “1” and “2” is smaller than 1.4; therefore, this item should be changed. Bond and Fox (2007) claimed that the margin of the threshold should be at least 1.4, while it should reach a maximum of 5.0; therefore, this item should certainly be changed, or the scores of “0” and “1” should be summed. Figures 6.2 and 6.3 show that the frequencies in the use of scores of “0,” “4,” and “5” are lower than those of the other scores; in other words, the raters of this study tended to mark the participants’ spoken performances at “1,” “2,” or “3,” and it can be said that the scores tended to be centered. The expected score ogives revealed that the threshold between scores of “2” and “3” was located at nearly theta = 0; therefore, the participants of this study who performed well were marked higher than “4,” and low-proficiency learners were marked at “0” to “3.” This result can be interpreted positively; however, a score of “1” measured a narrow band of ability.

Table 6.1

Category statistics: Communicative efficiency

Data					Quality Control			Rasch–Andrich Threshold		Expectation				
Score	Category Total	Counts Used	%	Cum.%	Avg. Meas.	Exp. Meas.	Outfit MNSQ	Measure	S.E.	Measure Category	at – 0.5	Most Probable from	Rasch–Thurstone Threshold	Cat. Peak Prob.
0	116	108	7%	7%	–2.66	–2.51	0.9			(–4.30)		Low	Low	
1	257	257	16%	23%	–1.77	-1.73	0.9	–2.98	0.11	–2.68	–3.57	–2.98	–3.28	41%
2	596	596	38%	61%	–1.00	-0.99	0.9	–2.20	0.07	–1.19	-1.96	–2.20	–2.05	53%
3	494	494	31%	92%	–0.16	-0.16	0.9	–0.40	0.06	0.77	-0.28	–0.40	–0.34	61%
4	100	100	6%	99%	1.30	0.97	0.7	1.96	0.11	2.83	1.83	1.96	1.88	53%
5	19	19	1%	100%	3.17	3.11	0.9	3.62	0.29	–4.82	3.96	3.62	3.76	100%

Note. Avg. Meas. = Average of measurement; Exp. Meas. = Expected value of the average measures, Cat. Peak Prob = Category peak probability.

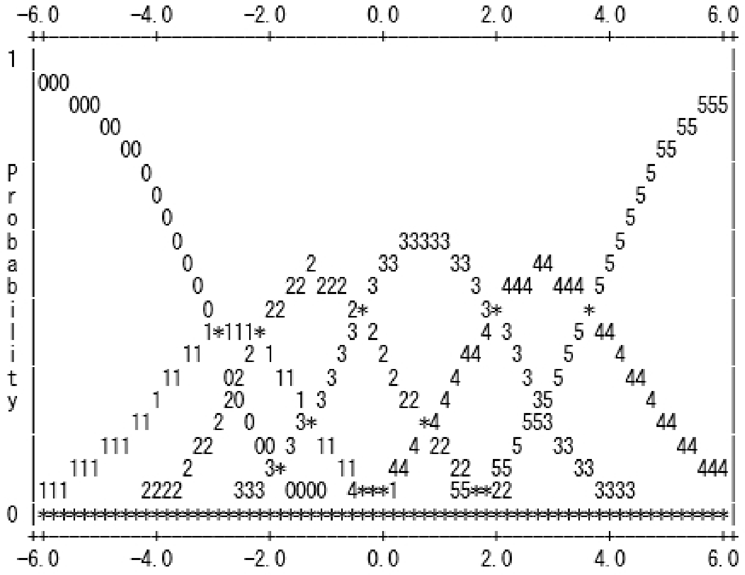


Figure 6.1. Probability curves for communicative efficiency.

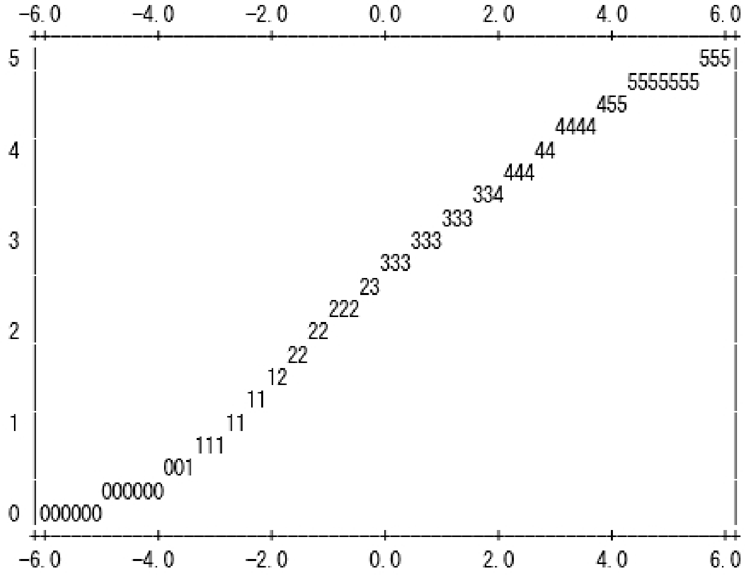


Figure 6.2. Expected score ogives (model ICC): Communicative efficiency.

Next, the item of grammar and vocabulary was checked. Figures 6.3 and 6.4 show the probability curves and expected score ogives for grammar and vocabulary. Figure 6.3 clearly shows that a score of “1” does not work, as there is no obvious peak for this score. In addition, the range of the score of “3” is wider than the ranges of the other scores; therefore, it can be said that there was also a central tendency on the grammar and vocabulary item. In addition, the MNSQ shown in Table 6.2 indicates the score of "1" in the grammar and vocabulary item is misfit. Hence, this item should also be revised.

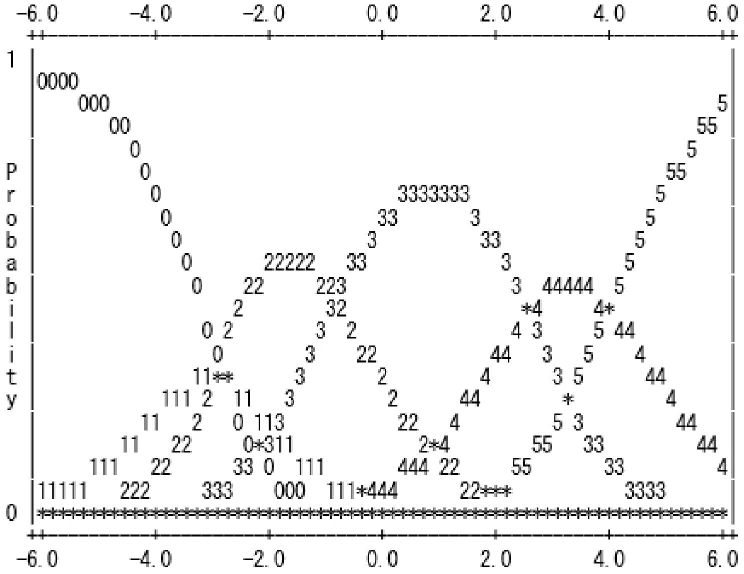


Figure 6.3. Probability curves for grammar and vocabulary.

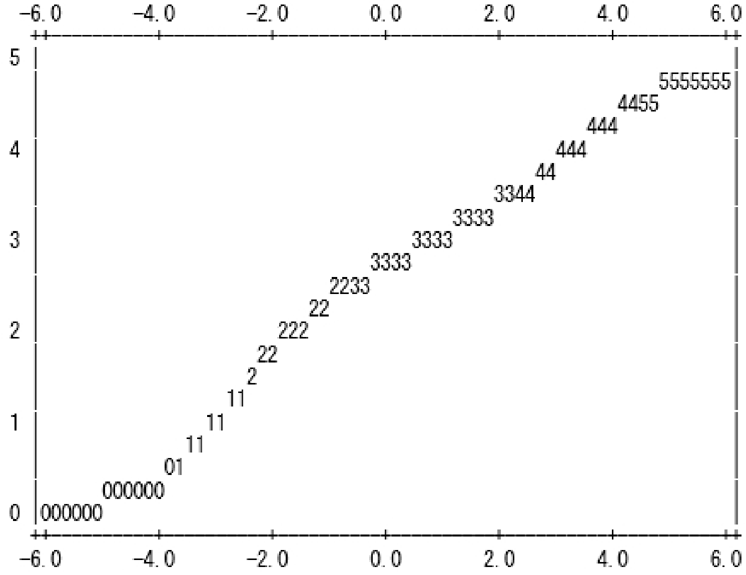


Figure 6.4. Expected score ogives (model ICC): Grammar and vocabulary.

Table 6.2

Category statistics: Grammar and vocabulary

Data					Quality Control			Rasch–Andrich Threshold		Expectation				
Score	Category Total	Counts Used	%	Cum.%	Avg. Meas.	Exp. Meas.	Outfit MNSQ	Measure	S.E.	Measure at Category	at -0.5	Most Probable from	Rasch–Thurstone Thresholds	Cat. Peak Prob.
0	116	108	7%	7%	–2.84	–2.76	1.0			(–4.32)		Low	Low	100%
1	156	156	10%	17%	–1.74	–2.00	1.4	–2.75	0.12	–2.98	-3.71		–3.33	29%
2	623	623	40%	56%	–1.22	–1.25	1.1	–3.01	0.08	–1.66	-2.38	–2.88	–2.57	56%
3	614	614	39%	95%	–0.48	–0.36	1.1	–0.81	0.06	0.85	–0.62	–0.81	–0.73	72%
4	62	62	4%	99%	1.47	1.06	0.8	2.56	0.14	3.30	2.27	2.56	2.41	51%
5	11	11	1%	100%	3.43	3.44	0.9	4.01	0.38	(–5.24)	4.40	4.01	4.18	100%

Note. Avg. Meas. = Average of measurement; Exp. Meas. = Expected value of the average measures, Cat. Peak Prob = Category peak probability.

The contents item was then observed, and this was found to be the worst item in this study. Figure 6.5 shows that a score of “1” did not work; moreover, a score of “4” did not work well either. There were two steep slopes in the expected score ogives, as shown in Figure 6.6. In addition, Table 6.3 shows that the MNSQ value for each score did not match Bond and Fox’s (2007) criteria; therefore, it is necessary to change this item.

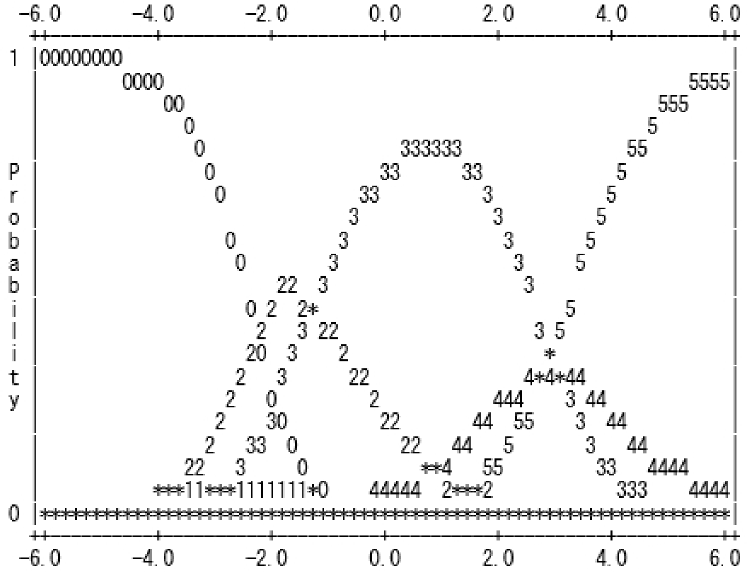


Figure 6.5. Probability curves for contents.

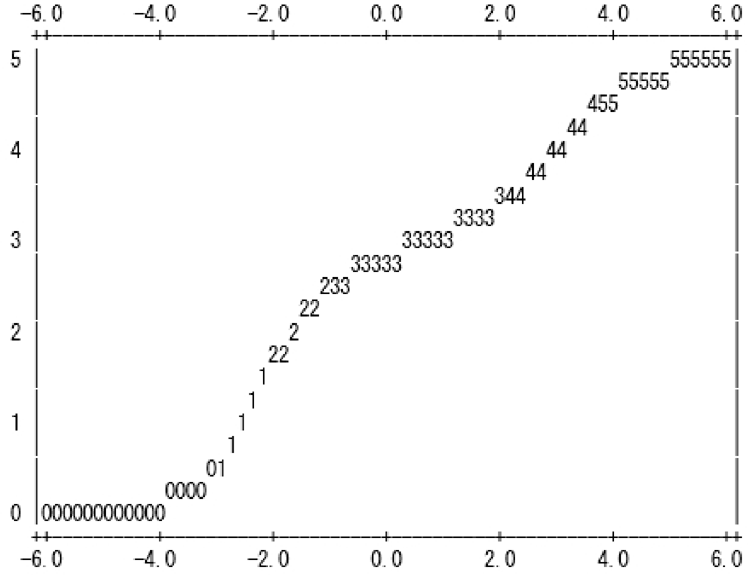


Figure 6.6. Expected score ogives (model ICC): Contents.

Table 6.3

Category statistics: Contents

Data					Quality Control			Rasch–Andrich Threshold		Expectation				
Score	Category Total	Counts Used	%	Cum.%	Avg. Meas.	Exp. Meas.	Outfit MNSQ	Measure	S.E.	Measure at Category	at -0.5	Most Probable from	Rasch–Thurstone Threshold	Cat. Peak Prob.
0	112	104	7%	7%	–2.19	–2.28	1.4			(–3.42)		Low	Low	100%
1	23	23	1%	8%	–1.25	–1.60	1.5	–0.44	0.13	–2.52	–3.00		–2.49	6%
2	411	411	26%	34%	–0.55	–0.88	1.4	–4.13	0.12	–1.68	–2.12	–2.29	–2.36	49%
3	926	926	59%	93%	–0.20	–0.01	1.4	–1.28	0.06	0.76	–0.94	–1.28	–1.15	80%
4	75	75	5%	98%	1.27	1.15	0.9	3.05	0.12	2.94	2.18		2.50	30%
5	35	35	2%	100%	3.52	3.12	0.4	2.80	0.23	–4.37	3.74	2.92	3.35	100%

Note. Avg. Meas. = Average of measurement; Exp. Meas. = Expected value of the average measures, Cat. Peak Prob = Category peak probability.

The pronunciation item was antithetically drawn to the ideal probability curves (Figures 6.7 and 6.8). Table 6.4 shows that there were no misfits, and the used scores were distributed ideally.

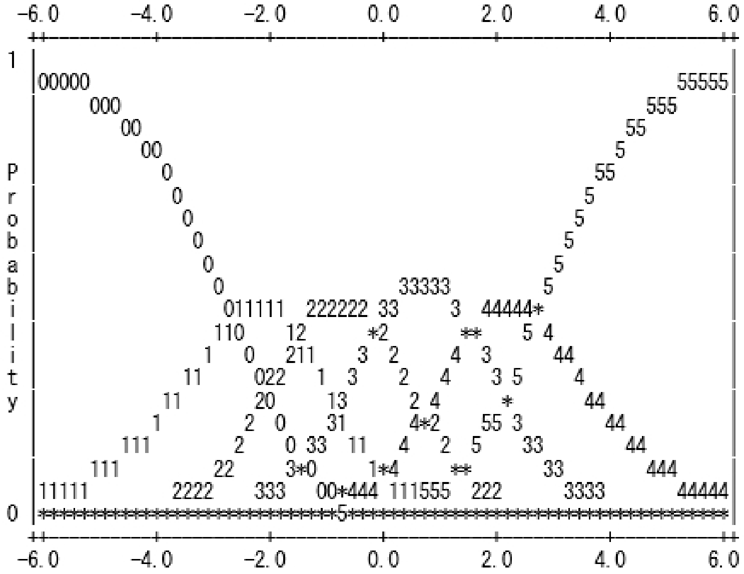


Figure 6.7. Probability curves for pronunciation.

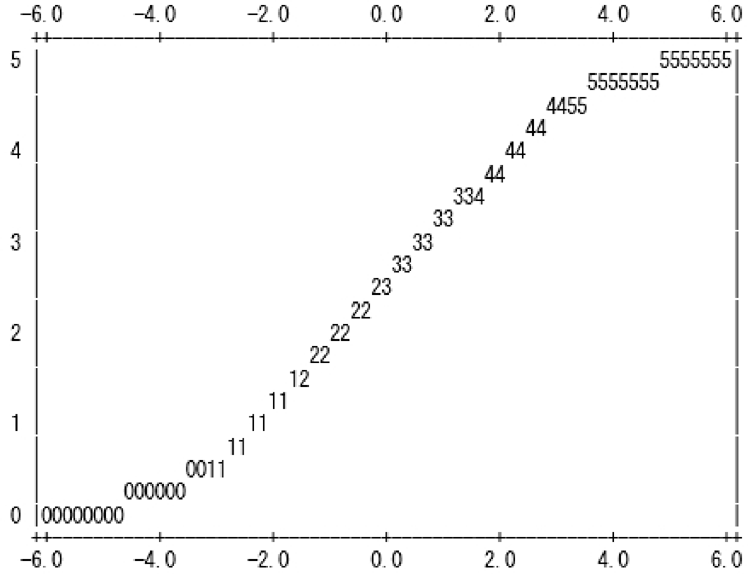


Figure 6.8. Expected score ogives (model ICC): Pronunciation.

Table 6.4

Category statistics: Pronunciation

Data					Quality Control			Rasch–Andrich Threshold		Expectation				
Score	Category Total	Counts Used	%	Cum.%	Avg. Meas.	Exp. Meas.	Outfit MNSQ	Measure	S.E.	Measure at Category	at –0.5	Most Probable from	Rasch–Thurstone Threshold	Cat. Peak Prob.
0	117	109	7%	7%	–2.23	–2.06	0.9			(–3.90)		Low	Low	100%
1	286	286	18%	25%	–1.37	–1.29	0.7	–2.63	0.11	–2.19	–3.12	–2.63	–2.87	45%
2	513	513	33%	58	–0.50	–0.6	1.1	–1.53	0.07	–0.77	–1.47	–1.53	–1.48	47%
3	448	448	28%	86%	0.14	0.13	1.1	–0.11	0.06	0.71	–0.05	–0.11	–0.07	51%
4	162	162	10%	96%	1.02	0.99	1.1	1.56	0.09	2.23	1.46	1.56	1.48	46%
5	56	56	4%	100%	2.14	2.52	1.3	2.7	0.17	(3.98)	3.19	2.70	2.93	100%

Note. Avg. Meas. = Average of measurement; Exp. Meas. = Expected value of the average measures, Cat. Peak Prob = Category peak probability.

Table 6.5

Category statistics: Holistic scale

Data					Quality Control			Rasch–Andrich Thresholds		Expectation				
Score	Category Total	Counts Used	%	Cum.%	Avg. Meas.	Exp. Meas.	Outfit MNSQ	Measure	S.E.	Measure at Category	at – 0.5	Most Probable from	Rasch–Thurstone Thresholds	Cat. Peak Prob.
0	120	112	7%	7%	–2.39	–2.25	0.9			(–3.96)		Low	Low	100%
1	230	230	15%	22%	–1.51	–1.49	0.8	–2.58	0.11	–2.42	–3.25	–2.58	–2.95	38%
2	583	583	37%	59%	–0.78	–0.77	0.9	–2.06	0.07	–1.01	–1.74	–2.06	–1.85	53%
3	496	496	32%	90%	0.08	0.02	0.9	–0.22	0.06	0.82	–0.14	–0.22	–0.18	57%
4	116	116	7%	98%	1.06	1	0.9	1.94	0.10	2.5	1.7	1.94	1.78	44%
5	37	37	2%	100%	2.69	2.78	1.1	2.92	0.21	(4.21)	3.45	2.92	3.17	100%

Note. Avg. Meas. = Average of measurement; Exp. Meas. = Expected value of the average measures, Cat. Peak Prob = Category peak probability.

Finally, the holistic scale item was observed, and it showed a tendency to avoid using a score of “1” (Figures 6.9 and 6.10). Table 6.5 indicates a small distance between scores of “1” and “2,” meaning that the scores did not work well.

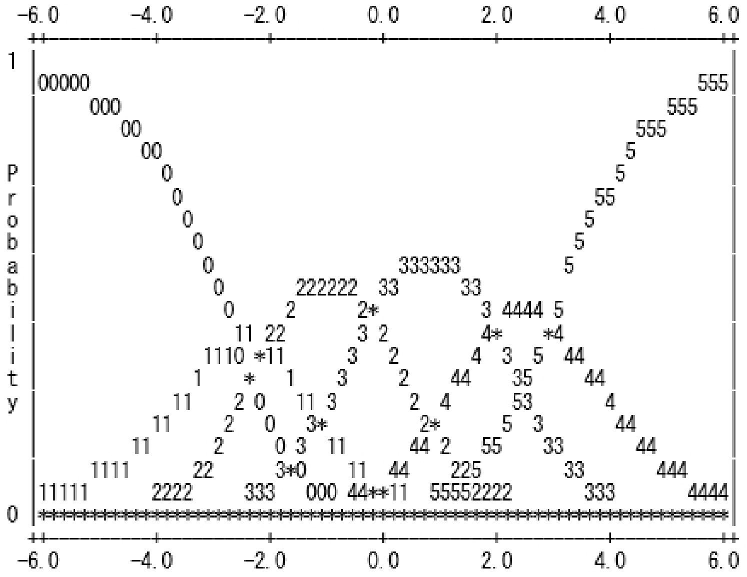


Figure 6.9. Probability curves for the holistic scale.

