Inference Generation Processes of Japanese EFL Learners:
Effects of Questioning on Their Reading Comprehension

A Dissertation

Submitted to the University of Tsukuba

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Linguistics

Maki SHIMIZU
Abstract

Inference Generation Processes of Japanese EFL Learners: Effects of Questioning on Their Reading Comprehension

By Maki Shimizu

When we focus on mental representations constructed by readers in the course of or after reading, it can be argued that the goal of reading is to construct a coherent “situation model” (Kintsch, 1988, 1994, 1998; van Dijk & Kintsch, 1983; van Oostendorp & Goldman, 1999; van Oostendorp, 2002); according to Kintsch (1994), “in the situation model the information provided by the text is elaborated from prior knowledge and is integrated with it” (p. 294). Furthermore, it is known that these construction and integration processes (Kintsch, 1988, 1998) largely rely on readers’ inference generation.

In the field of cognitive psychology, several theories of readers’ inference generation were vigorously proposed in the early 1990s, and a number of surveys followed to attempt to verify those theories. In contrast, we know only a small number of studies in the field of second language (L2) reading (Horiba, 2001). The knowledge obtained by the investigations into L2 learners’ inference generation will contribute substantially to the construction of L2 reading theory. Furthermore, it will also offer important pedagogical implications.

The main aims of the present study are as follows: (a) to clarify inference generation processes of Japanese EFL learners, (b) to investigate the effects of reading questions on learners’ inference making, (c) to examine some language tests for EFL (English as a foreign language) or ESL (English as a second language) learners, based on componential models of reading, in which inference generation processes are included. Four pieces of research, Studies 1, 2, 3, and 4, were conducted in pursuit of these aims. To be more specific, Study 1 and a part of Study 2 are related to the aim (a); a part of Study 2 and Study 3 are related to the aim (b); and finally, Study 4 is related to the aim (c). These results are summarized in the following.

With regard to inference generation processes, Study 1 revealed that the learners of higher L2 proficiency inferred more frequently than the learners of lower L2 proficiency. Furthermore, the obvious difference was found in the amount of “local connecting inferences” which are defined as inferences that connect two pieces of information written within one or two adjacent sentences in a text, but not in “elaborative inferences,” which are defined as inferences that elaborate information in the text by the use of readers’ prior knowledge.

Study 2 narrowed its target to a “causal antecedent inference” (Graesser, Singer, & Trabasso, 1994; Graesser & Kreuz, 1993), which is a kind of connecting inference, and more precisely investigated generation processes. The results of correlation analyses showed that there was a moderate correlation between generating causal antecedent inferences and
understanding the meaning of the propositions which were related to but not overtly targeted by the questions. Moreover, the examination of the relationships between learners’ L2 reading proficiency and “causal antecedent inference” generation showed a stronger relationship when the inferences were measured off-line than on-line, that is, not during reading but while undertaking a retrieval task. This seems to reveal an important characteristic of L2 reading processes. Inference generation ability of L2 learners may develop from off-line inference generation processes into on-line inference generation processes. This is because the difference in the amount of “causal antecedent inferences” between the different L2 reading proficiency groups was found in the off-line context only; in contrast, according to the previous studies on L1 reading, the inferences are considered to usually be generated on-line.

In addition, in Study 2 speculation was made of the effects of inference questions on learners’ comprehension. “Why” questions designed to make readers infer the possible causes of a particular event in a text (Graesser et al., 1994) were used in this study. I obtained the following findings: (a) Inference generation processes were significantly activated by the questions, and (b) the questions helped the learners identify propositional meanings of certain parts of a text as well. The second finding, in particular, appears to be interesting because the result can be considered to be one of the indirect effects.

My interests of Study 3 were in what order we should arrange a series of inference questions to make them more effective, and what relationships would be found between question sequence and learners’ L2 reading proficiency. For the learners of lower L2 reading proficiency, compared with the opposite sequence, the sequence in which they answered inferential questions during the early stages of the sequence, far from thematic questions which were presented at the end of sequence, was found to have a positive effect.