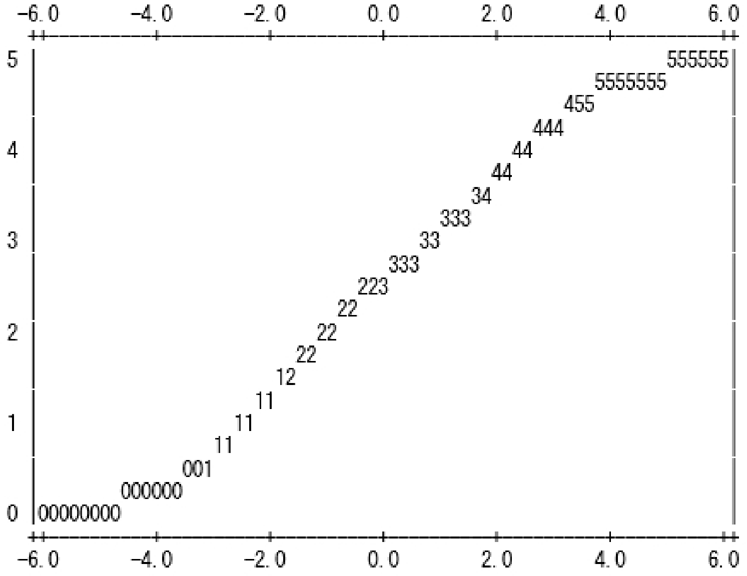


Figure 6.10. Expected score ogives (ICC model): Holistic rating.

The results of the analyses indicated that the new criteria used in this study are not sufficient;

hence, the score of “1” in each item tended not to work and the score of “0” was changed to “1.” In other words, the score range was changed from 0–5 to 1–5.

6.3.2 Validation of the Five-level EBB Scale

The same analysis was conducted as described above, and this revealed that the newest criteria – that is, the criteria score ranging from “1” to “5” (hereafter, the five-level scale) – worked better than the criteria analyzed in the previous part of this section. In fact, the contents item still included a misfit score, but the new score was relatively improved. In the latter parts of this section, the category statistics in each item are observed in advance of the other MFRM analysis.

Table 6.6, Figure 6.11, and Figure 6.12 show the category statistics, probability curves, and expected score ogives of the item of communicative efficiency. The results of the category statistics show that the fit value matches the criterion by Bond and Fox (2007). In addition, the margin of each scale is wider than 1.4 and smaller than 5.0. The probability curves shown in Figure 6.11 have peaks in each score; therefore, it can be said that this item matches the Rasch model, and it functions well.

Table 6.6

Category statistics: Communicative efficiency (2)

Data					Quality Control			Rasch–Andrich Threshold		Expectation				
										Measure		Most	Rasch–	Cat.
	Category	Counts			Avg.	Exp.	Outfit			at	at –	Probable	Thurstone	Peak
Score	Total	Used	%	Cum.%	Meas.	Meas.	MNSQ	Measure	S.E.	Category	0.5	from	Threshold	Prob.
1	373	365	23%	23%	–3.01	–2.9	0.9			(–4.04)		Low	Low	100%
2	596	596	38%	61%	–1.86	–1.87	0.8	–2.85	0.07	–2.06	–3.18	–2.85	–3.00	52%
3	494	494	31%	92%	–0.94	–0.95	0.8	–1.22	0.06	–0.01	–1.08	–1.22	–1.15	61%
4	100	100	6%	99%	0.54	0.21	0.7	1.19	0.11	2.07	1.06	1.19	1.11	53%
5	19	19	1%	100%	2.41	2.39	1.0	2.88	0.29	(4.08)	3.21	2.88	3.02	100%

Note. Avg. Meas. = Average of measurement; Exp. Meas. = Expected value of the average measures, Cat. Peak Prob = Category peak probability.

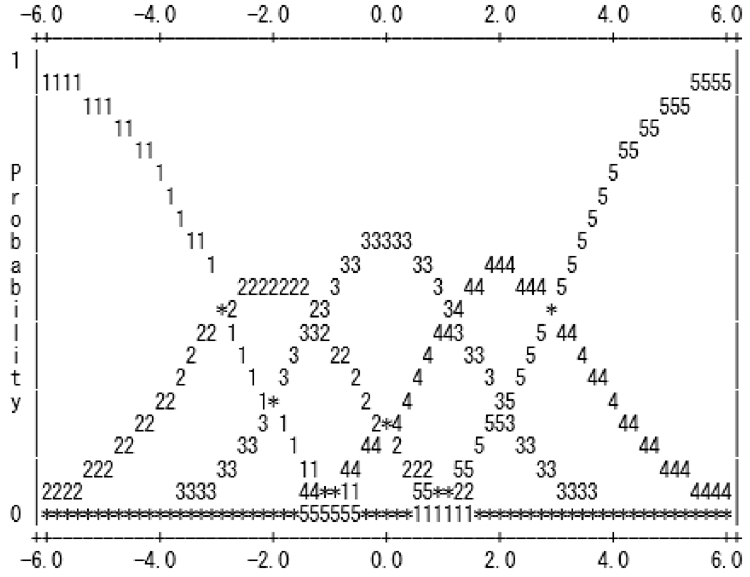


Figure 6.11. Probability curves for communicative efficiency (2).

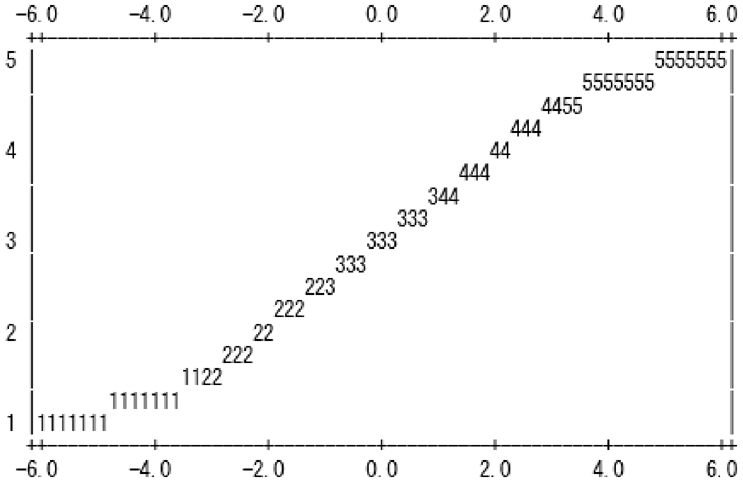


Figure 6.12. Expected score ogives (model ICC): Communicative efficiency (2).

Next, the grammar and vocabulary item was inspected. Table 6.7, Figure 6.13, and Figure 6.14 give the results of the category statistics for grammar and vocabulary, and the fit statistics show that the outfit MNSQ ranged from 0.7 to 1.3. In addition, the thresholds of each scale are separated, and the probability curves have peaks in each scale; therefore, it can be said that this item worked well.

Table 6.7

Category statistics: Grammar and vocabulary (2)

Data					Quality Control			Rasch–Andrich Threshold		Expectation				
Score	Category Total	Counts Used	%	Cum.%	Avg. Meas.	Exp. Meas.	Outfit MNSQ	Measure	S.E.	Measure at Category	at –0.5	Most Probable from	Rasch–Thurstone Thresholds	Cat. Peak Prob.
1	272	264	17%	17%	–3.14	–3.22	1.2			(–4.67)		Low	Low	100%
2	623	623	40%	56%	–2.08	–2.12	1.0	–3.51	0.08	–2.57	–3.78	–3.51	–3.63	56%
3	614	614	39%	95%	–1.24	–1.12	1.1	–1.62	0.06	0.08	–1.40	–1.62	–1.52	73%
4	62	62	4%	99%	0.72	0.34	0.8	1.82	0.14	2.58	1.53	1.82	1.67	51%
5	11	11	1%	100%	2.70	2.74	0.9	3.30	0.38	(4.52)	3.68	3.30	3.47	100%

Note. Avg. Meas. = Average of measurement; Exp. Meas. = Expected value of the average measures, Cat. Peak Prob = Category peak probability.

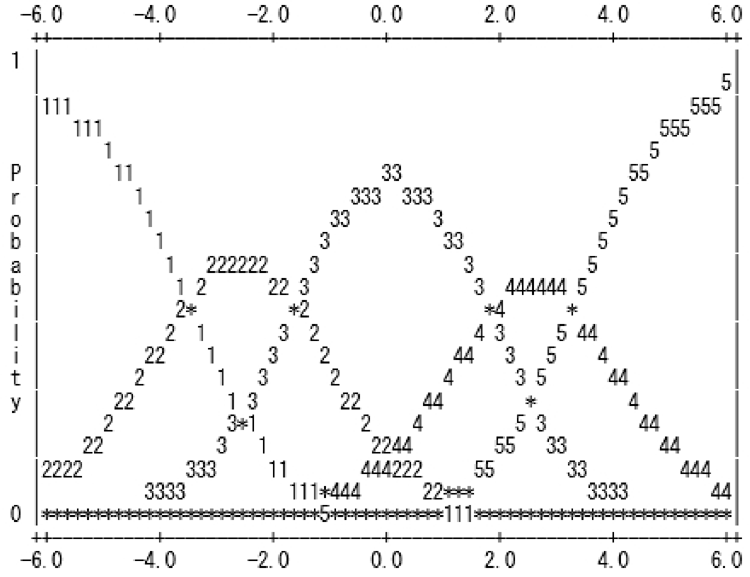


Figure 6.13. Probability curves for grammar and vocabulary (2).

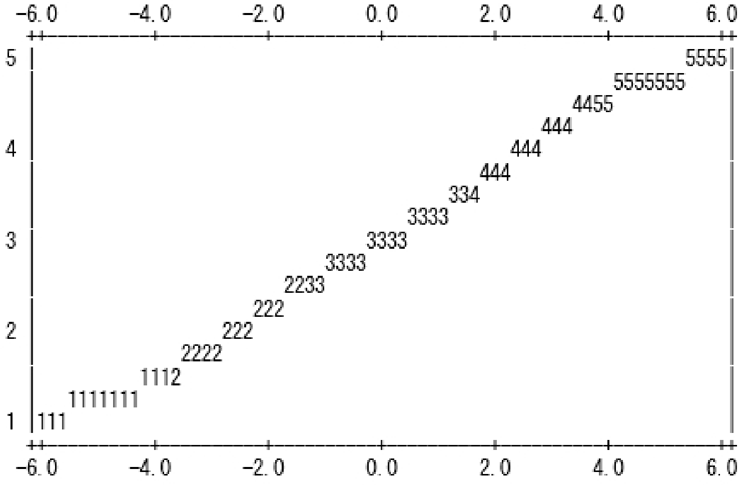


Figure 6.14. Expected score ogives (model ICC): grammar and vocabulary (2).

The contents item did not function in the former analysis, nor did it match the Rasch model. The scores of "2" and "3" misfit the model, while "5" is overfit to the Rasch model (Table 6.8). Figure 6.15 shows that the probability curves of a score of "4" do not have a peak, and the range of a score of "3" is extremely large in comparison with those of other scales. However, these results represent an improvement on the previous analysis.

Table 6.8

Category statistics: Contents (2)

	Data					Quality Control		Rasch–Andrich Threshold		Expectation				
Score	Category Total	Counts Used	%	Cum.%	Avg. Meas.	Exp. Meas.	Outfit MNSQ	Measure	S.E.	Measure at Category	at –0.5	Most Probable from	Rasch–Thurstone Thresholds	Cat. Peak Prob.
1	135	127	8%	8%	–2.59	–2.66	1.2			(–4.45)		Low	Low	100%
2	411	411	26%	34%	–1.09	-1.51	1.4	–3.24	0.11	–2.52	-3.62	–3.24	–3.42	51%
3	926	926	59%	93%	–0.69	-0.47	1.5	–1.8	0.06	0.31	–1.44	–1.80	–1.63	81%
4	75	75	5%	98%	0.89	0.75	0.9	2.63	0.12	2.54	1.77		2.09	31%
5	35	35	2%	100%	3.11	2.75	0.5	2.42	0.23	(3.98)	3.34	2.52	2.95	100%

Note. Avg. Meas. = Average of measurement; Exp. Meas. = Expected value of the average measures, Cat. Peak Prob = Category peak probability.

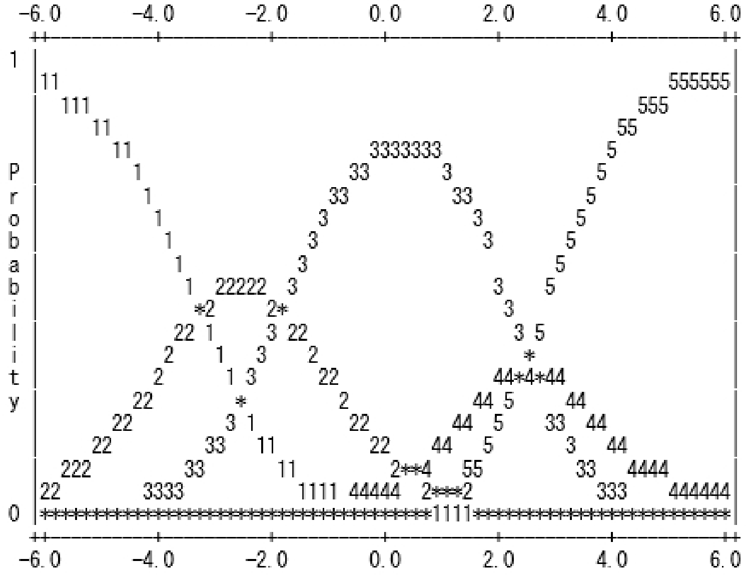


Figure 6.15. Probability curves for contents (2).

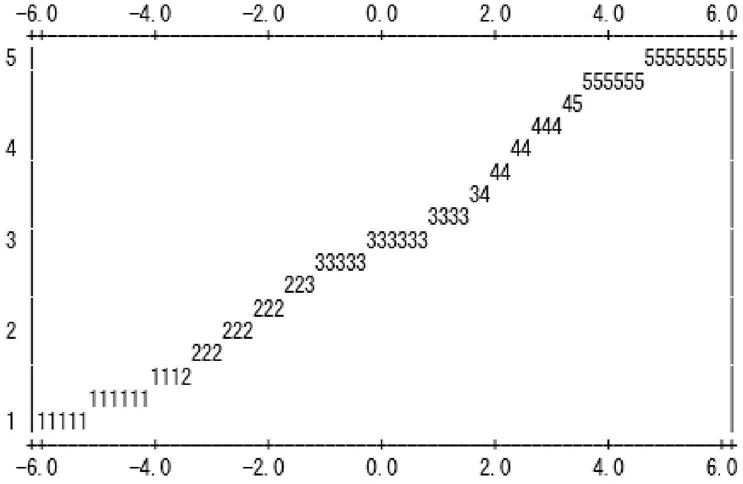


Figure 6.16. Expected score ogives (model ICC): contents.

The pronunciation item worked well when the six-level scale was used, and the five-level scale also fit the Rasch model. The probability curves shown in Figure 6.17 were ideally drawn.

Table 6.9

Category statistics: Pronunciation (2)

Data					Quality Control			Rasch–Andrich Threshold		Expectation				
Score	Category Total	Counts Used	%	Cum.%	Avg. Meas.	Exp. Meas.	Outfit MNSQ	Measure	S.E.	Measure at Category	at – 0.5	Most Probable from	Rasch–Thurstone Threshold	Cat. Peak Prob.
1	403	395	25%	25%	–2.48	–2.35	0.8			(–3.34)		Low	Low	100%
2	513	513	33%	58%	–1.25	–1.38	1.0	–2.09	0.07	–1.54	–2.54	–2.09	–2.31	47%
3	448	448	28%	86%	–0.56	–0.57	1.1	–0.84	0.06	0.02	–0.75	–0.84	–0.78	51%
4	162	162	10%	96%	0.36	0.33	1.0	0.88	0.09	1.56	0.78	0.88	0.80	46%
5	56	56	4%	100%	1.52	1.88	1.2	2.05	0.17	(3.32)	2.53	2.05	2.27	100%

Note. Avg. Meas. = Average of measurement; Exp. Meas. = Expected value of the average measures, Cat. Peak Prob = Category peak probability.

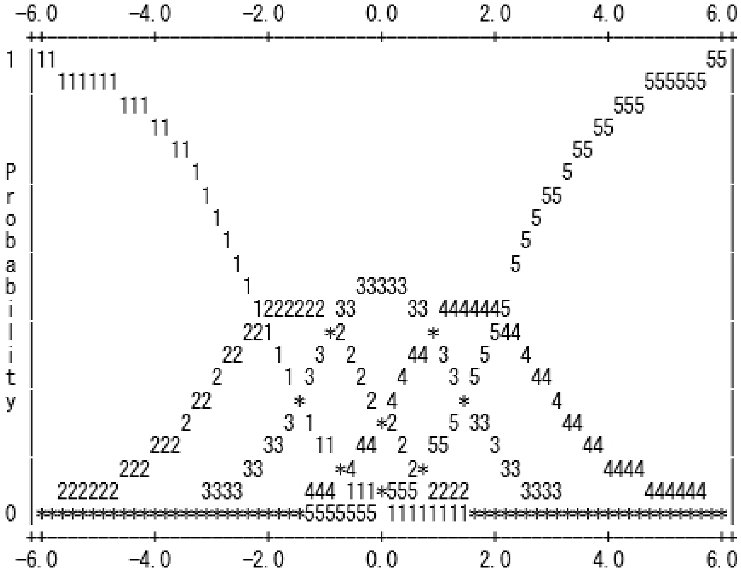


Figure 6.17. Probability curves for pronunciation (2).

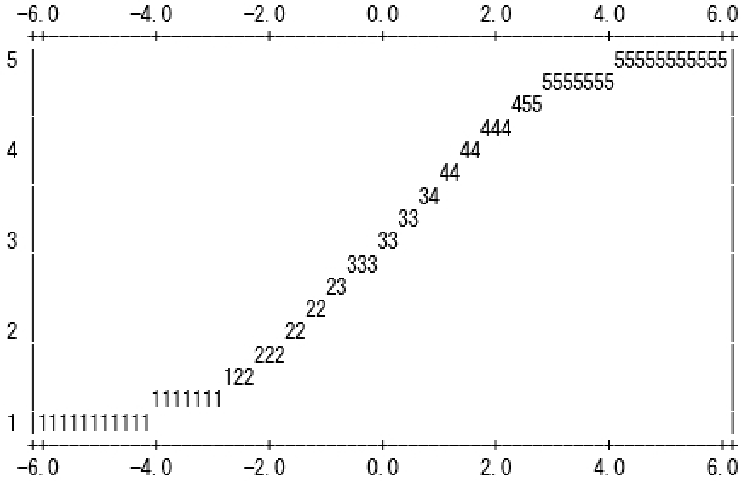


Figure 6.18. Expected score ogives (model ICC): pronunciation (2).

Finally, the holistic rating results are inspected. This item matched the Rasch model (see Table 6.10). The margin of the Rasch-Andrich thresholds between the scores of “4” and “5” was only 1.01; therefore, the criteria must still be changed. However, there are clear peaks for each score; thus, the five-level scale is more appropriate.

Table 6.10

Category statistics: Holistic rating (2)

Data					Quality Control			Rasch–Andrich Threshold		Expectation				
Score	Category Total	Counts Used	%	Cum.%	Avg. Meas.	Exp. Meas.	Outfit MNSQ	Measure	S.E.	Measure at Category	at – 0.5	Most Probable from	Rasch–Thurstone Threshold	Cat. Peak Prob.
1	350	342	22%	22%	–2.7	–2.58	0.9			(–3.77)		Low	Low	100%
2	583	583	37%	59%	–1.55	-1.56	0.9	–2.58	0.07	–1.80	–2.92	–2.58	–2.73	52%
3	496	496	32%	90%	–0.61	–0.67	0.8	–0.95	0.06	0.12	-0.85	–0.95	–0.89	58%
4	116	116	7%	98%	0.41	0.34	0.9	1.26	0.1	1.83	1.01	1.26	1.10	44%
5	37	37	2%	100%	2.02	2.14	1.1	2.27	0.21	(3.56)	2.79	2.27	2.51	100%

Note. Avg. Meas. = Average of measurement; Exp. Meas. = Expected value of the average measures, Cat. Peak Prob = Category peak probability.

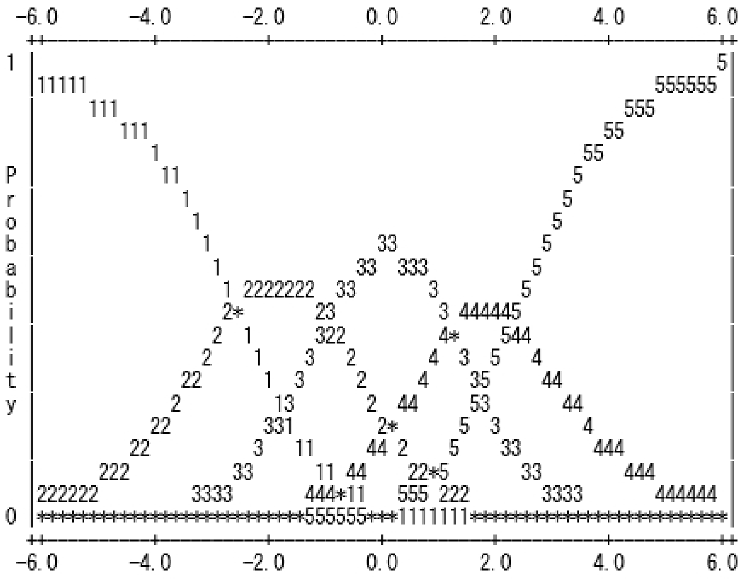


Figure 6.19. Probability curves for the holistic rating (2).

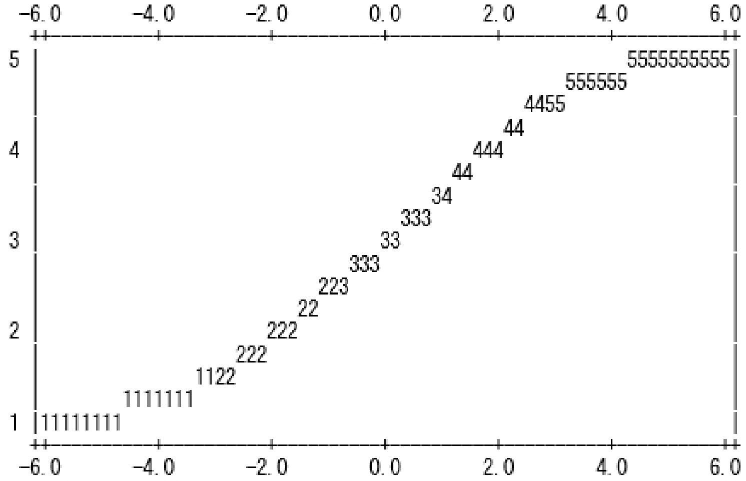


Figure 6.20. Expected score ogives (model ICC): holistic rating (2).

In total, the five-level scale is more effective for measuring the participants' ability. There is still some room to revise the criteria, but in this study, the five-level scale will be used. In the latter part of this chapter, the five- and three-faceted Rasch measurements are employed to assess the difficulty of the task and topic facets.

6.4 Three-Faceted Rasch Measurement: Participants × Raters × Text Topics

6.4.1 Analysis of the Data Fit

In the previous section of this chapter, whether a five- or six-level scale should be used in the analysis was considered. Below, a three-faceted Rasch measurement is conducted to investigate the difficulty of the text topics, while a second three-faceted Rasch measurement is operated to determine the difficulty of the task types. In other words, two 3-faceted Rasch measurements are performed. In the present section, the facets of participants × raters × text topics are treated. In this analysis, there are 128 participants, four raters, and seven text topics forming the subject of analysis. In the analyses below, the rating scale considered in the previous section is used, and the scale's total score is employed for the analysis; in other words, the scores range from 5 to 25.

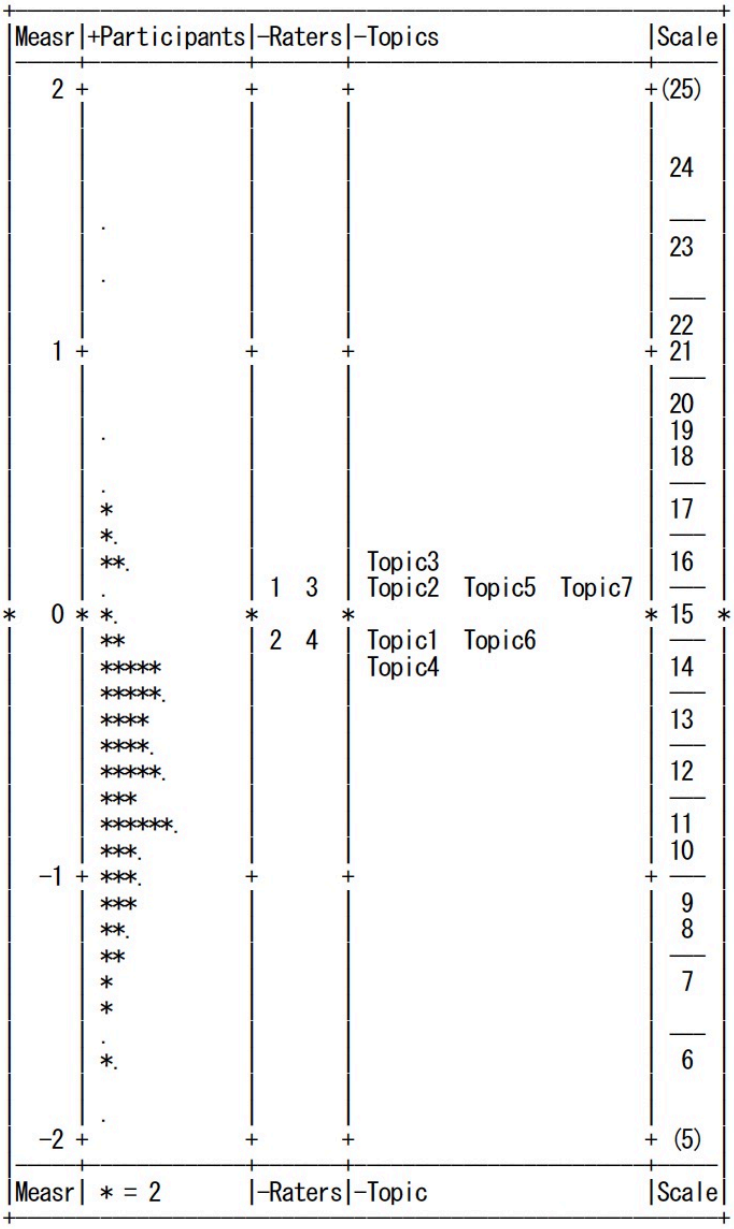
To conduct the three-faceted Rasch measurement, first, the fit statistics and unexpected responses must be assessed to determine whether the data fit the Rasch model. The results of the data log-likelihood chi-squared test were significant (data log-likelihood $\chi^2 = 7785.17$, $df = 1655$, $p < .00$); therefore, the data did not fit the Rasch model. However, Eckes (2015, p. 61) indicated that this method to observe data fit to Rasch model is too severe, and this result of data fit was not fatal. Another criterion of model fit is the ratio of unexpected responses in the full dataset. There were 80 unexpected responses in this dataset out of 1,811 responses (128 participants × four raters × seven text topics – missing data); therefore, the ratio of unexpected responses was about 4%. Furthermore, the fit statistics in the rater and topic facets were observed, and a misfit in the topic facet was identified. This will be interpreted in the latter part of this section, but it can be said here that the whole dataset is feasible for analysis.

Next, the unidimensionality assumption of the data should be confirmed. Here, the facets were used to calculate the variance explained by the Rasch measures, and the score was 64.64%. Engelhard (2013) claimed that if the score of variance explained by the Rasch measures exceeds 20%, the data satisfy the unidimensionality assumption; therefore, the present data fulfill the requirement.

6.4.2 Results of the Three-faceted Rasch Measurement: Participants × Raters × Text Topics

Figure 6.21 is a Wright map of the three-faceted Rasch measurements. This figure shows the calibration of the participants' proficiency, raters' severity, and difficulty of the topic facet.

The facet of participants, shown in the leftmost line in Figure 6.21, illustrates the logits (log odds) of participant ability. This result indicates that the participants had low ability concerning speaking skill performance with the text topics used in this experiment.



```
+----------------------------------------------------------+
|Measr|+Participants|-Raters|-Topics                 |Scale|
|-----+-------------+-------+------------------------+-----|
|   2 +             +       +                        +(25) |
|     |             |       |                        |     |
|     |             |       |                        |     |
|     |             |       |                        | 24  |
|     |             |       |                        |     |
|     | .           |       |                        | --- |
|     |             |       |                        | 23  |
|     | .           |       |                        |     |
|     |             |       |                        | --- |
|     |             |       |                        | 22  |
|   1 +             +       +                        + 21  |
|     |             |       |                        | --- |
|     |             |       |                        | 20  |
|     | .           |       |                        | 19  |
|     |             |       |                        | 18  |
|     | .           |       |                        | --- |
|     | *           |       |                        | 17  |
|     | *.          |       |                        | --- |
|     | **.         |       | Topic3                 | 16  |
|     | .           | 1  3  | Topic2  Topic5  Topic7 | --- |
*   0 * *.          *       *                        * 15  *
|     | **          | 2  4  | Topic1  Topic6         | --- |
|     | *****       |       | Topic4                 | 14  |
|     | *****.      |       |                        | --- |
|     | ****        |       |                        | 13  |
|     | ****.       |       |                        | --- |
|     | *****.      |       |                        | 12  |
|     | ***         |       |                        | --- |
|     | ******.     |       |                        | 11  |
|     | ***.        |       |                        | 10  |
|  -1 + ***.        +       +                        + --- |
|     | ***         |       |                        |  9  |
|     | **.         |       |                        |  8  |
|     | **          |       |                        | --- |
|     | *           |       |                        |  7  |
|     | *           |       |                        |     |
|     | .           |       |                        | --- |
|     | *.          |       |                        |  6  |
|     |             |       |                        |     |
|     | .           |       |                        |     |
|  -2 +             +       +                        + (5) |
|-----+-------------+-------+------------------------+-----|
|Measr| * = 2       |-Raters|-Topic                  |Scale|
+----------------------------------------------------------+
```

Figure 6.21. Wright map of the three-faceted Rasch measurement (participants × raters × topics).

Table 6.11 shows the summary of the measurement for the participants, and the full table of this analysis is given in Appendix R. The observed average was 11.72, and the average of the measure was –0.57; hence, the participants' ability was lower than the standard for this measurement. The fixed chi-squared test was significant; therefore, the study participants were not equal in their ability. The person separation reliability was .95; thus, their ranks were stable. The fit statistics shown in Table 6.11 indicate that the average of the participant facet matched the Rasch model. In fact, there were 18 misfits for the participant facet; however, this can be attributed to carelessness, miscoding of data, lucky guessing or special knowledge (Bond & Fox, 2007, p. 242). Such issues could have emerged because some participants did not complete the task; moreover, there were some unclear voice data, which may also cause misfits.

Next, the rater facet was observed. Table 6.12 reports the rater results. The infit and outfit MNSQs ranged from 0.7 to 1.3; therefore, this facet matched the model. The fixed chi-squared test was significant, and this means that the raters' severity was not equal. Ideally, the severity should be equal, but it is natural for ratings to disperse. However, the index of exact agreement was 30.8%, where the expected value was 14.3%; therefore, the rating quality was higher than the expected agreement.

The main part of this analysis examined the difficulty of the task topics for the retelling tasks. The topic difficulties are loosely divided into two groups in Figure 6.21. Table 6.13 shows the summary of the topic measurement; there was one topic that was misfit to the Rasch model. There were 21 unexpected responses in topic 4, and this means that the topic may have been difficult for the participants. Topic 4 was used in study 2B, and this was given in the first or last phase of the analysis; therefore, the practice effect may have influenced this result. The topic logits ranged from –0.15 to 0.16; hence, the effects of the topic on the performance seem to have been small.

Table 6.11

Summary of the Participant Measurement in Three-faceted Rasch Measurement: Participants × Raters × Topics

Total Score	Total Count	Observed Average	Fair (M) Average	Measure	Model S.E.	Infit MNSQ	Infit Zstd	Outfit MNSQ	Outfit Zstd	Estim. Discrim.	Corr. PtBis	Participants
166.7	14.1	11.72	11.89	–0.57	0.12	0.97	–0.3	0.97	–0.3		.07	Mean (Count: 128)
58.3	3.0	3.04	3.06	0.56	0.02	0.73	1.6	0.73	1.6		.30	S.D. (Population)
58.6	3.0	3.06	3.08	0.56	0.02	0.73	1.6	0.74	1.6		.30	S.D. (Sample)

Model, Populn: RMSE .12 Adj (True) S.D. .55 Separation 4.48 Strata 6.30 Reliability .95

Model, Sample: RMSE .12 Adj (True) S.D. .55 Separation 4.50 Strata 6.33 Reliability .95

Model, Fixed (all same) chi-square: 2710.7 d.f.: 127 significance (probability): .00

Model, Random (normal) chi-square: 120.1 d.f.: 126 significance (probability): .63

Note. Zstd = Infit z-standardized t-statistic, Estim. Discrim = estimated discrimination, Corr. PtBis = point-biserial correlation, Populn = population, RMSE = root mean square standard error, and Adj (True) S.D. = adjusted true S.D.

Table 6.12

Summary of the Rater Measurement in Three-faceted Rasch Measurement: Participants × Raters × Topics

Total Score	Total Count	Observed Average	Fair (M) Average	Measure	Model S.E.	Infit MNSQ	Infit Zstd	Outfit MNSQ	Outfit Zstd	Estim. Discrim.	Corr. PtBis	Exact Obs. %	Agree. Exp. %	Raters
5126.0	454.0	11.29	11.53	0.09	0.02	1.04	0.6	1.04	0.6	0.85	.54	35.7	14.4	3
5156.0	451.0	11.43	11.66	0.07	0.02	0.86	–2.2	0.86	–2.0	1.11	.58	26.5	14.5	1
5526.0	453.0	12.20	12.42	–0.08	0.02	1.04	0.6	1.06	0.8	0.87	.57	35.0	14.2	4
5527.0	453.0	12.20	12.44	–0.08	0.02	1.06	0.8	1.04	0.6	1.16	.59	25.8	14.2	2
5333.8	452.8	11.78	12.01	0.00	0.02	1.00	0.0	1.00	0.0		.57			Mean (Count: 4)
193.0	1.10	0.42	0.42	0.08	0.00	0.08	1.3	0.08	1.2		.02			S.D. (Population)
222.9	1.30	0.49	0.48	0.09	0.00	0.09	1.5	0.09	1.4		.02			S.D. (Sample)

Model, Populn: RMSE .02 Adj (True) S.D. .08 Separation 3.76 Strata 5.34 Reliability (not inter-rater) .93

Model, Sample: RMSE .02 Adj (True) S.D. .09 Separation 4.37 Strata 6.17 Reliability (not inter-rater) .95

Model, Fixed (All Same) Chi-square: 60.4 d.f.: 3 significance (probability): .00

Model, Random (normal) chi-square: 2.9 d.f.: 2 significance (probability): .24

Inter-rater agreement opportunities: 2711 Exact agreements: 834 = 30.8% Expected: 387.6 = 14.3%

Note. Zstd = Infit z-standardized t-statistic, Estim. Discrim = estimated discrimination, Corr. PtBis = point-biserial correlation, Exact Obs. % = observed %, Agree Exp. % = Exact %, Populn = population, RMSE = root mean square standard error, and Adj (True) S.D. = adjusted true S.D.

Table 6.13

Summary of the Topic Measurement in the Three-faceted Rasch Measurement: Participants × Raters × Topics

Total Score	Total Count	Observed Average	Fair(M) Average	Measure	Model S.E.	Infit MNSQ	Infit Zstd	Outfit MNSQ	Outfit Zstd	Estim. Discrim.	Corr. PtBis	Topic
4319.0	362.0	11.93	11.15	0.16	0.02	0.96	–0.5	0.95	–0.6	1.02	.52	Topic 3
5799.0	509.0	11.39	11.50	0.10	0.02	0.75	–4.3	0.75	–4.2	1.23	.57	Topic 2
835.0	72.0	11.60	11.70	0.06	0.05	1.09	0.5	1.05	0.3	0.97	.57	Topic 7
1389.0	148.0	9.39	11.73	0.05	0.04	1.15	1.3	1.09	0.7	1.15	.39	Topic 5
4429.0	360.0	12.30	12.42	–0.08	0.02	1.23	2.8	1.27	3.2	0.75	.50	Topic 1
1548.0	148.0	10.46	12.74	–0.14	0.03	0.66	–3.3	0.66	–3.3	1.24	.40	Topic 6
3016.0	212.0	14.23	12.81	–0.15	0.03	1.38	3.3	1.40	3.4	0.60	.40	Topic 4
3047.9	258.7	11.61	12.01	0.00	0.03	1.03	0.0	1.02	0.0		.48	Mean (Count: 7)
1731.7	144.0	1.40	0.60	0.11	0.01	0.24	2.7	0.25	2.8		.08	S.D. (Population)
1870.4	155.5	1.51	0.65	0.12	0.01	0.26	2.9	0.26	3.0		.08	S.D. (Sample)

Model, Populn: RMSE .03 Adj (True) S.D. .11 Separation 3.33 Strata 4.77 Reliability .92

Model, Sample: RMSE .03 Adj (True) S.D. .12 Separation 3.62 Strata 5.16 Reliability .93

Model, Fixed (All Same) Chi-square: 126.7 d.f.: 6 significance (probability): .00

Model, Random (Normal) Chi-square: 5.7 d.f.: 5 significance (probability): .34

Note. Zstd = Infit z-standardized t-statistic, Estim. Discrim = estimated discrimination, Corr. PtBis = point-biserial correlation, Populn = population, RMSE = root mean square standard error, and Adj (True) S.D. = adjusted true S.D.

6.4.3 Discussion of the Three-faceted Rasch Measurement: Participants × Raters × Text Topics

The results of this analysis revealed the participants' proficiency, raters' severity, and topic difficulty. The participant measurement revealed that the participants' average ability was –0.57; in other words, the participants in this study had inadequate ability to complete the retelling task, as the topics used in the retelling task were difficult for the participants. The maximum value of the logits in this analysis was 1.54 (Participant 9), and the next highest was 1.25 (Participant 74). The third largest logit value was 0.69; thus, the top two participants had remarkably high proficiency, and these two logits were the outliers in this analysis. To examine why those participants became outliers, the audio data were listened to. Participants 9 and 74 told the story and gave their opinions fluently, and the quality of their pronunciation was high; in addition, the contents of their utterances were sufficient for receiving scores of "5" or "4." These participants performed relatively highly compared with the average performance.

The rater severity was not equal. However, the logits of this facet ranged from –0.08 to 0.09; therefore, the differences in rating severity were not large. In fact, the separation and strata indices in the model, which were calculated for the entire population, indicated 3.76 and 5.34, respectively. Thus, the severity of raters was divided into about four. There were four raters in this study; therefore, this result is natural. Moreover, the quality of the rating was better than expected according to the Rasch model; hence, the raters rated the performances appropriately. The fit statistics of the rater facet could be used to interpret rater self-consistency; since there were no misfit or overfit items, the raters' self-consistency was sufficient. This result could be interpreted as showing that the rater training worked effectively.

Finally, the topic facet was considered. The fixed chi-squared test was significant ($\chi^2 = 126.7$, $p < .05$); in other words, the topics had different levels of difficulty. Although one misfit item and one overfit item were found in this facet, as mentioned in the previous section, the practice effect could have caused the misfit. The practice effect on the participants' performance was not considered in this study, and this point will be discussed in future. The separation and strata indices were marked as 3.33 and

4.77, respectively, and the fixed chi-square test was significant; therefore, the topics' difficulty could be divided into at least three levels.

It was expected that the topic difficulty would affect the participants' spoken performances in each study conducted in the previous chapters. The results of the analysis revealed that the topics had different difficulty levels; therefore, the topic facet certainly influenced performance.

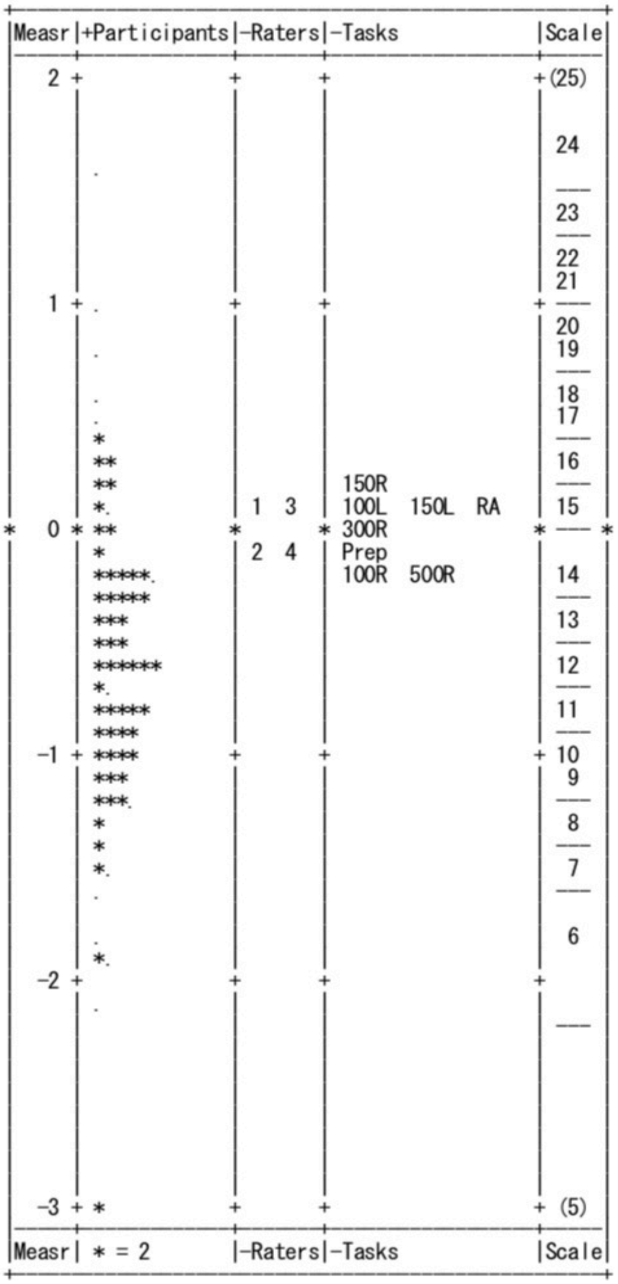
6.5 The Three-faceted Rasch Measurement: Participants × Raters × Tasks (Length × Mode)

6.5.1 Data Fit to the Rasch Model

As in section 6.4.1, the fits of the whole data and the unidimensionality assumption were first checked. There were 1,811 responses in total, 71 of which were unexpected; in sum, 3.9% of the data were unexpected (see Appendix U). Furthermore, the data log-likelihood chi-squared was significant (data log-likelihood $\chi^2 = 6697.95$, $df = 1{,}421$, $p = .00$). Hence, the full dataset of this analysis did not match the Rasch model, and the fit statistics of each facet indicated that there was one misfit item in the task facet; however, there was no misfit in the rater facet. The unidimensionality assumption result was 65.61%, satisfying the unidimensionality requirement.