Study 4 examined each item on the reading sections of three popular tests from the viewpoint of reading processes that have been identified in previous studies. This examination showed that TOEFL contained various types of questions, including inference questions; in contrast, most of the items in the STEP test favorably asked for the information explicitly stated in a text; in other words, inference questions were rarely found. Furthermore, only a moderate coefficient was obtained by the correlation analysis between the questions labelled as lower-level processing questions (i.e., questions ask for the information explicitly stated in a text) and those labelled as higher-level processing questions, which included inference questions as well as the other two types of questions. From these results, my claim that higher-level processing questions should be constructed for reading comprehension assessment, in addition to lower-level processing questions, was strengthened.

This paper is concluded by providing some implications for the area of theory development of L2 reading research, and that of exploring effective tasks in improving reading performance.
Acknowledgments

I owe a debt of gratitude to many people for their invaluable guidance and assist once over the course of my studies at the Graduate School of Humanities and Social Sciences, University of Tsukuba. Although it would be difficult to acknowledge everyone who has in various ways contributed to the research reported in this dissertation, the following people deserve special recognition.

First of all, I would like to thank my advisors. Akihiko Mochizuki, my chief advisor, has been graciously generous with his time and has provided a great deal of helpful advice. His comments have enlightened me considerably, in areas ranging from constructing a research design in the field of English language teaching to compiling the results into papers. His faith in me and his ability to bring me back on track at difficult times have been invaluable during these five years. Yuji Ushiro awakened my interest in second language learners’ reading processes just after I became a graduate student. What I learned and discussed in his class and his Reading Research Group at the University of Tsukuba inspired me to study this field more deeply. His invaluable comments also substantially helped me in this study. Akiyo Hirai has given me a number of insightful pieces of advice about writing academic papers, analyzing data, and organizing this dissertation. I have learned a great deal from her. Hirosada Iwasaki has also given me precious comments and showed his expertise about vocabulary. His suggestions considerably helped me to analyze the reading materials used in my research. Akira Kubota taught me how to deal with the data on inferences generated by second language learners. His insightful advice also helped my ideas to develop during the writing of this dissertation. Tsuneo Takanashi, Kyoto Notre Dame University, willingly accepted a request to be an external member of my dissertation committee. His helpful comments greatly assisted me in revising an earlier version of this dissertation.

Many thanks are also due to Akihiro Yamaguchi and Takeshi Furuya at Gunma University, for providing me with concrete suggestions on data analyses and interpretations from the standpoint of cognitive psychology.

I am very thankful to Rie Koizumi and Yo In’nami, who gave me many helpful comments on research design, statistical analyses, and procedures for test validation, through the dissertation support group headed by Mochizuki. I also would like to express my appreciation to Miki Hirose and all the other group members for their help and suggestive comments.

My special thanks go to Yuko Hijikata, Yuko Morimoto, and Chikako Nakagawa for their constructive suggestions on my research and their continuous support. Also, I would like to thank Koga Tsutomu and the other members of the Reading Research Group at the University of Tsukuba for sharing various experiences with me over the last four years. I also appreciate the many insightful comments offered by all the members of Ushiro’s dissertation.
support group. Many thanks are due to Ryuta Iseki, University of Tsukuba, for his valuable advice, especially as to first language reading research. His supports have contributed substantially to the current study.

I am also indebted to Sakae Okuma, Masaharu Shimada, and Koichi Miyakoshi, all of the University of Tsukuba, who have generously allowed me to ask their students to participate in the research for this dissertation. I am also indebted to the participants.

I wish to acknowledge my gratitude to Seiichiro Okajima and Ichiro Horimoto at Ota High School, Masanori Tanaka at Takasaki High School, Hirotaka Tagaya at Isesaki Higashi High School, Takeshi Nakajima at Takasaki Kita High School, Rika Okamoto at Ueno High School, and Fumiyasu Orihara at Moriya High School, for their support for conducting the research reported in this study. Thanks also go to their high school students for their cooperation.

I am indebted to Glen McCabe, a graduate student at the University of Tsukuba, for his continuous support of the present research, and to his friends, Kathryn Hackshaw and Mark Rosa, for proofreading this dissertation.

I am grateful to the Society for Testing English Proficiency for awarding a grant for a part of this study. I also appreciate the technical advice of Hiroshi Ikeda, my advisor on the portion of the research supported by the grant.

I would like to offer my sincere thanks to John Rippey at Gunma University, who read the earlier version of my master’s thesis, which was submitted to the University of Tsukuba in 2003 and served as the basis for this dissertation, and who provided me with valuable comments. My heartfelt appreciation also goes out to the now retired Minoru Takashima of Gunma University, and to Keiko Uehara at Gunma University, for providing me with a chance to observe their work in the academic field of English language education, when I was a research student at Gunma University. I am also grateful for the constant encouragement of Masato Kosuge at Maebashi High School, who was my English teacher in high school.

Finally, I would like to express my deep gratitude to my parents for their understanding, encouragement, and warm support. My father taught me the pleasure of learning. My mother always believed in me.

Maki Shimizu
Tsukuba, Japan
December, 2005
Table of Contents

Abstract ................................................................................................................................... ii
Acknowledgments .................................................................................................................... iv
Table of Contents .................................................................................................................... vi
List of Tables .............................................................................................................................. ix
List of Figures ............................................................................................................................ xii
List of Appendixes .................................................................................................................... xiv

Chapter
1. Introduction ................................................................................................................................. 1
2. Literature Review ....................................................................................................................... 6
   2.1 Reading Processes: Componential Models of Reading ....................................................... 6
   2.2 Inference Generation Processes ......................................................................................... 11
       2.2.1 Reading Models and L2 Reading Research ............................................................... 11
       2.2.2 Data Collection Methods for Inferences .................................................................. 31
   2.3 Effects of Questions on Inference Generations ................................................................. 37
   2.4 Present Study ....................................................................................................................... 44
3. Study 1: Mental Representations of EFL Learners of Different L2 Proficiency, With a Focus on Inferences ............................................................................................................. 47
   3.1 Purposes and Research Hypotheses ................................................................................... 47
   3.2 Method ................................................................................................................................ 49
       3.2.1 Participants .................................................................................................................. 49
       3.2.2 Materials ...................................................................................................................... 49
       3.2.3 Data Collection Procedures ....................................................................................... 52
       3.2.4 Scoring ......................................................................................................................... 54
       3.2.5 Data Analyses .............................................................................................................. 56
   3.3 Results .................................................................................................................................. 57
       3.3.1 Validity and Descriptive Statistics of L2 Proficiency Test ........................................... 57
       3.3.2 Results of a Recall Test, With a Focus on Inferences ................................................... 59
   3.4 Discussion ............................................................................................................................ 74
   3.5 Summary .............................................................................................................................. 78
4. Study 2: Causal Antecedent Inference Generation and Questioning .................................... 81
   4.1 Purposes and Research Hypotheses ................................................................................... 81
   4.2 Method ................................................................................................................................ 84
       4.2.1 Participants .................................................................................................................. 84
       4.2.2 Materials ...................................................................................................................... 86
       4.2.3 Data Collection Procedures ....................................................................................... 89
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.2.4 Transcribing and Scoring</td>
<td>90</td>
</tr>
<tr>
<td>4.2.5 Data Analyses</td>
<td>93</td>
</tr>
<tr>
<td>4.3 Results</td>
<td>94</td>
</tr>
<tr>
<td>4.3.1 Participants’ Responses to a Questionnaire</td>
<td>94</td>
</tr>
<tr>
<td>4.3.2 Validity and Descriptive Statistics of L2 Reading Proficiency Test</td>
<td>98</td>
</tr>
<tr>
<td>4.3.3 Results of Think-Aloud Protocols</td>
<td>101</td>
</tr>
<tr>
<td>4.4 Discussion</td>
<td>107</td>
</tr>
<tr>
<td>4.5 Summary</td>
<td>116</td>
</tr>
<tr>
<td>5. Study 3: Effects of Inference Question Sequence on Reading Comprehension</td>
<td>118</td>
</tr>
<tr>
<td>5.1 Purposes and Research Hypotheses</td>
<td>118</td>
</tr>
<tr>
<td>5.2 Pilot Study 1</td>
<td>122</td>
</tr>
<tr>
<td>5.3 Pilot Study 2</td>
<td>125</td>
</tr>
<tr>
<td>5.4 Main Study</td>
<td>127</td>
</tr>
<tr>
<td>5.4.1 Method</td>
<td>127</td>
</tr>
<tr>
<td>5.4.1.1 Participants</td>
<td>127</td>
</tr>
<tr>
<td>5.4.1.2 Materials</td>
<td>129</td>
</tr>
<tr>
<td>5.4.1.3 Data Collection Procedures</td>
<td>132</td>
</tr>
<tr>
<td>5.4.1.4 Data Analyses</td>
<td>133</td>
</tr>
<tr>
<td>5.4.2 Results</td>
<td>134</td>
</tr>
<tr>
<td>5.4.2.1 Validity and Descriptive Statistics</td>
<td>134</td>
</tr>
<tr>
<td>5.4.2.1.1 Society for Testing English Proficiency</td>
<td>134</td>
</tr>
<tr>
<td>5.4.2.2 Tests of the National Center for University Entrance Examinations</td>
<td>140</td>
</tr>
<tr>
<td>5.4.2.2.2 Comparisons of Different Sequences of Answering Questions: Interactions With Learners’ L2 Reading Proficiency</td>
<td>148</td>
</tr>
<tr>
<td>5.4.3 Discussion</td>
<td>157</td>
</tr>
<tr>
<td>5.5 Summary</td>
<td>167</td>
</tr>
<tr>
<td>6. Study 4: Questions on Reading Tests Which Measure Higher-Level Processes and Lower-Level Processes</td>
<td>169</td>
</tr>
<tr>
<td>6.1 Purposes and Research Hypotheses</td>
<td>169</td>
</tr>
<tr>
<td>6.2 Pilot Study</td>
<td>171</td>
</tr>
<tr>
<td>6.2.1 Method</td>
<td>171</td>
</tr>
<tr>
<td>6.2.2 Results and Discussion</td>
<td>174</td>
</tr>
<tr>
<td>6.3 Main Study</td>
<td>181</td>
</tr>
<tr>
<td>6.3.1 Method</td>
<td>181</td>
</tr>
<tr>
<td>6.3.1.1 Participants</td>
<td>181</td>
</tr>
<tr>
<td>6.3.1.2 Materials</td>
<td>183</td>
</tr>
<tr>
<td>6.3.1.3 Data Collection Procedures and Data Analyses</td>
<td>185</td>
</tr>
<tr>
<td>6.3.2 Results and Discussion</td>
<td>186</td>
</tr>
</tbody>
</table>
List of Tables