6.5.2 Results and Discussion for the Three-faceted Rasch Measurement: Participants × Raters × Tasks (Length × Mode)

Figure 6.22 shows the Wright map of this analysis, and the rater column demonstrates that the retelling tasks used in this study tended to be difficult for the participants. Table 6.14 shows the summary of the participant measurement report in the three-faceted (participants × raters × tasks) Rasch measurement. The results were comparable to those reported in the previous section. There were two responses that expected the minimum score; therefore, those responses were automatically excluded from the analysis. The average logit value of the participant facet was –0.63, and the same two participants were the outliers as in the former analysis. The full data on the participant measurement are given in Appendix T, and 19 misfit participants were detected. Seventeen of the 19 participants who were misfit to the model were detected in the unexpected responses. The 19 participants' utterances were listened to, and it was found that they all performed well on two or three tasks; however, most performed poorly on one or two tasks. This could be an explanation for why these 19 participants produced unexpected responses.



```
+-----------------------------------------------------+
|Measr|+Participants|-Raters|-Tasks             |Scale|
|-----+-------------+-------+-------------------+-----|
|   2 +             +       +                   +(25) |
|     |             |       |                   |     |
|     |             |       |                   |     |
|     |             |       |                   | 24  |
|     | .           |       |                   |     |
|     |             |       |                   | --- |
|     |             |       |                   | 23  |
|     |             |       |                   | --- |
|     |             |       |                   | 22  |
|     |             |       |                   | 21  |
|   1 + .           +       +                   + --- |
|     |             |       |                   | 20  |
|     | .           |       |                   | 19  |
|     |             |       |                   | --- |
|     | .           |       |                   | 18  |
|     | .           |       |                   | 17  |
|     | *           |       |                   | --- |
|     | **          |       |                   | 16  |
|     | **          |       | 150R              | --- |
|     | *.          | 1   3 | 100L  150L  RA    | 15  |
*   0 * **          *       * 300R              * --- *
|     | *           | 2   4 | Prep              |     |
|     | *****.      |       | 100R  500R        | 14  |
|     | *****       |       |                   | --- |
|     | ***         |       |                   | 13  |
|     | ***         |       |                   | --- |
|     | ******      |       |                   | 12  |
|     | *.          |       |                   | --- |
|     | *****       |       |                   | 11  |
|     | ****        |       |                   | --- |
|  -1 + ****        +       +                   + 10  |
|     | ***         |       |                   |  9  |
|     | ***.        |       |                   | --- |
|     | *           |       |                   |  8  |
|     | *           |       |                   | --- |
|     | *.          |       |                   |  7  |
|     | .           |       |                   | --- |
|     |             |       |                   |     |
|     | .           |       |                   |  6  |
|     | *.          |       |                   |     |
|  -2 +             +       +                   +     |
|     | .           |       |                   |     |
|     |             |       |                   | --- |
|     |             |       |                   |     |
|     |             |       |                   |     |
|     |             |       |                   |     |
|     |             |       |                   |     |
|     |             |       |                   |     |
|     |             |       |                   |     |
|     |             |       |                   |     |
|  -3 + *           +       +                   + (5) |
|-----+-------------+-------+-------------------+-----|
|Measr| * = 2       |-Raters|-Tasks             |Scale|
+-----------------------------------------------------+
```

Figure 6.22. Wright map of the three-faceted Rasch measurement (participants × raters × tasks).

In the rater facet shown in Table 6.15, there were no misfit raters, and the raters' logits ranged from –0.10 to 0.11. The dispersion of the logits in the rater facet was wider than that in the former analysis; however, the change was modest in size. The fit statistics of this measurement ranged from 0.7 to 1.3; therefore, the raters were fit to the Rasch model. This meant that the rater self-consistency was sufficient. The separation and strata results were 4.01 and 5.68. Finally, the indices of exact and expected agreement were checked, and these results were identical to those in the previous measurement.

The logits of the task facet ranged from –0.23 to 0.16, and there were two overfit tasks (preparation: the retelling task included 1 minute of preparation time; and RA (reading aloud): the retelling task included the pre-task activity of reading the stimuli aloud). In contrast, the retelling task in which the 100-word reading material was given was misfit. Two overfit tasks were given to fewer participants in this study. However, the 100-word reading task was used most frequently. This frequency in the use of each task may have caused the result. The notable point of this analysis is that the difficulty levels of the 100-word reading task and 500-word reading condition were quite close.

Table 6.14

Summary of Participant Measurement in the Three-faceted Rasch Measurement: Participants × Raters × Tasks

Total Score	Total Count	Observed Average	Fair (M) Average	Measure	Model S.E.	Infit MNSQ	Infit Zstd	Outfit MNSQ	Outfit Zstd	Estim. Discrim.	Corr. PtBis	Participants
149.0	12.6	11.79	11.81	–0.63	0.17	0.85	–0.5	0.85	–0.5		.12	Mean (Count: 126)
76.7	5.4	3.10	3.18	0.70	0.18	0.81	1.7	0.83	1.7		.29	S.D. (Population)
77.0	5.4	3.11	3.19	0.70	0.18	0.81	1.7	0.83	1.7		.29	S.D. (Sample)

With extremes, Model, Populn: RMSE .24 Adj (True) S.D. .66 Separation 2.72 Strata 3.96 Reliability .88
With extremes, Model, Sample: RMSE .24 Adj (True) S.D. .66 Separation 2.73 Strata 3.97 Reliability .88
Without extremes, Model, Populn: RMSE .16 Adj (True) S.D. .59 Separation 3.82 Strata 5.43 Reliability .94
Without extremes, Model, Sample: RMSE .16 Adj (True) S.D. .60 Separation 3.84 Strata 5.45 Reliability .94
With extremes, Model, Fixed (All Same) Chi-square: 2709.9 d.f.: 125 significance (probability): .00
With extremes, Model, Random (Normal) Chi-square: 95.7 d.f.: 124 significance (probability): .97

Note. Zstd = Infit z-standardized t-statistic, Estim. Discrim = estimated discrimination, Corr. PtBis = point-biserial correlation, Populn = population, RMSE = root mean square standard error, and Adj (True) S.D. = adjusted true S.D.

Table 6.15

Summary of Rater Measurement in the Three-faceted Rasch Measurement: Participants × Raters × Tasks

Total Score	Total Count	Observed Average	Fair (M) Average	Measure	Model S.E.	Infit MNSQ	Infit Zstd	Outfit MNSQ	Outfit Zstd	Estim. Discrim.	Corr. PtBis	Exact Obs. %	Agree. Exp. %	Raters
4478.0	396.0	11.31	11.50	0.11	0.02	1.01	0.1	1.01	0.1	0.90	.55	34.6	14.9	3
4545.0	395.0	11.51	11.71	0.07	0.02	0.93	–1.0	0.93	–0.9	1.04	.59	26.1	15.0	1
4858.0	396.0	12.27	12.44	–0.08	0.02	1.04	0.5	1.05	0.7	0.88	.58	34.7	14.8	4
4888.0	395.0	12.37	12.55	–0.10	0.02	1.06	0.8	1.05	0.7	1.18	.60	25.6	14.7	2
4692.3	395.5	11.86	12.05	0.00	0.02	1.01	0.1	1.01	0.2		.58			Mean (Count: 4)
182.6	0.5	0.46	0.46	0.09	0.00	0.05	0.7	0.05	0.7		.02			S.D. (Population)
210.9	0.6	0.54	0.53	0.11	0.00	0.06	0.8	0.06	0.8		.02			S.D. (Sample)

Model, Populn: RMSE .02 Adj (True) S.D. .09 Separation 4.01 Strata 5.68 Reliability (not inter-rater) .94

Model, Sample: RMSE .02 Adj (True) S.D. .11 Separation 4.67 Strata 6.55 Reliability (not inter-rater) .96

Model, Fixed (All Same) Chi-square: 68.3 d.f.: 3 significance (probability): .00

Model, Random (Normal) Chi-square: 2.9 d.f.: 2 significance (probability): .23

Inter-rater agreement opportunities: 2,371 Exact agreements: 717 = 30.2% Expected: 351.8 = 14.8%

Note. Zstd = Infit z-standardized t-statistic, Estim. Discrim = estimated discrimination, Corr. PtBis = point-biserial correlation, Exact Obs. % = observed %, Agree Exp. % = Exact %, Populn = population, RMSE = root mean square standard error, and Adj (True) S.D. = adjusted true S.D.

Table 6.16

Summary of Topic Measurement in Three-Facet Rasch Measurement: Participants × Rater × Tasks

Total Score	Total Count	Observed Average	Fair (M) Average	Measure	Model S.E.	Infit MNSQ	Infit Zstd	Outfit MNSQ	Outfit Zstd	Estim. Discrim.	Corr. PtBis	Topics
3981.0	360.0	11.06	11.24	0.16	0.02	0.95	–0.6	0.92	–1.0	1.09	.57	150R
824.0	72.0	11.44	11.47	0.12	0.05	0.15	–8.7	0.15	–8.4	1.60	.61	Reading aloud
2767.0	212.0	13.05	11.56	0.10	0.03	1.00	0.0	0.98	–0.1	0.99	.51	100L
2747.0	210.0	13.08	11.61	0.09	0.03	0.96	–0.4	0.96	–0.3	1.07	.44	150L
1389.0	148.0	9.39	12.00	0.01	0.04	1.24	2.0	1.17	1.3	1.09	.38	300R
835.0	72.0	11.60	12.33	–0.06	0.06	0.29	–5.3	0.36	–4.4	1.40	.67	Preparation
1548.0	148.0	10.46	12.98	–0.20	0.04	0.73	–2.6	0.73	–2.5	1.20	.39	500R
4678.0	360.0	12.99	13.13	–0.23	0.02	1.42	4.9	1.49	5.5	0.58	.49	100R
2346.1	197.8	11.63	12.04	0.00	0.04	0.84	–1.3	0.84	–1.3		.51	Mean (Count: 8)
1351.0	105.8	1.26	0.66	0.14	0.01	0.41	4.0	0.40	3.9		.10	S.D. (Population)
1444.3	113.1	1.35	0.71	0.15	0.01	0.44	4.2	0.43	4.1		.10	S.D. (Sample)

Model, Populn: RMSE .04 Adj (True) S.D. .13 Separation 3.45 Strata 4.93 Reliability .92
Model, Sample: RMSE .04 Adj (True) S.D. .14 Separation 3.71 Strata 5.28 Reliability .93
Model, Fixed (All Same) Chi-square: 193.2 d.f.: 7 significance (probability): .00
Model, Random (Normal) Chi-square: 6.7 d.f.: 6 significance (probability): .35

Note. Zstd = Infit z-standardized t-statistic, Estim. Discrim = estimated discrimination, Corr. PtBis = point-biserial correlation, Populn = population, RMSE = root mean square standard error, and Adj (True) S.D. = adjusted true S.D.

The results of study 2 in this series revealed that when the stimuli were given visually—in other words, when reading stimuli were given—the participants tended to produce longer utterances. In addition, study 3 revealed that the participants could produce many utterances when a pre-task reading aloud activity was assigned to the participants before they performed the retelling tasks, or when planning time was given, in comparison with the no-treatment case. Therefore, it was expected that the 100- or 150-word listening tasks would be the hardest for the participants. However, the three-faceted Rasch measurement of participants × raters × tasks indicated that the order of task difficulties was as follows: 150-word reading, reading aloud, 100-word listening, 150-word listening, 300-word reading, prepared retelling, 500-word reading, and 100-word reading. Therefore, this result did not agree with the findings of the previous studies; however, this can be explained in that the previous studies considered the complexity, accuracy, and fluency of the spoken performances. The participants in these studies could produce longer utterances when they were assigned to read the stimuli; moreover, they could produce more words when they were assigned to prepare or read the stimuli aloud before carrying out the retelling tasks. However, the longer utterances may have included more errors than the shorter ones did, and if the quality of pronunciation was poor, the longer utterances may have been caused by halo effects. Nevertheless, considering how to elicit test takers' performance is an important point for training their speaking ability, since their errors could be sources of instruction. Considering which retelling task types elicit more sophisticated performance is also important, as learners could be assessed by the performance on the task. Therefore, it is necessary for users of retelling tasks—that is, teachers—to consider how to elicit both longer and higher quality performances, depending on the use situation of the retelling task.

6.6 Summary of Study 4

The purpose of study 4 was to observe the difficulty of the text topics and task conditions in the retelling tasks, as well as synthesizing the results of studies 1, 2, and 3. The results of the three-faceted

Rasch measurement (participants × raters × topics) revealed that the text topics had different difficulty levels, and these differences affected the participants' performance. Another analysis of MFRM, which was designed as participants × raters × tasks, indicated that the task difficulties were certainly different, and this result was not in accordance with the previous studies in this series.

Chapter 7
General Discussion and Conclusion

7.1 Overview of Findings and Conclusion

The aim of Study 1 was to determine the differences in spoken performance for retelling and summarizing tasks. The motivation for this study was Kissner (2006), who claimed that retelling and summarizing are similar tasks, and even veteran teachers cannot distinguish between these two tasks; therefore, the author considered that there is a room for investigation to identify the differences between these two tasks and attempt to define them clearly.

In this study, a pilot study and main experiment were conducted. The pilot study was conducted to observe the differences in spoken performance for the two different tasks. The results of this pilot study indicate no difference between the two task conditions; therefore, the performance on these tasks was essentially the same. In other words, the naming of the task did not affect the spoken performance.

The purpose of the main experiment was to determine the differences in spoken performance when different task directions were given to the participants; in addition, the effects of the length of text to spoken performances were also investigated. Generally, people summarize stories when they receive longer sentences as the stimulus; therefore, the author considered that the length of input material affected the spoken performances. To observe the effects of the task direction, the following directions were given to answer the retelling task: “You can use the words or expressions directly if you can recall them, but if you cannot remember the exact expressions, phrases, or sentences, just use your own expressions.” On the other hand, the directions for the summarizing task were: “You should paraphrase the sentence or expressions that you have read as much as possible, but if you cannot find an alternative expression, you can use the expressions provided in the material.”

A two-way ANOVA was conducted to identify the interactions of task direction and text length, and the results showed that there was no significant interaction in all CAF indices . However, significant main effects were found in the factor of text length. For the results of the analysis of main effects, the following indices were greater when the participants answered short texts: number of word types, number of syllables, words per minute, syllables per second, number of disfluency markers, and Guiraud index. This means the participants in this experiment could tell more various words fluently, but they tend to use many disfluency markers when short material was given. On the other hand, significant main effects were found in the indices of the number of word tokens, pause length, type token ratio, words per AS-unit, and frequency of reproduction. The participants of this experiment performed many words under the retelling task condition, and the total length of the silent pause was short while they performed their speaking skills under retelling tasks. In addition, they tended to use complex structure and the same expressions of the original text in the retelling condition.

The participants tried to use the expressions written in the original text in the retelling condition; simultaneously, they tried to paraphrase the sentence in the summarizing condition. In all cases, the task direction affected the use of expressions. This result indicates that if a task user (i.e., language teacher or test developer) wants language learners or test takers to learn expressions through retelling or summarizing tasks, clear directions, such as those mentioned above, should be given.

Both retelling and summarizing tasks should be used for different purposes in teaching. As mentioned above, learners tended to try to use expressions shown in the stimuli in retelling tasks. Therefore, when teachers want students to use target expressions in their utterances, retelling tasks should be used. In reality, “keyword retelling” is used as a post-reading activity in secondary schools (e.g., Yamashita, 2014). On the other hand, learners need to obtain the skills for paraphrasing or compressing long documents, so college students should be assigned the summarizing task. As mentioned in chapter 2, Cumming (2013) claimed that integrated tasks are suitable for university students because they need to listen to lectures and discuss the contents of the topics. In addition, they need to read a large amount of texts and write a report or thesis in their curriculum. Thus, those two tasks should be used properly, depending on L2 developmental stages.

The purpose of Study 2 was to observe the effects of text length, difficulty, and input mode on the spoken performance in retelling. There were two experiments in this study: Experiment 2A aimed to determine the effects of the factors mentioned above; however, there was a limitation in terms of the experiment design; therefore, Experiment 2B was conducted to overcome this limitation. The results of both Experiments 2A and 2B revealed that the participants performed more words when they obtained information from written material than audio material. In other words, the reading-speaking retelling task elicited longer performance. On the other hand, when longer listening material was given, the participants could not perform well; therefore, this task condition seemed to be the most difficult of the four task conditions (short-reading, long-reading, short-listening, and long-listening).

Spoken performance in terms of accuracy and complexity was not affected largely as the length and number of utterances and fluency. L2 speakers need to use incomplete grammar and vocabulary knowledge when they say something; therefore, the accuracy and complexity of utterances are not changed in many cases of task condition. As Bygate (1987) claimed, one’s skills to produce spoken language depend on one’s knowledge of language. It is difficult to obtain language knowledge when learners are struggling with tasks; on the other hand, they may notice skills or strategies to perform their skills when they speak. Thus, the accuracy and complexity of utterances do not differ greatly in a short time, but the fluency and the amount of utterances tend to be improved in accordance with experience of answering tasks.

The results of Study 3 revealed that the pre-task planning and pre-task reading aloud task partially affect participants’ oral performances in terms of fluency; however, as with the results of Study 2, the accuracy and complexity did not change, even if pre-task treatment was given to the participants. Previous researchers, such as Foster and Skehan (1996) and Mehnert (1998), have reported that the quality of utterances improved if pre-task preparation time was given; however, participants were given 10 minutes of preparation. In this study, only one-minute preparation time was given; therefore, this short planning time might have affected the results of this study. However, there is almost no preparation time to communicate with others in the real world; hence, such a long preparation time is inauthentic for speaking practice.

A pre-task reading aloud task might be a good practice for pronunciation, but in actuality, the spoken performances in this task condition in terms of the pronunciation did not differ from the performance of the control group. Fluency was improved as well in Experiment 3A; however, other factors of speech, such as accuracy and complexity, did not change.

Both Skehan (2009) and Robinson (1995, 2005) explained the effects of task load for L2 spoken performances based on the speech model advocated by Levelt (1989). Robinson (1995) claimed that a heavy load task, such as the time pressured task, elicits higher quality oral performances because learners plan the utterances and the contents of the speech thoroughly in the phase of conceptualization, and such deep thinking activates subsequent lexical and grammatical encoding phases. Therefore, the heavy load task can elicit more higher quality spoken performances than when an easy task is assigned. On the other hand, Skehan (2009) argued that differences of task features affect different phases of speech production. The results of this study revealed that the accuracy and complexity of utterances were not changed in the planned and unplanned groups. There were no differences between the group assigned to read the stimuli aloud before responding to the retelling task and the group not assigned to read aloud. However, participant fluency was improved in the planned group and the group assigned the pre-task reading aloud activity. In sum, low-load tasks in Experiments 3A and 3B elicited longer and more fluent utterances. Therefore, the results of those two experiments support Skehan's limited capacity hypothesis.

The results of Study 4 revealed that difficulties of the topics differ, and it supported the idea that differences in the topics of the stimuli affect performance. In addition, the task difficulties also differed, but the results of this study did not match the results of Study 2 and Study 3. In this study, the most difficult task was a 150-word reading, and the easiest task was a 100-word reading among eight tasks. On the other hand, the author expected the 100-word listening task or the 150-word listening task to be more difficult than the tasks that instructed participants to read input materials. However, the difficulty levels of the 100-word and 150-word listening tasks were third and fourth out of eight tasks.

7.2 Limitations of the Current Study and Suggestions for Further Work

The largest limitation is a lack of proficiency measurement in terms of reading, listening, and vocabulary skills and knowledge. In Study 2, the effects of an input mode on spoken performance were observed; however, the effects of a participant's language abilities were not considered. In a future study, the relationship between proficiency and task conditions should be assessed.

Another limitation is a lack of proficiency measurement in terms of reading, listening, and vocabulary skills and knowledge. In Study 2, the effects of an input mode on spoken performance were observed; however, the effects of a participant's language abilities were not considered. In a future study, the relationship between proficiency and task conditions should be assessed.