Table                                           Page

2.1. Brief Descriptions and Examples of Knowledge-Based Inferences……………………13
2.2. Predictions of On-Line Inference Processing (Graesser et al., 1994, p. 384)…………15
2.3. Definitions of van den Broek et al.’s (1993) Inference Types…………………………22
2.4. Examples of Inference Types Proposed by van den Broek et al. (1993)……………….23
2.5. Summary of Reading Models…………………………………………………………………29
2.6. Tasks for Data Collection…………………………………………………………..…………32
2.7. Summary of Question Types by Ikeno (2000)………………………………………………38
3.1. Examples of L2 Proficiency Test Used in Study 1……………………………….………..50
3.2. Passages Used in Study 1……………………………………………………………………51
3.3. Results of a Pilot Study……………………………………………………………………52
3.4. Allotted Times for a Recall Task in Study 1………………………………………..……53
3.5. Means and SDs of the L2 Proficiency Test………………………………………………58
3.6. Inter-Rater Reliability in the Scoring of Idea Units by Passage…………………………59
3.7. Descriptive Statistics of Idea Units Produced by Written Recall Protocols…………………60
3.8. Inter-Rater Reliability in the Scoring of Inferences by Passage………………………….61
3.9. Results of a Repeated-Measures ANOVA: Tests of Within-Subjects Effects……………….63
3.10. Results of a Repeated-Measures ANOVA: Tests of Between-Subjects Effects………….65
3.11. T-Tests for Average Number of Inferences of Two Groups…………………………….66
3.12. Repeated-Measures One-Way ANOVAs for Effects of L2 Proficiency on Inference Types……69
3.13. Results of Tukey’s HSD Test (1)………………………………………………………….70
3.14. Descriptive Statistics of Three Types of Inferences by Passage…………………………….71
3.15. Repeated-Measures One-Way ANOVAs for Effects of Passage on Inference Types (1)………….72
3.16. Results of Tukey’s HSD Test (2)……………………………………………………………..73
3.17. Repeated-Measures One-Way ANOVAs for Effects of Passage on Inference Types (2)………….73
3.18. Results of Tukey’s HSD Test (3)…………………………………………………………….74
4.1. Academic Fields of the Participants in Study 2………………………………………….85
4.2. Narrative Passages for a Think-Aloud Task in Study 2……………………………………87
4.3. Think-Aloud Procedures……………………………………………………………………90
4.4. Frequency Levels of Words in Target IUs………………………………………………92
6.3. Subcategories Used for “Inference Questions” in Pilot Study .......................174
6.4. Types of Questions in STEP Test, DNC Test and TOEFL ............................175
6.5. Number of Items and Proportions by Question Type and Test .....................177
6.6. Academic Fields of the Participants in Study 4 .........................................182
6.7. Reading Tests Used in Study 4 ..................................................................184
6.8. Administered Questions on the STEP Test and TOEFL in Study 4 .................185
6.9. Point Biserial Correlation Coefficients of STEP Test Used in Study 4 ..........188
6.10. Point Biserial Correlation Coefficients of TOEFL Used in Study 4 ..........188
6.11. Means and SDs of STEP Test Used in Study 4 .......................................189
6.12. Means and SDs of TOEFL by Question Type Used in Study 4 ...............189
6.13. Correlation Coefficients Among “Paraphrase Questions” on TOEFL,
      “Paraphrase Questions” on STEP Test, and Higher-Level Processing
      Questions on TOEFL ..............................................................................190
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Relationships among the four studies reported in this dissertation</td>
<td>5</td>
</tr>
<tr>
<td>2.1 A two-sentence text fragment with its textbase and situation model (Kintsch, 1994, 295)</td>
<td>9</td>
</tr>
<tr>
<td>2.2. Types of inferences in reading (van den Broek et al., 1993, p. 171)</td>
<td>21</td>
</tr>
<tr>
<td>2.3. A fragment of the associative net for BANK (Kintsch, 1988, p. 165)</td>
<td>24</td>
</tr>
<tr>
<td>2.4. Memory model (van Someren et al., 1994, p. 19)</td>
<td>35</td>
</tr>
<tr>
<td>3.1. A line graph describing an interaction between inference type and L2 proficiency</td>
<td>64</td>
</tr>
<tr>
<td>3.2. A line graph describing an interaction between passage and inference type</td>
<td>64</td>
</tr>
<tr>
<td>3.3. Means of inferences by type and group in Passage 1</td>
<td>67</td>
</tr>
<tr>
<td>3.4. Means of inferences by type and group in Passage 2</td>
<td>67</td>
</tr>
<tr>
<td>3.5. Means of inferences by type and group in Passage 3</td>
<td>68</td>
</tr>
<tr>
<td>3.6. Memory model (van Someren et al., 1994, p. 19)</td>
<td>35</td>
</tr>
<tr>
<td>4.1. Participants’ attitudes toward think-aloud protocols: Question 6</td>
<td>96</td>
</tr>
<tr>
<td>4.2. Inferences generated before “why” questions and those generated after “why” questions</td>
<td>103</td>
</tr>
<tr>
<td>4.3. Numbers of IUs understood during first reading without “why” questions and second reading with “why” questions</td>
<td>106</td>
</tr>
<tr>
<td>4.5. A: The text representation at time t, consisting of six nodes and their connections. B: The text representation at time t + 1, consisting of the original six nodes and their connections from time t, plus a new node and its connections (Langston &amp; Trabasso, 1999, p. 39)</td>
<td>111</td>
</tr>
<tr>
<td>5.1. Mean scores of tests of DNC by L2 reading proficiency and question sequence</td>
<td>149</td>
</tr>
<tr>
<td>5.2. Cumulative scores of “Lower” on the fiscal 2003 DNC test (Passage 1)</td>
<td>151</td>
</tr>
<tr>
<td>5.3. Cumulative scores of “Lower” on the fiscal 2004 DNC test (Passage 2)</td>
<td>151</td>
</tr>
<tr>
<td>5.4. Cumulative scores of “Upper” on the fiscal 2003 DNC test (Passage 1)</td>
<td>152</td>
</tr>
<tr>
<td>5.5. Cumulative scores of “Upper” on the fiscal 2004 DNC test (Passage 2)</td>
<td>152</td>
</tr>
<tr>
<td>5.6. The dendrograms of the “Lower” Group A’s and “Lower” Group B’s fiscal 2003 DNC test</td>
<td>155</td>
</tr>
<tr>
<td>5.7. The dendrograms of the “Lower” Group A’s and “Lower” Group B’s fiscal 2004 DNC test</td>
<td>155</td>
</tr>
</tbody>
</table>
5.8. The dendrograms of the “Upper” Group A’s and “Upper” Group B’s fiscal 2003 DNC test……………………………………………………………..…………………156