The last limitation is the design of the data collection. The author tried to synthesize the results of Studies 1, 2, and 3, but it was impossible to observe the relationship between the topic and the task facets because there were many missing values. Because of the missing values, it became impossible to conduct a generalizability theory to observe the effects of the facets on task difficulty. Therefore, the author plans to collect new data with a fully crossed design in the next study to try to observe the effects of the topics of stimuli or input mode on the difficulty of the task.

7.3 Pedagogical Implications

Retelling tasks is certainly practical because test developers can control their material. The motivation for this study is to ascertain what factors affect the difficulty of tasks and how the utterances were changed in different task conditions. The results of this study indicates that participants were able to perform well while the material was given in a written format and most task conditions do not affect performance in terms of accuracy and complexity. From the perspectives of language education, accuracy should be the most important factor. L2 speakers can speak fluently if they become accustomed to a task (Bygate, 1987), and bringing out the accuracy of their speech is difficult for them if they do not concentrate their efforts on learning grammar and vocabulary.

The participants in this study could produce the same phrases used in input materials, while a task director asks to use the same expressions of material. Therefore, test developers or teachers should consider what kind of grammatical or lexical elements they should focus on eliciting, since learners can pay attention to using target expressions and can learn words, phrases, or grammatical rules. The effects of such directions may be small, but if this kind of activity is repeated, learners can improve their speaking skills effectively.

As mentioned above, the spoken performances were not greatly changed in most of the task conditions; rather, the retelling task is reliable task to elicit EFL learners' spoken performances. The speaking ability of Japanese learners of English is very low (Ishilawa et al., 2009; MEXT, 2015). As referred to in Experiment 2A, the task demands of the retelling task match the can-do descriptor of B1.2 of CEFR-J. Hence, this task can be used as a benchmark of achievement for Japanese senior-high or university students. Therefore, this type of task should be used as a pedagogical task or a test task in speaking lessons.

7.4 Concluding Remarks

The current study found that spoken performances, in terms of accuracy and complexity, are not very different while the task condition is changed; on the other hand, fluency and amounts of utterance tend to change. If the amounts of utterance are increased, the opportunities to give feedback are increased. Therefore, even if the quality of a spoken performance in terms of accuracy and complexity is unchanged, trying to change the task condition and elicit more utterances might be more important than paying attention to the accuracy and complexity of an observed performance.

References

American Council on the Teaching of Foreign Languages (ACTFL). (2012). *ACTFL proficiency guidelines 2012.* Retrieved from https://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf

Aoshima, T. (2014). *Japanese EFL learners' production of collocations in story retelling tests.* (Unpublished master's thesis). University of Tsukuba, Japan.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19*(4), 453–476. doi:10.1191/0265532202lt240oa

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

Barkaoui, K., Brooks, L., Swain, M., & Lapkin, S. (2013). Test-takers' strategic behaviors in independent and integrated speaking tasks. *Applied Linguistics, 34*, 304–324. doi:10.1093/applin/ams046

Blutė, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 21–46). Amsterdam: Benjamins.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York, NY: Routledge.

Bosker, H. R., Pinget, A. F., Quené, H., Sanders, T., & De Jong, N. H. (2012). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing, 30*, 159–175. doi:10.1177/0265532212455394

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English for academic purposes speaking tasks* (TOEFL Monograph No. MS-29). Princeton, NJ: Educational Testing Service.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. Boston, MA: Allyn & Bacon.

Bygate, M. (1987). *Speaking*. Oxford: Oxford University Press.

Clarke, M. A. (1979). Reading in Spanish and English: Evidence from adult ESL students. *Language Learning, 29*, 121–150. doi:10.1111/j.1467-1770.1979.tb01055.x

Cumming, A. (2013). Assessing integrated skills. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 216–229). Boston, MA: John Wiley & Sons Ltd.

De Bot, K. (1992). A bilingual production model: Levelt's 'Speaking' model adapted. *Applied Linguistics, 13*(1), 1–24. doi:10.1093/applin/13.1.1

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing, 11*, 125–144. doi:10.1177/0265532211424479

Dörnyei, Z., & Scott, M. L. (1997). Communication strategies in a second language: Definitions and taxonomies. *Language Learning, 47*(1), 173–210. doi:10.1111/0023-8333.51997005

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt am Main: Peter Lang.

Educational Testing Service. (2016). *Test and score data summary for TOEFL iBT® Tests: January 2016–December 2016 Test Data*. Retrieved from https://www.ets.org/s/toefl/pdf/94227_unlweb.pdf

Engelhard, Jr. G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York, NY: Routledge.

Ferris, D., & Tagg, T. (1996). Academic listening/speaking tasks for ESL students: Problems, suggestions, and implications. *TESOL Quarterly, 30*(2), 297–320. doi:10.2307/3588145

Foster, P., & Skehan, P. (1996). The influence of planning and task-type on second language performance. *Studies in Second Language Acquisition, 18*, 299–323. doi:10.1017/S0272263100015047

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21*, 354–375.

Frost, K., Elder, C., & Wigglesworth, G. (2011). Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers' oral performances. *Language Testing, 29*(3), 345–369. doi:10.1177/0265532211424479

Fulcher, G. (2003). *Testing second language speaking*. Harlow, UK: Pearson Education.

Fulcher, G. (2015). Assessing second language speaking. *Language Teaching, 48*, 198–216 doi:10.1017/S0261444814000391

Fulcher, G., & Davidson, F. (2012). *The Routledge handbook of language testing*. New York: Routledge.

Gilabert, R., Baron, J., & Levkina, M. (2011). Manipulating task complexity across task types and modes. In P. Robinson (Ed.), *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (pp. 91–104). Amsterdam: John Benjamins.

Hickling, R., & Ichikawa, Y. (2010). English charge! Tokyo: Kinseido.

Hickling, R., & Osaki, S. (2013). English upload. Tokyo: Kinseido.

Higgs, T. V., & Clifford, R. (1982). The push toward communication. In T. V. Higgs (Ed.), *Curriculum, competence and the foreign language teacher* (pp. 243–265). Skokie, IL: National Textbook Company.

Hirai, A. (2015). Jugyou o ikasu sutorii riterinngu tesuto no katsuyou [Application of the Story Retelling Test utilize class], *Otsuka Forum, 33*, 49–69.

Hirai, A., & Koizumi, R. (2008). Validation of an EBB scale: A case of the Story Retelling Speaking Test. *JLTA (Japan Language Testing Association) Journal, 11*, 1–20.

Hirai, A., & Koizumi, R. (2009). Development of a practical speaking test with a positive impact on learning using a story retelling technique. *Language Assessment Quarterly, 6*, 151–167. doi:10.1080/15434300902801925

Hirai, A., & Koizumi, R. (2013). Validation of empirically derived rating scales for a story retelling speaking test. *Language Assessment Quarterly, 10*, 398–422. doi:10.1080/15434303.2013.824973

Hirai, A., & O’ki, T. (2011). Comprehensibility and naturalness of text-to-speech synthetic materials for EFL listeners, *JACET Journal, 53*, 1–17.

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 1–20). Amsterdam: John Benjamins.

Huang, H. T. D., Hung, S. T. A., & Hong, H. T. V. (2016). Test-Taker Characteristics and Integrated Speaking Test Performance: A Path-Analytic Study. *Language Assessment Quarterly, 13*(4), 283–301. https://doi.org/10.1080/15434303.2016.1236111

Iino, A., & Yabuta, Y. (2013). Relationship among reading aloud, shadowing, and speaking. *Chubu chiku eigo kyoiku gakkai kiyo, 42*, 139–146.

Ishikawa, S., Nakamura, Y., Ito, Y., Schneider, S., & Sugimori, N. (2009). *A study on measurement and assessment for oral English proficiency.*

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*(1), 24–49.

Jong, H. N., Steinel, P. M., Florjin, F. A., Schoonen, R., & Hulstjin, H. J. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition, 34*, 5–34.

Kissner, E. (2006). *Summarizing, paraphrasing, and retelling: Skills for better reading, writing, and test taking.* Portsmouth: Heinemann.

Klingner, J. K. (2004). Assessing reading comprehension. *Assessment for Effective Intervention, 29*(4), 59–70.

Koizumi, R., & Hirai, A. (2010). Exploring the quality of the Story Retelling Speaking Test: Roles of story length, comprehension questions, keywords, and opinions. *ARELE (Annual Review of English Language Education in Japan), 21,* 211–220.

Koizumi, R., & Hirai, A. (2012). Comparing the story retelling speaking test with other speaking tests. *JALT (Japan Association for Language Teaching) Journal, 34,* 35–59.

Kormos, J. (2014). *Speech production and second language acquisition*. London: Routledge.

Kuiken, F., & Vedder, I. (2011). Task complexity and linguistic performance in L2 writing and speaking. In P. Robinson (Ed.), *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (pp. 91–104). Amsterdam: John Benjamins.

Levelt, W. J. M. (1989). S*peaking: From intention to articulation.* Cambridge, MA: Massachusetts Institution of Technology Press.

Lewkowicz, J. A. (1997). The integrated testing of a second language. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education: Language testing and assessment 7* (pp. 121–130). Dordrecht, the Netherlands: Kluwer.

Li, L., Chen, J., & Sun, L. (2014). The effects of different lengths of pretask planning time on L2 learners' oral test performance. *TESOL Quarterly, 49,* 38–66. doi: 10.1002/tesq.159

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

McNamara, T. (1996). *Measuring second language performance*. Essex: Addison Wesley Longman Limited.

Mehnert, U. (1998). The effects of different length of time for planning on second language performance. *Studies in Second Language Acquisition, 20,* 83–108. doi:10.1017/S0272263198001041

Ministry of Education, Culture, Sports, Science and Technology. (2008). The course of study for junior school: Foreign languages. Retrieved from http://www.mext.go.jp/a_menu/shotou/new-cs/youryou/chu/gai.htm

Ministry of Education, Culture, Sports, Science & Technology. (2009). *The course of study in higher school: Foreign languages.* Retrieved from http://www.mext.go.jp/a_menu/shotou/new-cs/youryou/1304427.htm

Ministry of Education, Culture, Sports, Science and Technology. (2011). Heisei 24 nendo kouritsu koutougakkou ni okeru kyouikukatei no hensei, jisshi joukyou tyousa (B hyou) no kekka ni tsuite [The result of the survey on the curriculum organization and implementation at public senior high schools in 2011 (Part B)]. Retrieved from http://www.mext.go.jp/a_menu/shotou/new-cs/__icsFiles/afieldfile/2011/01/25/1301650_2_1.pdf

Ministry of Education, Culture, Sports, Science and Technology. (2013). English education reform plan corresponding to globalization. Retrieved fromhttp://www.mext.go.jp/english/topics/1343591.htm

Ministry of Education, Culture, Sports, Science and Technology. (2017). Daigaku nyuugakusha sennbatsu kaikaku [Reformation of entrance examination for university], Retrieved from http://www.mext.go.jp/b_menu/houdou/29/07/__icsFiles/afieldfile/2017/07/18/1388089_002_1.pdf

Ministry of Education, Culture, Sports, Science and Technology. (2017). *The course of study for junior high school.* Retrieved from http://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2017/06/21/1384661_5.pdf

Ministry of Education, Culture, Sports, Science and Technology. (n.d.). Eigo kyouiku kaizen purann [Plan for the improvement of English education]. Retrieved from http://www.mext.go.jp/a_menu/kokusai/gaikokugo/1371433.htm

Mizumoto, A. (n.d.). *Koukaryou [effect size]*. Retrieved from www.mizumot.com/stats/effectsize.xls

Murphy, J. M. (1991). Oral communication in TESOL integrating speaking, listening, and pronunciation. *TESOL Quarterly, 25*, 51–75. doi:10.2307/3587028

Negishi, M. (2013, December). Zenkoku 47 todouhuken no koukounyushibunseki kara kangaeru tesuto dezain to tyuugakkou 3nennkan no shidou [Considering test design and instruction in three years of junior high school through 47 prefectural entrance examinations for high school]. In M. Negishi (Proposer), Symposium conducted at the meeting of the ARCLE (Action Research Center for Language Education). Retrieved from http://www.arcle.jp/report/2013/0005_2.html

Oller, J. (1979). *Language tests at school: A pragmatic Approach*. Essex: Longman.

Onoda, S. (2014). An exploration of effective teaching approaches for enhancing the oral fluency of EFL students. In T. Muller, J. Adamson, P. S. Brown, & S. Herder (Eds.), *Exploring EFL fluency in Asia* (pp. 87–108). London: Palgrave Macmillan.

Oxford University Press. (2016). *Oxford dictionary of English*. Oxford: Author.

Plakans, L. (2007). *Second language writing and reading-to-write assessment tasks: A process study* (Unpublished doctoral dissertation). The University of Iowa, Iowa City, Iowa.

Plakans, L. (2012). Writing integrated items. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 249–261). NY: Routledge.

Plakans, L. (2013). Assessment of integrated skills. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 204–212). Hoboken, NJ: Wiley-Blackwell.

Qian, D. D. (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers. *Language Assessment Quarterly, 6*(2), 113–125. doi:10.1080/15434300902800059

Rausch, A. (2012). Complexity, accuracy and fluency: Toward a conceptual model of communicative and sociolinguistic frameworks, *Bulletin of the Faculty of Education, Hirosaki University, 108,* 19–30.

Reed, D. K., & Vaughn, S. (2012). Retell as an indicator of reading comprehension. *Scientific Studies of*

Reading, 16(3), 187–217.

Riggenbach, H. (1991). Towards an understanding of fluency: A microanalysis of nonnative speaker conversation. *Discourse Processes, 14,* 423–441. doi:10.1080/01638539109544795

Robinson, P. (1995). Task complexity and second language narrative discourse. *Language Learning, 45,* 99–140. doi: 10.1111/j.1467-1770.1995.tb00964.x

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics, 43,* 1–32. doi: 10.1515/iral.2005.43.1.1

Roelofs, A. (2003). Shared phonological encoding processes and representations of languages in bilingual speakers. *Language and Cognitive Processes, 18*(2), 175–204. doi:10.1080/01690960143000515

Shimizu, M. (2009). An examination of reliability of an oral reading test for EFL learners using generalizability theory. *ARELE (Annual Review of English Language Education in Japan), 20,* 181–190.

Shimizu, S. (2004). *JACET 8000 analysis program: V8an* (Revised web ed.). Retrieved from http://www.tcp-ip.or.jp/~shim/j8web/j8web.cgi

Shishido, M., Allen, B., & Takahashi, M. (2012). *AFP world news report: Looking at the world through AFP news*. Tokyo: Seibido.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing, 11,* 99–123. doi:10.1177/026553229401100202

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, P. (2001). Tasks and language performance assessment. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 167–185). Essex, UK: Pearson Education.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*(4), 510–532. doi:10.1093/applin/amp047

Syllable Counter & Word Count [Computer software]. Retrieved from http://www.wordcalc.com/

Takahashi, M. (2013). Role of oral reading in the development of reading ability. *The Japanese Journal of Educational Psychology, 61,* 95–111.

Takahashi, N., & Matsumoto, S. (2016). *Tyu koutougakkou eigoka ni okeru hanasu chikara wo takamerutame no shidou no arikata ni kansuru kenkyu [A study of ideal methods of instruction to increase English speaking skills in junior and senior high school English classes]*. Retrieved from http://www1.iwate-ed.jp/kenkyu/siryou/h27/h27_0904_1.pdf

Takayama, Y. (2007). Effects of self-monitored intensive reading-aloud program on the participants' oral performance. *Bulletin of Tokyo Gakugei University, Humanities and Social Sciences, 1,* 58, 37–44.

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–273). Amsterdam: John Benjamins.

Tokyo Academy (Ed.). (2012). *Opun sesame siriizu 2014 nenndo kyouin saiyou shiken mondaishuu 2: Ippankyouyou 1 [Open sesame series workbook volume 2 for teacher employment examination: General education 1]*. Tokyo: Shichiken Shuppan.

2014 nenndo kyouin saiyou shaken mondaishuu 2, Ippankyouyou 1 [A workbook 2

Tono, Y. (Ed.). (2013). *The CEFR-J handbook: A resource book for using CAN-DO descriptions for English language teaching.* Tokyo, Japan: Taishukan-shoten.

Uenaka, R., & Korechika, S. (2013). *Fast pass for the TOEIC ® test.* Tokyo: Cengage Learning.

Van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversations. *TESOL Quarterly, 23,* 489–508. doi:10.2307/3586922

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing, 29,* 325–344. doi:10.1177/0265532211424478

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, *17*(1), 65–83.

Weir, C. J., & Wu, J. R. W. (2006). Establishing test form and individual task comparability: A case study of a semi-direct speaking test. *Language Testing*, *23*(2), 167–197. doi:10.1191/0265532206lt326oa

Wall, D., & Horák, T. (2006). *The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe: Phase 1, the baseline study* (TOEFL Monograph Series MS-34). Princeton, NJ: Educational Testing Service.

Wall, D., & Horák, T. (2007). Using baseline studies in the investigation of test impact. *Assessment in Education: Principles, Policy & Practice*, *14*(1), 99–116. doi:10.1080/09695940701272922

Wall, D., & Horák, T. (2008). *The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe: Phase 2, coping with change* (TOEFL iBT Research Series No. 5). Princeton, NJ: Educational Testing Service.

Yamashita, Y. (2014). *Tasuku gata riidinngu jugyou ni yoru tyuu koutougakkou eigoka no jugyou zukuri [Planning English classes for junior and senior high school using "task-based reading teaching"]*. Retrieved from http://www.shiga-ec.ed.jp/www/contents/1438647296290/files/16kiyou.pdf

Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity, and accuracy in L2 monologic oral production. *Applied Linguistics*, *24*, 1–27. doi:10.1093/applin/24.1.1

Appendices

Appendix A. Format of Test Paper for Reading-Speaking Retelling Tasks

今から、下記の英文を◯分間で読んでください。読んでいる最中、声を出したり、メモを取ったりしてはいけません。黙読時間が終わったら、すぐにテスト用紙を裏返してください。その後、読んだ内容について話し、それに対する意見・感想を述べてもらいます。

Reading Material

※ 上記下線部はタスク条件によって表現が変わります。

---<Next page>---

Retell as much of the story as you can in English within two minutes and half. You can check the keywords while you are talking. At the end of your speech, be sure to include your opinions about the story. You will hear a signal 30 seconds before closing.

今読んだ内容をできるだけ詳しく、2分30秒間英語で話してください。話しながら、キーワードを見てもかまいません。読んだ内容を話し終えたら、必ず、その内容についての感想や意見を英語で述べてください。終了30秒前に合図をしますので、感想を始める目安にできます。話す前に、名前と学籍番号を言ってから話してください。

Keywords

Appendix B. Format of Test Paper for Listening-Speaking Retelling Tasks

今から英文を 2 度流します。聞いた英文の内容を話し、それに続いて聞いた内容に対する意見・感想を述べてください。持ち時間は 2 分 30 秒間です。音声が流れている間、メモを取ることはできません。

音声終了後、指示を出しますのですぐにテスト用紙を裏返してください。

---<Next page>---

Retell as much of the story as you can in English within two minutes and half. You can check the keywords while you are talking. At the end of your speech, be sure to include your opinions about the story. You will hear a signal 30 seconds before closing.

今聞いた内容をできるだけ詳しく、2 分 30 秒間英語で話してください。話しながら、キーワードを見てもかまいません。内容を話し終えたら、必ず、その内容についての感想や意見を英語で述べてください。終了 30 秒前に合図をしますので、感想を始める目安にできます。話す前に、名前と学籍番号を言ってから話してください。

Keywords:

Appendix C. Text of Material 2A-1

Most people who use the Internet, email, and cell phones believe we should be able to use this technology freely, in privacy, and without censorship or spying. More and more, however, we are finding that there is no such thing as privacy in cyberspace. Various groups – ranging from private hackers to corporate spies and government security agencies – are gaining access to our messages, calls, and Internet sites. Although governments sometimes try to protect our privacy with laws prohibiting cyber-spying, the fact is that in many cases the governments themselves are among the biggest spies and censors of cyberspace.

This text is quoted from *AFP world news report: Looking at the world through AFP news* (Shishido, Allen & Takahashi, 2012, p. 58).

Appendix D. Text of Material 2A-2

Often multinational corporations have tried to “encourage” people in poorer countries to adopt Western style diets. Hamburgers, fried chicken, French fries, and Coca Cola can easily gain a status appeal among people in developing countries, yet such foods can have hidden dangers. Their production can be damaging to the environment, since the raising of cattle, in particular, requires large amounts of water, land, and energy. Moreover, these foods can be damaging to health. They can cause problems of obesity, diabetes, and behavioral disorders such as ADHD. Especially in places where people already eat mainly vegetarian diets, derived from ecologically sustainable crops, it is unwise to urge people to adopt a Western style diet. Rather, it would be better for us in richer countries to change our dietary habits and consume more of the ecologically-friendly, healthy foods eaten by people in other countries.

This text is quoted from *AFP world news report: Looking at the world through AFP news* (Shishido, Allen & Takahashi, 2012, p. 58).