5.9. The dendrograms of the “Upper” Group A’s and “Upper” Group B’s fiscal 2004 DNC test……………………………………………………………..…………………156
## List of Appendixes

<table>
<thead>
<tr>
<th>Appendix</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-A. L2 Proficiency Test Used in Study 1</td>
<td>218</td>
</tr>
<tr>
<td>3-B. Reading Passages Used for Recall Tasks in Study 1</td>
<td>224</td>
</tr>
<tr>
<td>3-C. Rubric of Recall Tasks</td>
<td>227</td>
</tr>
<tr>
<td>3-D. Point Biserial Correlation Coefficients of the STEP Test Used</td>
<td>228</td>
</tr>
<tr>
<td>4-A. Reading Passages for the Think-Aloud Protocol</td>
<td>229</td>
</tr>
<tr>
<td>4-B. Questionnaire Used in Study 2</td>
<td>231</td>
</tr>
<tr>
<td>4-C. Rubric of the Think-Aloud Tasks</td>
<td>232</td>
</tr>
<tr>
<td>4-D. Passage Used for a Demonstration</td>
<td>233</td>
</tr>
<tr>
<td>4-E. Passage Used for a Practice</td>
<td>233</td>
</tr>
<tr>
<td>4-F. Scoring Guide Used in Study 2</td>
<td>234</td>
</tr>
<tr>
<td>4-G. Criteria Used for Dividing Passages Into Idea Units</td>
<td>237</td>
</tr>
<tr>
<td>4-H. Students’ Questionnaire Responses to Question 7</td>
<td>239</td>
</tr>
<tr>
<td>(Difficulty in Thinking Aloud)</td>
<td></td>
</tr>
<tr>
<td>4-I. Point Biserial Correlation Coefficients of the STEP Test Used in Study 2</td>
<td>242</td>
</tr>
<tr>
<td>5-A. Questionnaire Administered in Pilot Study 2 of Study 3</td>
<td>243</td>
</tr>
<tr>
<td>5-B. English Tests of the National Center for University Entrance</td>
<td></td>
</tr>
<tr>
<td>Examinations Used in Study 3</td>
<td>244</td>
</tr>
<tr>
<td>5-C. Society for Testing English Proficiency 2nd Grade Test Used in Study 3</td>
<td>248</td>
</tr>
<tr>
<td>5-D. Questionnaire Administered After Tests in Study 3</td>
<td>250</td>
</tr>
<tr>
<td>5-E. Rubric Given to Administrators in Study 3</td>
<td>251</td>
</tr>
<tr>
<td>6-A. Test of Society for Testing English Proficiency in Study 4</td>
<td>254</td>
</tr>
<tr>
<td>6-B. Sample of Section 3, TOEFL Practice Test A (Passages 1-3)</td>
<td>257</td>
</tr>
</tbody>
</table>
1. Introduction

Most researchers now acknowledge that reading is an active rather than a passive activity. Behind this change in thinking, there has been a significant shift of emphasis from the text itself to the reader as an individual. It is thought that the advent of schema theory (e.g., Bransford & Johnson, 1973; Sanford & Garrod, 1981; Schank & Abelson, 1977) played an important role in taking us to this stage. Schema theory considers readers’ prior knowledge, such as their domain knowledge concerning a topic of a passage, and their knowledge of how a language or a text is organized from the viewpoint of rhetoric. The former knowledge has been called “content schemata,” and the latter knowledge has been called “formal schemata.” Inference generation constitutes one of the reading processes that readers advance and is closely related to readers’ knowledge. Therefore, the focus of the present study is on readers rather than on texts.