Appendix E. Text of Material 2A-3

Recently, we have been facing serious economic problems. The world economy is getting worse every year. Because of this, more and more people want to save money. This causes a cooling down of the economy. The employment rates of many countries are also still low. Some young people have tight schedules: For example, they seek work and go to interviews almost every day. However, those who run businesses do not want to employ many people because the affairs of their companies are in pretty bad shape and some are even going broke. As a result, the number of unemployed people and part-time workers is increasing. The government of each country needs to meet the demands for employment in order to make the economy better.

This text is quoted from Aoshima (2014) p.105.

Appendix F. Text of Material 2A-4

We have to make efforts to stay healthy otherwise we might get sick or have severe pain in parts of our body. If we get sick or feel pain due to unhealthy habits, we have to be careful to take medicine or to see a doctor. Ill health also leads us to get angry easily. This sometimes results in an undesirable, tense atmosphere among people. To avoid these problems, we should have a good night's sleep. There is sufficient evidence that people who have a good sleep are healthier than people who do not. By getting enough sleep, we can stay healthy. We can avoid missing work for personal reasons, such as poor health. In this way, we can take responsibility for our own lives.

This text is quoted from Aoshima (2014) p.105.

Appendix G. Text of Material 1B-1 and 2B-4 (Topic 1)

Studying can be very tiring and stressful. These days many students are choosing to study in coffee shops instead of libraries. Drinking their favorite coffee and sitting in comfortable chairs makes it easy for them to relax. They can also enjoy talking to their friends. These are things they cannot do in libraries. Many coffee shops have free Internet connection. Students just need to bring their laptop computers. Internet cafes are even better because computers are already there. Libraries also have computers with Internet connection, but students often have to wait before they are able to use them.

The original text is quoted from Hickling and Ichikawa (2010). The words written in parentheses are not shown in the original text.

Studying can be very tiring and stressful. These days many students are choosing (to study) in coffee shops instead of libraries. (Drinking) their favorite coffee and sitting in comfortable chairs makes it easy for them to relax. They can also enjoy (talking) to their friends. These are things they cannot do in libraries. Many coffee shops have free Internet connection. Students just need (to bring) their laptop computers. Internet cafes are even better because computers are already there. Libraries also have computers with Internet connection, but students often have to wait before they are able (to use) them.

Appendix H. Text of Material 1B-2, 2B-2, 3A-2, and 3B-2 (Topic 2)

Grocery stores sell fruits and vegetables from many different places. Have you ever wondered which grocery store to choose and why? Here are a few good reasons to buy locally produced food.

Firstly, locally produced food is fresher and tastier than the food produced in other regions or countries. Food grown in your community was picked within the past day or two. It is loaded with flavor and has not lost nutrients due to shipping. In addition, locally produced food is safer than the food produced in other places. You have the power to ensure that the food you buy is free of harmful chemicals for people. Moreover, eating locally produced food has a positive effect on the environment. Transportation of locally grown food does not consume much fossil fuel, which means less pollution. Seeking out locally produced food at your supermarket will save the environment and our health.

The following text is the original of topic 2, quoted from Uenaka and Korechila (2013). Some expressions were changed from the original to control the text length and difficulty. The expressions of underlined are changed in the material.

Grocery stores sell fruits and vegetables from many different places. Have you ever wondered which to choose and why? Here are some good reasons to buy locally produced food.

Firstly, locally produced food is fresher and tastier than the food produced in other regions or countries. Food grown in your community was picked within the past day or two. It is loaded with flavor and has not lost nutrients due to shipping. In addition, locally produced food is safer than the food produced in other countries. You have the power to ensure that the food you buy is free or pesticides, hormones, and antibiotics. Moreover, eating locally produced food has a positive effect on the environment. Transportation of food grown locally does not consume much fossil fuel, which means less pollution. So, seek out locally produced food at your supermarket.

Appendix I. Text of Material 1A-1, 2B-1, 3A-1, and 3B-1 (Topic 3)

Some people think they have an answer to the problems of automobile crowding and air pollution in large cities. Their answer to solve these problems is using bicycles, or “bike.”

In large cities, hundreds of people ride bicycles to work every day. In New York City, some bike riders have even formed a group called Bike for a Better City. They say that if more bicycles are used as a commuter tool, there will be fewer automobiles in the center of the city, and then less dirty air from car engines. For several years, this group has been asking the city government to help bicycle riders. For example, they order the city to place bicycle paths. Using the separate bicycle lanes will decrease traffic accidents. Then, more people will use bikes instead of vehicles. This way, the issue of air pollution and traffic accidents might be changed for better.

The following text is original of topic 3, quoted from Tokyo Academy (2014). Some expressions were changed from the original to control the text length and difficulty. The expressions of underlined are changed in the material.

Some people think they have an answer to the problems of automobile crowding and air pollution in large cities. Their answer is the bicycle, or “bike”.

In a great many cities, hundreds of people ride bicycles to work every day. In New York City, some bike riders have even formed a group called Bike for a Better City. They say that if more people ride bicycles to work, there will be fewer automobiles in the center of the city, and then less dirty air from car engines. For several years this group has been asking the city government to help bicycle riders. For example, they want the city to paint special lanes-- or bicycles only—on some of the main streets because when bicycle riders must use the same lanes as cars, there may be accidents. Bike for a Better City feels if there are special lanes, more people will use bikes.

Appendix J. Text of Material 2B-3 (Topic 4)

The number of old people in Japan will continue to increase. But it will be difficult to build enough homes for older people or find enough professional care workers for everyone. Who is going to help them? The question may not be who, but what. One possibility is care robots. Many older people, however, do not like the idea of a robot taking care of them. In the future, we are going to see many more practical things for elderly people – things like remote-control beds that can become wheelchairs, and high-tech toilets that can monitor a person's health.

The original text is quoted from Hickling and Osaki (2013). Some expressions were changed from the original text to control text length and difficulty, and those of them are underlined.

The number of old people in Japan will continue to grow in the future. But it would be difficult to build enough homes for elderly people or find enough professional care workers for everyone. Who, then, is going to help them? The question may not be who, but what. One possibility is care robots. Many elderly people, however, do not like the idea of a robot taking care of them. In the future, we are going to see many more practical things for elderly people—things like remote-control beds that can become wheelchairs, and high-tech toilets that can monitor a person's health.

Appendix K. Text of Material 1B-4 (Topic 5)

Managing your weight is important to living a healthy life. However, many people have difficulty controlling their weight. Whether a person hopes to lose or gain weight, moderate exercise is very important. This article will introduce weight control through the methods of daily exercise and change of diet.

First, moderate exercise is important to controlling your weight. Except for heavy class martial artists, most athletes do not have extra fat because they work out every day. Their hard training contributes not only to losing extra fat, but also to developing muscles. Muscles are heavier than fat; therefore, some athletes look heavier than they are. However, they are probably healthier than a person who does not exercise. Muscles accelerate energy consumption; accordingly, a person with a well-developed musculature can consume extra fat effectively. Thus, developed muscles are really important to losing weight. Of course, performing a heavy exercise routine may be harmful to your body, especially your joints or your back. You need consider the appropriate exercise intensity before you start a training program.

Second, changing your daily habits is helpful to change your body. If you always use escalators or elevators, your leg muscles may shrink and weaken. While commuting to your school or office, you can do small exercises. You can ride a bicycle instead of driving a car or getting on a train. If you need to ride a train, why don’t you ignore empty seats? You practice balance while standing in a train. If that is easy for you, you can try to stand on tiptoe, changing the exercise intensity. Those tiny exercises can also be useful for someone who wants to gain weight, because they increase the bulk of muscle. Then, your weight will increase, which may arouse your appetite.

To summarize the above two points: do moderate exercise and change your daily habits. However, there is also the question of diet. People who want to gain weight can enjoy eating different foods. However, restricting meals is very difficult for dieters.

A person who wants to lose extra weight should cut high-calorie foods such as fried chicken or French fries. One gram of fat includes nine calories, but carbohydrates and protein include four calories per gram. Thus, if you cut oily foods, you can avoid taking extra calories. Some dieters may want to try fasting. These choices cannot be recommend, because your body and brain refuse such trials. If you fast, you may want to eat too much as a reaction to your hunger, which causes a rebound. The most important factor in restricting your diet is to avoid over-exertion, and cutting high-calorie food. So, why don't you try calorie calculation?

In conclusion, there are three important factors in controlling your weight: taking moderate exercise, changing your daily habits, and cutting high-calorie food. In any case, don't over-exert yourself. Your body can be changed step by step.

Appendix L. Text of Material 1B-3 (Topic 6)

Devices such as smart phones, tablet PCs, and laptops are used around the world. People can use them anytime, anywhere to find information or to communicate with others. Many people benefit from them, but some people, especially older people, say that those items are poison.

Information terminal is essential for modern life. Businesspeople and students use it to make documents. In addition, those items can use to communicate with others using e-mail or other SNS services. Sometimes, people use those devices as a camera, a voice recorder, or a musical instrument. Therefore, it can be said that mobile devices are the all-in-one necessities in the modern world.

On the other hand, some accidents happen because of these items. While people use mobile items, they can get into traffic accidents or fall from platforms. Any child knows to be careful of cars and other dangers outside. However, people of many ages use information terminals while walking in public areas. It can be said that modern people forget common sense using mobile items.

In addition, criminals use the Internet. The main targets are people who don’t understand it. Old people, who have not learned about the Internet, can be caught up in this kind of crime. Then some old people become prejudiced against multimedia equipment.

All people should keep in mind that information terminals is just a device. And it is not your friend, business partner, or servant. When you make an own food for the first time, you probably learn recipes and you may learn how to use cooking tools. Just like this, you need to learn how to use mobile devices in terms of the principal of information literacy. Then, you may be able to avoid cyber-criminals. At the same time, we need to acquire the common sense using information terminals.

Appendix M. Text of Material 1A-2 (Topic 7)

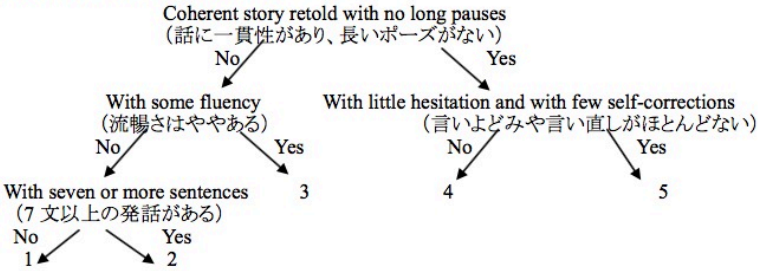
How do you define ‘happiness’? Everyone has a different idea of happiness. Some think earning money is the most important aspect of life, while others may say that living with family is essential for their happiness.

There is a famous proverb, and many people sympathize with it: “Let him that would be happy for a day, go to the barber; for a week, marry a wife; for a month, buy him a new car; for a year, build him a new house; for all his lifetime, be an honest man.”

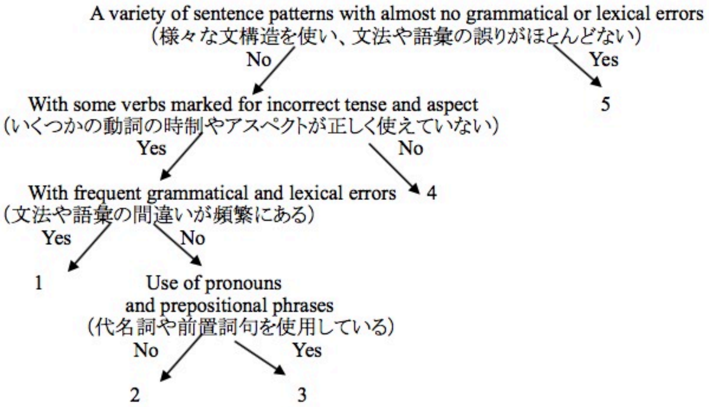
An original of above proverb came from the United Kingdom. It is said that cultures affect ideas of happiness. Although the details of each culture differ from that of the others, everyone can consider what “happiness” means and come to a conclusion. Happiness can be obtained from mental stability. Therefore, the proverb’s conclusion is that being an honest person is important for living happily.

Appendix N. The EBB Scale for the SRST by Hirai & Koizumi (2008)

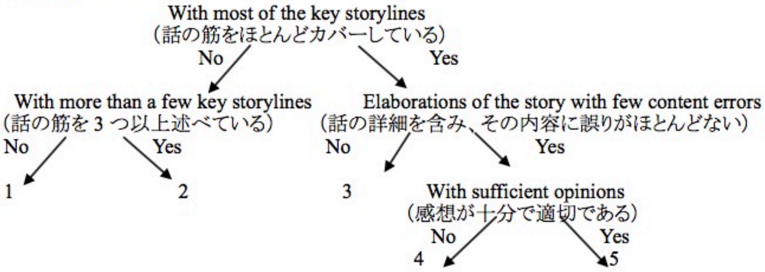
1. Communicative Efficiency（伝達能力）


Coherent story retold with no long pauses
（話に一貫性があり、長いポーズがない）
No
Yes
With some fluency
（流暢さはややある）
No
Yes
With little hesitation and with few self-corrections
（言いよどみや言い直しがほとんどない）
No
Yes
With seven or more sentences
（7 文以上の発話がある）
No
Yes
1
2
3
4
5

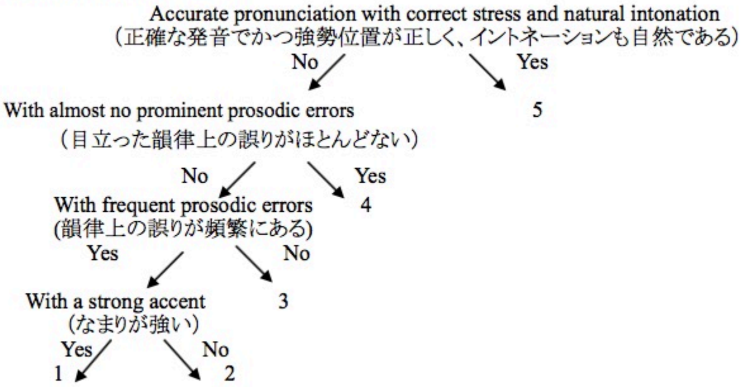
2. Grammar & Vocabulary（文法と語彙）


A variety of sentence patterns with almost no grammatical or lexical errors
（様々な文構造を使い、文法や語彙の誤りがほとんどない）
No
Yes
With some verbs marked for incorrect tense and aspect
（いくつかの動詞の時制やアスペクトが正しく使えていない）
Yes
No
5
With frequent grammatical and lexical errors
（文法や語彙の間違いが頻繁にある）
4
Yes
No
1
Use of pronouns
and prepositional phrases
（代名詞や前置詞句を使用している）
No
Yes
2
3

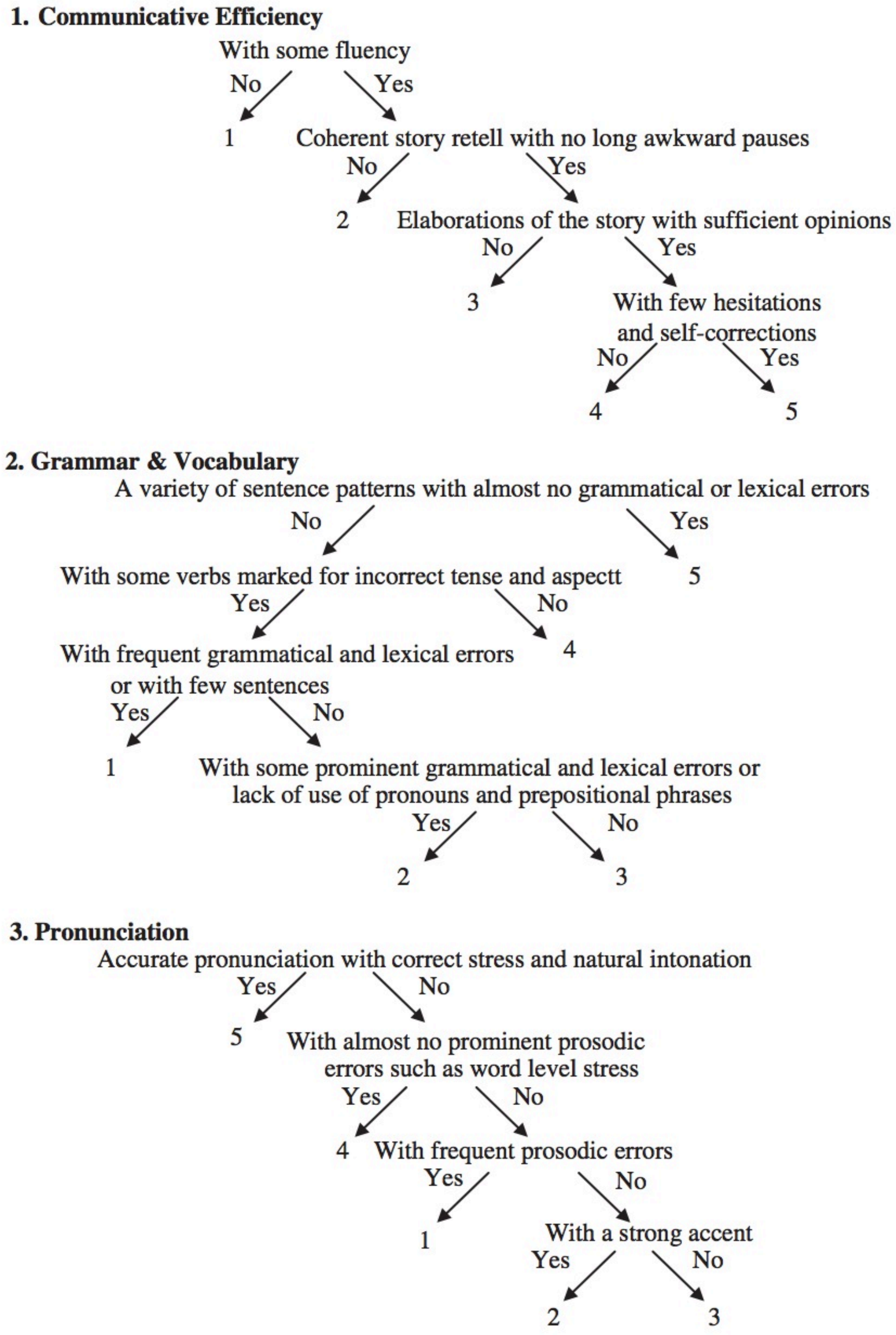
3. Content（内容）


With most of the key storylines
（話の筋をほとんどカバーしている）
No
Yes
With more than a few key storylines
（話の筋を 3 つ以上述べている）
No
Yes
Elaborations of the story with few content errors
（話の詳細を含み、その内容に誤りがほとんどない）
No
Yes
1
2
3
With sufficient opinions
（感想が十分で適切である）
No
Yes
4
5

4. Pronunciation（発音）


Accurate pronunciation with correct stress and natural intonation
（正確な発音でかつ強勢位置が正しく、イントネーションも自然である）
No
Yes
With almost no prominent prosodic errors
5
（目立った韻律上の誤りがほとんどない）
No
Yes
With frequent prosodic errors
(韻律上の誤りが頻繁にある)
4
Yes
No
With a strong accent
3
（なまりが強い）
Yes
No
1
2

Appendix O. The EBB Scale for the SRST by Hirai & Koizumi (2013)


1. Communicative Efficiency
With some fluency
No
Yes
1
Coherent story retell with no long awkward pauses
No
Yes
2
Elaborations of the story with sufficient opinions
No
Yes
3
With few hesitations and self-corrections
No
Yes
4
5
2. Grammar & Vocabulary
A variety of sentence patterns with almost no grammatical or lexical errors
No
Yes
With some verbs marked for incorrect tense and aspectt
5
Yes
No
With frequent grammatical and lexical errors or with few sentences
4
Yes
No
1
With some prominent grammatical and lexical errors or lack of use of pronouns and prepositional phrases
Yes
No
2
3
3. Pronunciation
Accurate pronunciation with correct stress and natural intonation
Yes
No
5
With almost no prominent prosodic errors such as word level stress
Yes
No
4
With frequent prosodic errors
Yes
No
1
With a strong accent
Yes
No
2
3

Appendix P. Note of Caution For Rating

注 1：Hirai and Koizumi (2008)の評価基準 1〜4 の記述を良く読み、発話が条件に適合しているかどうかを考えて評価を出してください。

注 2：発話が極端に少ない、成立しない場合には、評価基準にはありませんが、「0」をつけてください。

注 3：”no long awkward pauses” 及び “with some fluency”の判断基準は一貫性があれば個人の判断に委ねます。

注 4：文法や語彙の誤りを自己修正し、正しく表現できている場合には、その箇所は誤りとみなしません。

注 5：「なまり」とは極端な発音のことで、いわゆる日本人らしい平坦な発音に関してはこれに該当しません。

Appendix Q. Holistic Rating: ACTFL Proficiency Guideline 2012—Speaking (Descriptions were changed and translated by the author)

注：以下に description はありますが、おおまかなものですのである程度直感的に評価してください。

5: (> Intermediate-high) 時制の使い分けをしながら、段落レベルの発話をすることができる。細かなエラーやポーズ、自己修正は見られるが、致命的なエラーはない。

4: (Intermediate-mid) 知っている言葉やインプット文から与えられた表現など組み合わせたり言い換えたりして自分の言葉でやや複雑な文を使って表現することができる。ポーズや言い直し、自己修正はみられるが、総じて表現・文法のエラーの深刻性は低く、非母語話者の発話に慣れていない聞き手であっても発話内容をある程度理解できる程度である

3: (Intermediate-low) インプットとして与えられた文の表現を組み合わせたり言い換えたりして自分の言葉を使って表現する。言いよどみや不正確な表現は多いが、短いポーズ、非効率的な言い直し、自己訂正が多い。表現・文法のエラーにおける深刻度は深くなく、内容の理解に大きなストレスを感じない程度である。

2: (Novice-high) ポーズや言い直しが頻繁に生じ、不完全な文を使うことが多い。正しい文を使う場合にも、現在形を頻繁に利用し（もしくは同じ時制を繰り返す）、1 文ずつが短い。言い直しやポーズが多いため流暢さが低いものの、定型表現に関しては比較的流暢かつ正確に発話することができる。（聞き手が）発話の内容を正しく理解できないこともあるが、概ね意味は通じる。

1: (Novice-mid) 簡単な語彙を探そうとしてポーズや繰り返しが非常に多くなる。表現・文法のエラーが多いため（聞き手が）発話の内容理解に困難を感じ、（無意味に）母語の単語を使用したり、黙りこんでしまう。

0: (Novice-low) 発話がほとんどない、もしくは全く関係ないことを話している。

Appendix R. Participants Measurement Report in Three-Faceted Rasch Measurement: Participants × Rater × Topics

Total Score	Total Count	Observed Average	Fair(M) Average	Measure	Model S.E.	Infit MNSQ	Infit Zstd	Outfit MNSQ	Outfit Zstd	Estim. Discrm	Corr. PtBis	Participant
376	16	23.50	23.58	1.54	0.15	1.13	0.4	1.15	0.4	0.95	.00	9
264	12	22.00	22.58	1.25	0.12	0.46	-1.6	0.45	-1.6	1.41	.17	74
302	16	18.88	18.89	0.69	0.09	1.03	0.1	0.99	0.0	1.02	.47	5
286	16	17.88	17.87	0.54	0.10	1.01	0.1	1.01	0.1	0.95	.37	4
273	16	17.06	17.05	0.41	0.10	1.16	0.5	1.18	0.5	0.66	.05	6
268	16	16.75	16.74	0.36	0.10	1.14	0.4	1.23	0.6	0.86	-.03	49
191	12	15.92	16.48	0.31	0.13	0.54	-1.1	0.55	-1.0	1.12	.04	72
260	16	16.25	16.25	0.27	0.11	0.29	-2.5	0.30	-2.5	1.64	.45	41
258	16	16.13	16.12	0.25	0.11	0.53	-1.4	0.53	-1.3	1.57	.32	43
257	16	16.06	15.95	0.22	0.11	0.70	-0.7	0.67	-0.8	1.05	-.24	123
254	16	15.88	15.88	0.20	0.11	0.84	-0.3	0.85	-0.2	1.11	.13	40
253	16	15.81	15.82	0.19	0.11	1.01	0.1	1.01	0.1	0.91	.16	1
120	8	15.00	15.63	0.16	0.16	0.51	-0.9	0.50	-0.9	1.46	-.28	55
250	16	15.63	15.63	0.16	0.11	0.49	-1.5	0.50	-1.4	1.36	.35	52
245	16	15.31	15.32	0.10	0.11	1.01	0.1	1.01	0.1	0.93	-.27	24
212	14	15.14	15.06	0.04	0.12	0.18	-3.0	0.18	-3.0	1.75	.33	53
241	16	15.06	14.97	0.02	0.11	2.97	3.5	2.91	3.4	-0.32	.41	113
171	12	14.25	14.75	-0.02	0.13	0.63	-0.8	0.64	-0.8	1.30	.03	84
233	16	14.56	14.59	-0.05	0.11	0.47	-1.6	0.47	-1.6	1.42	.53	33

(Table continues)

Appendix R—Continued

166	12	13.83	14.34	-0.11	0.13	1.27	0.7	1.27	0.7	0.82	.14	91
109	8	13.63	14.24	-0.13	0.16	0.30	-1.6	0.30	-1.6	1.35	-.39	54
226	16	14.13	14.16	-0.14	0.11	0.81	-0.4	0.81	-0.4	1.20	.11	17
108	8	13.50	14.12	-0.15	0.16	0.21	-2.1	0.21	-2.1	1.52	-.27	64
162	12	13.50	14.01	-0.18	0.13	3.62	3.8	3.62	3.8	-1.18	.27	75
107	8	13.38	13.99	-0.18	0.16	0.20	-2.2	0.20	-2.2	1.62	-.38	56
107	8	13.38	13.99	-0.18	0.16	1.63	1.1	1.64	1.1	0.42	-.30	59
223	16	13.94	13.97	-0.18	0.11	2.02	2.2	2.02	2.2	0.12	.56	31
221	16	13.81	13.85	-0.21	0.11	0.94	0.0	0.92	-0.1	1.14	.43	36
221	16	13.81	13.85	-0.21	0.11	0.61	-1.0	0.62	-1.0	1.54	.23	50
220	16	13.75	13.79	-0.22	0.11	0.90	-0.1	0.91	-0.1	1.13	-.02	12
105	8	13.13	13.75	-0.23	0.16	0.42	-1.2	0.42	-1.2	1.59	.11	70
219	16	13.69	13.72	-0.23	0.11	0.86	-0.2	0.86	-0.2	1.06	.07	11
217	16	13.56	13.60	-0.26	0.11	0.50	-1.5	0.50	-1.5	1.49	.46	28
218	16	13.63	13.55	-0.27	0.11	0.43	-1.8	0.42	-1.8	1.51	-.19	94
216	16	13.50	13.54	-0.27	0.11	0.90	-0.1	0.91	-0.1	1.12	.08	16
216	16	13.50	13.54	-0.27	0.11	1.55	1.3	1.55	1.3	0.42	-.26	23
156	12	13.00	13.52	-0.28	0.13	0.53	-1.1	0.54	-1.1	1.10	-.13	90
103	8	12.88	13.51	-0.28	0.16	0.33	-1.5	0.33	-1.5	1.48	-.13	65
215	16	13.44	13.48	-0.29	0.11	0.89	-0.1	0.90	-0.1	1.10	-.18	13
101	8	12.63	13.27	-0.33	0.16	0.91	0.0	0.90	0.0	1.31	.15	61
211	16	13.19	13.23	-0.34	0.11	1.05	0.2	1.06	0.2	1.06	-.08	8

(Table continues)

Appendix R—Continued

211	16	13.19	13.23	-0.34	0.11	1.23	0.6	1.26	0.7	0.74	-.20	14
211	16	13.19	13.23	-0.34	0.11	0.86	-0.2	0.87	-0.2	0.94	.09	15
209	16	13.06	13.11	-0.36	0.11	5.02	5.8	5.10	5.9	-1.70	.10	3
209	16	13.06	13.11	-0.36	0.11	1.11	0.3	1.11	0.4	0.83	.18	48
98	8	12.25	12.91	-0.40	0.15	0.29	-1.7	0.29	-1.7	1.57	.47	69
204	16	12.75	12.80	-0.42	0.11	0.63	-1.0	0.64	-1.0	1.41	.28	26
204	16	12.75	12.80	-0.42	0.11	0.67	-0.8	0.64	-0.9	1.26	.07	42
147	12	12.25	12.80	-0.42	0.13	0.32	-2.0	0.32	-2.0	1.50	-.24	86
203	16	12.69	12.74	-0.44	0.11	1.62	1.5	1.66	1.6	0.61	-.25	22
202	16	12.63	12.68	-0.45	0.11	1.31	0.8	1.32	0.9	0.76	.07	25
144	12	12.00	12.56	-0.47	0.12	0.70	-0.6	0.71	-0.6	1.26	-.21	78
95	8	11.88	12.55	-0.47	0.15	0.28	-1.7	0.28	-1.7	1.85	-.42	63
199	16	12.44	12.49	-0.48	0.11	1.20	0.6	1.20	0.6	0.88	.29	27
197	16	12.31	12.37	-0.51	0.11	0.16	-3.5	0.16	-3.5	1.74	.39	38
199	16	12.44	12.37	-0.51	0.11	0.60	-1.1	0.61	-1.1	1.57	-.35	124
93	8	11.63	12.32	-0.52	0.15	0.81	-0.2	0.83	-0.1	0.97	-.14	67
198	16	12.38	12.30	-0.52	0.11	0.93	0.0	0.93	0.0	1.12	-.52	128
194	16	12.13	12.18	-0.54	0.11	1.17	0.5	1.22	0.6	0.76	-.18	20
194	16	12.13	12.18	-0.54	0.11	2.48	2.9	2.54	3.0	-0.57	.57	35
91	8	11.38	12.08	-0.56	0.15	0.74	-0.3	0.74	-0.3	0.90	-.32	62
192	16	12.00	12.06	-0.57	0.11	1.23	0.6	1.21	0.6	0.51	-.05	21
190	16	11.88	11.94	-0.59	0.11	0.92	0.0	0.91	-0.1	0.91	-.01	19

(Table continues)

Appendix R—Continued

136	12	11.33	11.92	-0.59	0.12	0.68	-0.7	0.65	-0.8	1.48	-.07	85
189	16	11.81	11.88	-0.60	0.11	0.77	-0.5	0.77	-0.5	1.30	.43	29
189	16	11.81	11.88	-0.60	0.11	0.75	-0.6	0.75	-0.6	1.19	.33	32
188	16	11.75	11.81	-0.61	0.11	2.49	3.0	2.41	2.9	0.11	.16	2
186	16	11.63	11.69	-0.64	0.11	2.05	2.3	2.05	2.3	-0.32	.49	30
186	16	11.63	11.69	-0.64	0.11	1.16	0.5	1.24	0.7	0.40	.51	47
133	12	11.08	11.69	-0.64	0.12	1.84	1.7	1.83	1.7	0.07	.05	77
185	16	11.56	11.63	-0.65	0.11	0.87	-0.2	0.92	-0.1	0.93	.02	18
86	8	10.75	11.50	-0.67	0.14	0.55	-0.9	0.56	-0.9	1.18	.48	71
184	16	11.50	11.43	-0.68	0.11	0.36	-2.2	0.35	-2.2	1.51	-.06	121
183	16	11.44	11.37	-0.69	0.11	0.43	-1.9	0.43	-1.9	1.43	.33	102
84	8	10.50	11.27	-0.71	0.14	0.15	-2.8	0.15	-2.7	1.84	-.08	68
179	16	11.19	11.26	-0.71	0.10	0.28	-2.7	0.28	-2.7	1.66	.45	39
126	12	10.50	11.14	-0.73	0.12	0.34	-2.1	0.34	-2.1	1.69	.22	73
175	16	10.94	11.01	-0.76	0.10	3.80	4.9	3.90	5.0	-1.61	-.50	44
177	16	11.06	10.99	-0.76	0.10	0.85	-0.3	0.85	-0.3	1.27	-.33	126
173	16	10.81	10.88	-0.78	0.10	0.97	0.0	0.98	0.0	1.03	.11	10
172	16	10.75	10.82	-0.79	0.10	0.41	-2.1	0.41	-2.1	1.45	.53	37
80	8	10.00	10.80	-0.79	0.14	0.16	-2.9	0.16	-2.8	1.87	-.29	58
170	16	10.63	10.70	-0.81	0.10	1.25	0.7	1.23	0.7	0.92	.17	7
170	16	10.63	10.70	-0.81	0.10	1.19	0.6	1.19	0.6	0.63	.55	34
79	8	9.88	10.68	-0.81	0.14	0.67	-0.7	0.67	-0.7	1.53	-.44	57

(Table continues)

Appendix R—Continued

120	12	10.00	10.66	-0.81	0.12	0.34	-2.3	0.34	-2.3	1.28	.00	87
169	16	10.56	10.63	-0.82	0.10	0.65	-1.0	0.64	-1.1	1.46	.43	51
171	16	10.69	10.61	-0.82	0.10	0.43	-2.0	0.41	-2.1	2.07	.25	106
171	16	10.69	10.61	-0.82	0.10	0.25	-3.1	0.25	-3.1	1.78	.21	112
170	16	10.63	10.55	-0.83	0.10	0.99	0.0	0.98	0.0	1.20	.51	98
168	16	10.50	10.42	-0.85	0.10	0.55	-1.5	0.55	-1.4	1.15	-.13	96
164	16	10.25	10.17	-0.89	0.10	0.77	-0.6	0.75	-0.7	1.48	.32	100
163	16	10.19	10.10	-0.90	0.10	0.67	-1.0	0.68	-1.0	1.39	.03	122
162	16	10.13	10.04	-0.91	0.10	0.52	-1.7	0.51	-1.7	2.01	.21	97
162	16	10.13	10.04	-0.91	0.10	1.42	1.2	1.42	1.2	0.43	.46	111
160	16	10.00	9.91	-0.93	0.10	1.20	0.6	1.18	0.6	0.61	-.49	119
72	8	9.00	9.82	-0.95	0.14	0.55	-1.1	0.55	-1.2	1.56	.22	66
156	16	9.75	9.81	-0.95	0.10	2.01	2.6	2.00	2.5	-0.21	.18	45
100	11	9.09	9.73	-0.96	0.12	2.87	3.7	2.86	3.6	-0.63	-.20	83
156	16	9.75	9.65	-0.97	0.10	0.56	-1.5	0.57	-1.5	1.11	.22	108
156	16	9.75	9.65	-0.97	0.10	1.20	0.7	1.19	0.6	1.20	.44	110
154	16	9.63	9.53	-0.99	0.10	1.29	0.9	1.30	0.9	0.90	.53	107
152	16	9.50	9.40	-1.01	0.10	1.20	0.7	1.21	0.7	0.49	-.32	127
151	16	9.44	9.33	-1.02	0.10	0.83	-0.4	0.84	-0.4	1.10	.20	95
147	16	9.19	9.07	-1.06	0.10	0.62	-1.3	0.62	-1.3	1.09	-.17	114
145	16	9.06	8.95	-1.08	0.10	0.86	-0.4	0.85	-0.4	1.89	-.29	125
141	16	8.81	8.83	-1.10	0.10	1.38	1.2	1.39	1.2	0.98	.46	46

(Table contines)

Appendix R—Continued

142	16	8.88	8.75	-1.11	0.10	0.85	-0.4	0.85	-0.4	0.43	-.42	117
140	16	8.75	8.62	-1.13	0.10	0.76	-0.8	0.75	-0.8	1.54	.31	109
140	16	8.75	8.62	-1.13	0.10	1.16	0.6	1.16	0.6	0.10	-.37	116
93	12	7.75	8.34	-1.18	0.12	1.05	0.2	1.01	0.1	1.19	-.01	76
93	12	7.75	8.34	-1.18	0.12	0.34	-2.4	0.35	-2.3	0.87	.18	80
133	16	8.31	8.18	-1.21	0.10	1.46	1.4	1.49	1.5	0.56	-.49	120
132	16	8.25	8.11	-1.22	0.10	0.81	-0.5	0.81	-0.5	1.46	.10	93
130	16	8.13	7.99	-1.24	0.10	0.98	0.0	0.92	-0.1	0.65	-.39	118
127	16	7.94	7.80	-1.27	0.11	1.28	0.9	1.29	0.9	1.37	.34	92
87	12	7.25	7.76	-1.28	0.13	1.54	1.3	1.68	1.5	0.89	.15	89
86	12	7.17	7.67	-1.30	0.13	1.04	0.2	1.01	0.1	1.11	-.03	88
42	6	7.00	7.66	-1.30	0.20	1.72	1.2	1.70	1.1	1.28	-.40	82
117	16	7.31	7.17	-1.39	0.11	0.53	-1.6	0.54	-1.5	1.32	-.01	105
115	16	7.19	7.05	-1.42	0.12	0.46	-1.8	0.50	-1.6	1.04	.31	104
111	16	6.94	6.81	-1.48	0.12	0.82	-0.4	0.80	-0.4	1.28	.22	99
51	8	6.38	6.80	-1.48	0.20	0.78	-0.1	0.74	-0.2	1.05	.49	60
48	8	6.00	6.31	-1.62	0.24	0.93	0.1	0.99	0.2	1.08	-.46	81
101	16	6.31	6.22	-1.65	0.15	0.82	-0.2	0.70	-0.5	0.97	-.05	115
71	12	5.92	6.13	-1.68	0.20	0.29	-1.3	0.31	-1.2	0.87	.07	79
99	16	6.19	6.10	-1.70	0.15	0.95	0.0	1.00	0.1	1.06	.33	101
92	16	5.75	5.70	-1.91	0.20	1.20	0.5	1.04	0.2	1.11	.41	103

(Table continues)

Appendix R—Continued

166.7	14.1	11.72	11.89	-0.57	0.12	0.97	-0.3	0.97	-0.3		.07	Mean (Count: 128)
58.3	3	3.04	3.06	0.56	0.02	0.73	1.6	0.73	1.6		.30	S.D. (Popuation)
58.6	3	3.06	3.08	0.56	0.02	0.73	1.6	0.74	1.6		.30	S.D. (Sample)

Model, Populn: RMSE .12   Adj (True) S.D. .55   Separation 4.48   Strata 6.30   Reliability .95

Model, Sample: RMSE .12   Adj (True) S.D. .55   Separation 4.50   Strata 6.33   Reliability .95

Model, Fixed (all same) chi-square: 2710.7   d.f.: 127   significance (probability): .00

Model,   Random (normal) chi-square: 120.1   d.f.: 126   significance (probability): .63

Appendix S. Unexpected Responses in Three-Faceted Rasch Measurement: Participants × Rater × Topics

Score	Exp.	Resd	StRes	Participant	Rater	Topic
16	10.8	5.2	2.1	77	3	Topic7
9	13.6	-4.6	-2.1	90	2	Topic7
5	10.5	-5.5	-2.2	97	1	Topic6
13	7.8	5.2	2.2	116	3	Topic5
5	10.2	-5.2	-2.1	111	2	Topic5
12	7.4	4.6	2.1	120	3	Topic5
10	14.4	-4.4	-2.0	113	1	Topic5
5	12.3	-7.3	-3.2	44	2	Topic4
7	13.9	-6.9	-3.1	22	4	Topic4
21	14.3	6.7	3.1	31	3	Topic4
7	13.5	-6.5	-3.0	8	3	Topic4
19	12.5	6.5	2.9	35	3	Topic4
5	11.5	-6.5	-2.8	44	1	Topic4
7	13.0	-6.0	-2.7	22	3	Topic4
7	13.0	-6.0	-2.7	25	3	Topic4
5	11.4	-6.4	-2.7	44	3	Topic4
6	12.3	-6.3	-2.7	44	4	Topic4
19	12.9	6.1	2.7	47	4	Topic4
19	13.4	5.6	2.6	35	2	Topic4
19	13.3	5.7	2.6	35	4	Topic4
8	13.3	-5.3	-2.4	20	4	Topic4
18	12.8	5.2	2.3	27	3	Topic4
11	16.1	-5.1	-2.2	1	3	Topic4
12	17.4	-5.4	-2.2	6	3	Topic4
20	15.1	4.9	2.2	31	4	Topic4
10	14.6	-4.6	-2.1	13	4	Topic4
10	14.6	-4.6	-2.1	16	4	Topic4
10	14.6	-4.6	-2.1	23	4	Topic4
21	13.6	7.4	3.4	75	4	Topic3
19	12.8	6.2	2.8	75	3	Topic3
5	11.4	-6.4	-2.7	2	2	Topic3

(Table continues)

Appendix S—Continued

5	11.4	-6.4	-2.7	2	4	Topic3
12	6.7	5.3	2.7	89	1	Topic3
17	10.5	6.5	2.6	44	4	Topic3
7	12.6	-5.6	-2.5	36	3	Topic3
17	11.3	5.7	2.4	30	2	Topic3
6	11.5	-5.5	-2.3	42	3	Topic3
5	10.5	-5.5	-2.2	2	1	Topic3
5	10.4	-5.4	-2.2	2	3	Topic3
5	10.4	-5.4	-2.2	77	1	Topic3
5	10.2	-5.2	-2.1	7	2	Topic3
5	10.2	-5.2	-2.1	7	4	Topic3
15	9.6	5.4	2.1	44	1	Topic3
17	12.4	4.6	2.0	23	1	Topic3
7	11.7	-4.7	-2.0	35	4	Topic3
8	13.9	-5.9	-2.7	75	2	Topic2
15	8.5	6.5	2.6	83	3	Topic2
18	12.3	5.7	2.5	3	1	Topic2
21	15.7	5.3	2.3	49	3	Topic2
8	13.1	-5.1	-2.3	75	3	Topic2
22	16.6	5.4	2.2	49	4	Topic2
9	13.9	-4.9	-2.2	75	4	Topic2
9	5.9	3.1	2.2	81	1	Topic2
19	14.4	4.6	2.1	24	3	Topic2
11	15.8	-4.8	-2.1	40	2	Topic2
15	9.7	5.3	2.1	45	2	Topic2
11	6.8	4.2	2.1	82	3	Topic2
15	9.6	5.4	2.1	83	2	Topic2
15	9.6	5.4	2.1	83	4	Topic2
5	13.9	-8.9	-4.0	3	2	Topic1
5	13.9	-8.9	-4.0	3	4	Topic1
5	13.2	-8.2	-3.7	3	1	Topic1
5	13.0	-8.0	-3.6	3	3	Topic1
23	15.8	7.2	3.2	113	2	Topic1
22	14.9	7.1	3.2	113	3	Topic1

(Table continues)

Appendix S—Continued

23	15.7	7.3	3.2	113	4	Topic1
11	6.1	4.9	3.1	101	1	Topic1
10	6.0	4.0	2.7	103	4	Topic1
18	12.5	5.5	2.4	18	2	Topic1
6	11.6	-5.6	-2.4	30	3	Topic1
5	10.8	-5.8	-2.4	45	2	Topic1
18	13.0	5.0	2.3	20	2	Topic1
18	12.9	5.1	2.3	21	2	Topic1
25	19.0	6.0	2.2	4	2	Topic1
21	23.9	-2.9	-2.2	9	4	Topic1
8	12.9	-4.9	-2.2	26	1	Topic1
9	13.9	-4.9	-2.2	48	4	Topic1
13	7.9	5.1	2.2	92	1	Topic1
16	11.0	5.0	2.1	111	4	Topic1
5	10.2	-5.2	-2.1	122	1	Topic1

Appendix T. Participant Measurement in the Three-faceted Rasch Measurement: Participants × Raters × Tasks (Length × Mode)

Total	Total	Observed	Fair(M)		Model	Infit	Infit	Outfit	Outfit	Estim.	Corr.	
Score	Count	Average	Average	Measure	S.E.	MNSQ	Zstd	MNSQ	Zstd	Discrm	PtBis	Participant
376	16	23.50	23.69	1.59	0.14	1.18	0.5	1.60	1.1	0.81	.10	9
84	4	21.00	20.61	1.00	0.19	0.22	-1.8	0.21	-1.9	1.89	.13	74
302	16	18.88	19.04	0.79	0.10	1.65	1.7	1.61	1.6	0.34	.28	5
286	16	17.88	17.99	0.64	0.10	1.11	0.4	1.09	0.3	0.82	.16	4
273	16	17.06	17.16	0.50	0.11	1.30	0.8	1.28	0.8	0.61	.02	6
268	16	16.75	16.85	0.44	0.11	1.52	1.3	1.63	1.5	0.49	-.31	49
260	16	16.25	16.35	0.35	0.11	0.19	-3.2	0.20	-3.1	1.72	.38	41
258	16	16.13	16.23	0.33	0.11	0.66	-0.8	0.68	-0.8	1.44	.19	43
254	16	15.88	15.98	0.28	0.11	0.91	-0.1	0.92	0.0	1.05	-.01	40
65	4	16.25	15.96	0.27	0.22	0.59	-0.3	0.61	-0.3	0.99	.28	72
253	16	15.81	15.92	0.26	0.11	0.88	-0.1	0.90	-0.1	1.02	.05	1
250	16	15.63	15.73	0.22	0.11	0.80	-0.4	0.83	-0.3	1.11	.22	52
64	4	16.00	15.72	0.22	0.23	0.28	-1.1	0.28	-1.1	1.68	-.13	84
257	16	16.06	15.72	0.22	0.11	0.91	-0.1	0.86	-0.2	0.89	-.28	123
245	16	15.31	15.43	0.16	0.12	0.86	-0.2	0.88	-0.1	1.05	.29	24
59	4	14.75	15.27	0.12	0.24	0.04	-2.2	0.05	-2.2	2.10	.25	55
62	4	15.50	15.24	0.12	0.23	0.11	-1.7	0.11	-1.7	1.83	-.12	77
212	14	15.14	15.23	0.12	0.12	0.25	-2.5	0.23	-2.7	1.71	.14	53
241	16	15.06	14.76	0.01	0.12	2.91	3.4	2.80	3.3	-0.26	.48	113

(Table continues)

Appendix T—Continues

233	16	14.56	14.69	-0.01	0.12	0.40	-1.9	0.40	-1.9	1.49	.19	33
56	4	14.00	14.52	-0.04	0.24	0.92	0.1	0.92	0.1	1.47	-.40	61
56	4	14.00	14.52	-0.04	0.24	0.04	-2.4	0.04	-2.4	1.84	.00	70
226	16	14.13	14.26	-0.10	0.12	0.65	-0.9	0.64	-0.9	1.34	-.04	17
223	16	13.94	14.08	-0.14	0.12	1.92	2.0	1.91	2.0	0.18	.23	31
221	16	13.81	13.95	-0.17	0.12	1.09	0.3	1.08	0.3	1.01	.26	36
221	16	13.81	13.95	-0.17	0.12	0.55	-1.3	0.56	-1.3	1.59	.25	50
220	16	13.75	13.89	-0.19	0.12	0.80	-0.4	0.80	-0.4	1.22	.00	12
219	16	13.69	13.83	-0.20	0.12	0.84	-0.3	0.84	-0.3	1.07	.33	11
53	4	13.25	13.78	-0.21	0.23	0.03	-2.5	0.03	-2.5	1.39	.36	64
53	4	13.25	13.78	-0.21	0.23	0.03	-2.5	0.03	-2.5	1.39	.36	69
56	4	14.00	13.76	-0.22	0.24	0.49	-0.5	0.50	-0.5	1.66	-.36	75
56	4	14.00	13.76	-0.22	0.24	0.32	-1.0	0.32	-1.0	1.81	.13	78
217	16	13.56	13.71	-0.23	0.12	0.49	-1.5	0.49	-1.5	1.51	.14	28
216	16	13.50	13.65	-0.24	0.12	0.95	0.0	0.95	0.0	1.06	.04	16
216	16	13.50	13.65	-0.24	0.12	1.19	0.5	1.18	0.5	0.73	-.11	23
215	16	13.44	13.58	-0.25	0.12	0.57	-1.2	0.57	-1.2	1.37	-.02	13
52	4	13.00	13.53	-0.27	0.23	0.16	-1.6	0.16	-1.5	1.46	-.06	54
218	16	13.63	13.35	-0.31	0.12	0.45	-1.7	0.44	-1.7	1.49	-.09	94
211	16	13.19	13.34	-0.31	0.12	0.92	0.0	0.91	-0.1	1.18	-.04	8
211	16	13.19	13.34	-0.31	0.12	0.81	-0.4	0.82	-0.3	1.11	.11	14
211	16	13.19	13.34	-0.31	0.12	0.74	-0.6	0.73	-0.6	1.07	.37	15

(Table continues)

Appendix T—Continues

51	4	12.75	13.29	-0.32	0.23	0.13	-1.7	0.13	-1.7	1.33	-.38	65
54	4	13.50	13.26	-0.33	0.24	0.14	-1.7	0.13	-1.7	1.79	.27	86
209	16	13.06	13.22	-0.34	0.12	6.42	7.0	6.59	7.1	-2.91	-.14	3
209	16	13.06	13.22	-0.34	0.12	0.97	0.0	0.97	0.0	0.93	.27	48
50	4	12.50	13.05	-0.37	0.23	0.09	-1.9	0.09	-1.9	1.55	.02	56
204	16	12.75	12.91	-0.40	0.12	0.94	0.0	0.97	0.0	1.15	.15	26
204	16	12.75	12.91	-0.40	0.12	0.92	0.0	0.89	-0.1	1.08	-.12	42
203	16	12.69	12.85	-0.42	0.11	1.48	1.2	1.47	1.2	0.74	.10	22
49	4	12.25	12.81	-0.43	0.23	0.17	-1.5	0.18	-1.5	1.88	.32	71
202	16	12.63	12.79	-0.43	0.11	1.14	0.4	1.13	0.4	0.93	.41	25
199	16	12.44	12.61	-0.47	0.11	1.07	0.3	1.09	0.3	0.96	.30	27
197	16	12.31	12.48	-0.49	0.11	0.12	-3.9	0.12	-3.9	1.78	.17	38
47	4	11.75	12.33	-0.53	0.22	0.13	-1.7	0.13	-1.7	1.98	-.39	63
194	16	12.13	12.30	-0.53	0.11	0.91	-0.1	0.93	0.0	1.05	.17	20
194	16	12.13	12.30	-0.53	0.11	2.60	3.1	2.69	3.2	-0.64	.32	35
50	4	12.50	12.24	-0.55	0.23	0.11	-1.8	0.11	-1.8	1.53	-.12	85
192	16	12.00	12.18	-0.56	0.11	1.08	0.3	1.06	0.2	0.68	.32	21
199	16	12.44	12.17	-0.56	0.11	0.76	-0.5	0.77	-0.5	1.43	-.36	124
198	16	12.38	12.10	-0.57	0.11	1.11	0.4	1.13	0.4	0.93	-.52	128
190	16	11.88	12.06	-0.58	0.11	0.83	-0.3	0.82	-0.3	1.01	.26	19
189	16	11.81	12.00	-0.60	0.11	1.00	0.1	1.01	0.1	1.10	.16	29
189	16	11.81	12.00	-0.60	0.11	0.70	-0.8	0.69	-0.8	1.25	.27	32

(Table continues)

Appendix T—Continued

188	16	11.75	11.94	-0.61	0.11	2.76	3.4	2.70	3.3	-0.13	.58	2
186	16	11.63	11.82	-0.63	0.11	1.78	1.8	1.78	1.8	0.02	.38	30
186	16	11.63	11.82	-0.63	0.11	0.90	-0.1	0.97	0.0	0.64	.31	47
185	16	11.56	11.76	-0.64	0.11	0.60	-1.1	0.65	-0.9	1.19	.28	18
48	4	12.00	11.74	-0.65	0.22	0.04	-2.3	0.04	-2.3	1.97	.00	87
48	4	12.00	11.74	-0.65	0.22	1.03	0.2	1.05	0.3	0.67	-.57	90
179	16	11.19	11.39	-0.72	0.11	0.29	-2.7	0.27	-2.8	1.65	.18	39
46	4	11.50	11.22	-0.75	0.22	0.00	-3.3	0.00	-3.3	2.19	.64	73
184	16	11.50	11.22	-0.75	0.11	0.59	-1.2	0.58	-1.2	1.27	-.18	121
42	4	10.50	11.16	-0.76	0.21	0.00	-3.6	0.00	-3.6	2.15	.58	59
183	16	11.44	11.15	-0.76	0.11	0.48	-1.7	0.47	-1.7	1.40	.34	102
175	16	10.94	11.15	-0.76	0.11	4.06	5.2	4.33	5.5	-1.86	-.21	44
173	16	10.81	11.03	-0.78	0.11	0.90	-0.2	0.91	-0.1	1.14	.34	10
172	16	10.75	10.97	-0.80	0.11	0.31	-2.7	0.31	-2.7	1.59	.29	37
41	4	10.25	10.93	-0.80	0.21	0.03	-2.9	0.03	-2.8	1.94	.37	66
170	16	10.63	10.84	-0.82	0.11	1.33	1.0	1.32	0.9	0.82	.46	7
170	16	10.63	10.84	-0.82	0.11	0.98	0.0	0.98	0.0	0.92	.39	34
169	16	10.56	10.78	-0.83	0.11	0.68	-0.9	0.67	-1.0	1.44	.52	51
177	16	11.06	10.77	-0.83	0.11	1.00	0.1	1.01	0.1	1.11	-.31	126
43	4	10.75	10.45	-0.89	0.21	0.03	-2.7	0.03	-2.7	1.98	.59	91
39	4	9.75	10.45	-0.89	0.20	0.18	-1.9	0.18	-1.8	1.16	.36	62
171	16	10.69	10.38	-0.90	0.11	0.58	-1.3	0.54	-1.5	1.87	.22	106

(Table continues)

Appendix T—Continued

171	16	10.69	10.38	-0.90	0.11	0.30	-2.8	0.29	-2.8	1.72	.24	112
170	16	10.63	10.31	-0.91	0.11	0.96	0.0	0.96	0.0	1.23	.56	98
38	4	9.50	10.21	-0.93	0.20	0.07	-2.6	0.07	-2.6	1.39	.37	67
38	4	9.50	10.21	-0.93	0.20	0.07	-2.6	0.07	-2.6	1.39	.37	68
168	16	10.50	10.18	-0.93	0.11	0.63	-1.1	0.64	-1.0	1.06	-.11	96
156	16	9.75	9.96	-0.97	0.10	1.81	2.2	1.78	2.1	0.09	-.28	45
37	4	9.25	9.96	-0.97	0.20	0.21	-1.8	0.21	-1.8	1.20	-.20	58
41	4	10.25	9.94	-0.97	0.21	0.03	-2.9	0.03	-2.8	1.94	.41	88
164	16	10.25	9.92	-0.98	0.10	0.98	0.0	0.95	0.0	1.26	.27	100
163	16	10.19	9.85	-0.99	0.10	0.90	-0.2	0.92	-0.1	1.10	-.02	122
162	16	10.13	9.78	-1.00	0.10	0.51	-1.7	0.50	-1.7	1.96	.27	97
162	16	10.13	9.78	-1.00	0.10	1.36	1.1	1.37	1.1	0.57	.52	111
160	16	10.00	9.65	-1.02	0.10	1.33	1.0	1.32	1.0	0.52	-.44	119
156	16	9.75	9.38	-1.06	0.10	0.80	-0.5	0.82	-0.4	0.79	.14	108
156	16	9.75	9.38	-1.06	0.10	1.25	0.8	1.23	0.7	1.11	.46	110
154	16	9.63	9.25	-1.09	0.10	1.37	1.1	1.38	1.1	0.78	.55	107
152	16	9.50	9.12	-1.11	0.10	1.65	1.8	1.70	1.9	0.02	-.41	127
151	16	9.44	9.05	-1.12	0.10	0.85	-0.3	0.88	-0.3	1.11	.27	95
141	16	8.81	8.98	-1.13	0.10	1.09	0.3	1.09	0.3	1.36	.40	46
147	16	9.19	8.79	-1.16	0.10	0.77	-0.7	0.77	-0.7	0.91	-.16	114
145	16	9.06	8.66	-1.18	0.10	0.88	-0.3	0.86	-0.3	1.81	-.21	125
32	4	8.00	8.65	-1.19	0.21	0.39	-1.1	0.40	-1.1	1.31	-.21	57

(Table continues)

Appendix T—Continued

142	16	8.88	8.46	-1.22	0.10	0.95	0.0	0.94	0.0	0.45	-.36	117
31	4	7.75	8.37	-1.23	0.22	0.03	-3.1	0.03	-3.1	1.38	.55	60
140	16	8.75	8.33	-1.24	0.11	0.85	-0.4	0.82	-0.5	1.42	.31	109
140	16	8.75	8.33	-1.24	0.11	1.32	1.0	1.29	0.9	0.12	-.32	116
133	16	8.31	7.89	-1.32	0.11	1.71	2.0	1.69	1.9	0.36	-.46	120
132	16	8.25	7.83	-1.33	0.11	0.75	-0.7	0.74	-0.8	1.58	.17	93
130	16	8.13	7.70	-1.35	0.11	1.19	0.6	1.08	0.3	0.55	-.38	118
127	16	7.94	7.52	-1.39	0.11	1.13	0.4	1.10	0.4	1.50	.44	92
30	4	7.50	7.21	-1.45	0.23	0.30	-1.3	0.29	-1.3	0.36	-.45	80
117	16	7.31	6.93	-1.52	0.12	0.48	-1.7	0.50	-1.6	1.40	.09	105
115	16	7.19	6.82	-1.55	0.12	0.58	-1.3	0.70	-0.7	0.99	.26	104
111	16	6.94	6.60	-1.61	0.13	0.77	-0.5	0.76	-0.5	1.34	.29	99
101	16	6.31	6.07	-1.80	0.15	1.00	0.1	0.86	-0.1	0.93	-.10	115
99	16	6.19	5.97	-1.85	0.16	0.89	0.0	0.85	-0.1	1.14	.42	101
24	4	6.00	5.88	-1.90	0.35	0.17	-0.9	0.17	-0.9	0.63	.43	79
24	4	6.00	5.88	-1.90	0.35	1.62	0.8	1.82	0.9	0.94	-.34	89
92	16	5.75	5.62	-2.07	0.20	1.14	0.4	0.88	0.0	1.16	.50	103
20	4	5.00	5.07	(-3.35	1.48)	Minimum					.00	76
20	4	5.00	5.07	(-3.35	1.48)	Minimum					.00	83

(Table continues)

Appendix T—Continued

149	12.6	11.79	11.81	-0.63	0.17	0.85	-0.5	0.85	-0.5		0.12	Mean (Count: 126)
76.7	5.4	3.10	3.18	0.70	0.18	0.81	1.7	0.83	1.7		0.29	S.D. (Population)
77	5.4	3.11	3.19	0.70	0.18	0.81	1.7	0.83	1.7		0.29	S.D. (Sample)

With extremes, Model, Populn: RMSE .24 Adj (True) S.D. .66 Separation 2.72 Strata 3.96 Reliability .88

With extremes, Model, Sample: RMSE .24 Adj (True) S.D. .66 Separation 2.73 Strata 3.97 Reliability .88

Without extremes, Model, Populn: RMSE .16 Adj (True) S.D. .59 Separation 3.82 Strata 5.43 Reliability .94

Without extremes, Model, Sample: RMSE .16 Adj (True) S.D. .60 Separation 3.84 Strata 5.45 Reliability .94

With extremes, Model, Fixed (all same) chi-square: 2709.9 d.f.: 125 significance (probability): .00

With extremes, Model, Random (normal) chi-square: 95.7 d.f.: 124 significance (probability): .97

Appendix U. Unexpected Responses in Three-Faceted Rasch Measurement: Participants × Raters × Tasks (Length × Mode)

Score	Exp.	Resd	StRes	Participant	Rater	Task
5	14.7	-9.7	-4.6	3	2	100R
5	14.6	-9.6	-4.6	3	4	100R
5	13.9	-8.9	-4.2	3	1	100R
5	13.7	-8.7	-4.1	3	3	100R
5	12.8	-7.8	-3.6	44	2	100R
21	14.6	6.4	3.1	31	3	100R
5	12	-7	-3.1	44	1	100R
6	12.7	-6.7	-3.1	44	4	100R
22	15.3	6.7	3.1	113	3	100R
5	11.7	-6.7	-3	44	3	100R
23	16.3	6.7	3	113	2	100R
23	16.2	6.8	3	113	4	100R
5	11.7	-6.7	-2.9	2	2	150R
5	11.5	-6.5	-2.9	2	4	150R
21	24.3	-3.3	-2.9	9	4	100R
19	12.8	6.2	2.9	35	3	100R
7	13	-6	-2.8	36	3	150L
11	6.3	4.7	2.8	101	1	100R
7	12.8	-5.8	-2.7	22	4	100L
8	13.6	-5.6	-2.7	26	1	100R
6	12	-6	-2.7	42	3	150L
19	13.3	5.7	2.7	47	4	100R
21	15.5	5.5	2.6	49	3	150R
18	12.5	5.5	2.5	3	1	150L
7	12.4	-5.4	-2.5	8	3	100L
19	13.8	5.2	2.5	35	2	100R
19	13.7	5.3	2.5	35	4	100R
17	11.1	5.9	2.5	44	4	150L
22	16.4	5.6	2.5	49	4	150R
10	6.2	3.8	2.5	103	4	100R
5	10.7	-5.7	-2.4	2	1	150R

(Table continues)

Appendix U—Continued

7	12.3	-5.3	-2.4	35	4	150L
9	5.8	3.2	2.4	89	1	Prep
13	7.7	5.3	2.4	116	3	300R
5	10.8	-5.8	-2.4	122	1	100R
5	10.5	-5.5	-2.3	2	3	150R
18	13.1	4.9	2.3	3	4	100L
14	20.2	-6.2	-2.3	5	1	100R
5	10.5	-5.5	-2.3	7	2	150R
21	24	-3	-2.3	9	1	100R
18	13.1	4.9	2.3	27	3	100R
17	11.9	5.1	2.3	30	2	150L
13	7.7	5.3	2.3	95	1	150R
5	10.5	-5.5	-2.3	97	1	500R
23	17.6	5.4	2.2	5	3	100L
5	10.4	-5.4	-2.2	7	4	150R
18	13.3	4.7	2.2	18	2	100R
7	11.8	-4.8	-2.2	22	3	100L
7	12	-5	-2.2	29	4	150L
11	15.7	-4.7	-2.2	40	2	150R
15	9.6	5.4	2.2	45	2	150R
5	10.3	-5.3	-2.2	111	2	300R
12	7.3	4.7	2.2	120	3	300R
12	17.5	-5.5	-2.2	123	2	100R
17	12.4	4.6	2.1	3	1	100L
19	14.5	4.5	2.1	24	3	150L
7	11.8	-4.8	-2.1	25	3	100L
20	15.4	4.6	2.1	31	4	100R
15	9.8	5.2	2.1	44	1	150R
12	16.8	-4.8	-2.1	49	4	150L
10	14.4	-4.4	-2.1	113	1	300R
14	8.9	5.1	2.1	119	3	300R
16	21.3	-5.3	-2	5	4	100R
13	18.1	-5.1	-2	6	1	100R
20	23.5	-3.5	-2	9	2	150R

(Table continues)

Appendix U—Continued

18	13.7	4.3	2	21	2	100R
6	10.8	-4.8	-2	21	3	150R
17	12.6	4.4	2	23	1	150R
20	15.6	4.4	2	40	4	150R
5	10	-5	-2	45	2	100L
5	9.9	-4.9	-2	110	2	300R