If we take a look at the area of first language (L1) reading research of the last two decades, in particular, within the domain of cognitive psychology, we can find a number of studies that have been conducted with a focus on inference generation. These surveys have been conducted in order to construct and make more sophisticated models for inference generation (e.g., Graesser & Kreuz, 1993; Graesser, Singer, & Trabasso, 1994; McKoon & Ratcliff, 1992; van den Broek, Fletcher, & Risden, 1993). As a result of such studies, a general overview of inference generation has emerged. However, what the inference generation processes of EFL (English as a foreign language) learners look like is still unknown, possibly due to a lack of related research. Because there are a number of differences between the reading processes of L1 readers and EFL readers, research which specifically targets EFL learners is certainly required. Thus, the first aim of the present study is, by conducting surveys, to investigate how EFL learners are engaged in this process.
Furthermore, to find a task that positively affects EFL learners’ inference generation processes is the second aim of the present study. Among a number of possible reading tasks, questioning was chosen for the present study. This is because, in the context of reading instruction, questions are frequently used by teachers; and many EFL reading textbooks also contain questions which function as guiding questions. Thus, posing a reading question has been a popular technique to enable students to comprehend a text more deeply. However, there is little empirical data to show the effect of a question on EFL learners’ inference generation processes. In the present study, I intend to delve into the potentiality of questioning.

Such were the aims of the present study, the findings of which will be presented as follows. The present study was conducted to achieve these aims. First of all, literature related to the present study is reviewed in Chapter 2. Theories and reading models, which have been developed with a focus on readers’ inference generation, as well as previous studies on reading questions, are carefully examined. Based on this review, areas that remain unanswered and need to be examined have been determined.

In Chapter 3, inferences generated by EFL learners in reading a text have been investigated. Their inferences were classified based on van den Broek et al.’s (1993) framework, and the amount of each inference type was measured. The results showed that there was a large difference between higher second language (L2) proficiency learners and lower L2 proficiency learners in the amount of the inferences which connect the focal point of a text to an earlier point in making a text locally coherent.

Examining these results, the “causal antecedent inferences” of EFL learners were surveyed in more detail in Study 2. “Causal antecedent inferences” refer to the inferences concerning causes which are related to certain effects, and function to connect the focal
point to the earlier point in text in order to create textual coherence. Moreover, I employed “why” questions, one of the “inference questions” proposed by Graesser and his colleagues (e.g., Graesser & Clark, 1985; Graesser & Goodman, 1985), and investigated their effectiveness on inferences learners make and to grasp the meanings of particular points of a text. Positive effects of questioning were revealed. All discussion of Study 2 can be found in Chapter 4.

Study 3 is discussed in Chapter 5. Recognizing the effectiveness of “inference questions” in Study 2, the purpose of Study 3 was determined as follows: in order to improve learners’ performance in reading comprehension, how “inference questions” should be presented while reading a passage. “Inference questions” were ranked beforehand in accordance with the degree of proximity to a text theme. Consequently, interaction between question sequence and learners’ L2 reading proficiency was found. To be more specific, lower L2 reading proficiency learners received benefits from the sequence in which they answered inferential questions during the early stages of the sequence, far from thematic questions which were presented at the end of sequence. There are two interpretations of this result. First, that the learners did not have enough vocabulary knowledge and grammatical knowledge to form a base from which to make inferences and understand a main theme of a passage; and that the local-to-global transition could guide their comprehension properly. Second, that to answer a main theme of a passage first of all was cognitively too demanding for such learners who are deficient in lower-level processing skills; that is, the second interpretation is from the perspective of distribution of readers’ cognitive resources and working memory capacity.

Study 4 is discussed in Chapter 6, and was conducted in order to find out whether English language tests for EFL learners contain enough questions to assess examinees’
higher-level processing skills, including questions which ask for inferences or a main theme of a passage. This was because in Study 3 it was found that lower L2 reading proficiency learners had difficulty in understanding these types of questions. The relationships between all of these studies, Studies 1, 2, 3, and 4, are represented visually on the next page (see Figure 1).

Finally, Chapter 7 is devoted to discussing and summarizing the findings of these studies, and to providing promising areas for further research, as well as reading instructions for EFL classroom in the future. Some of the limitations of the present study are also discussed.
Study 1: Mental Representations of EFL Learners of Different L2 Proficiency, With a Focus on Inferences

Study 4: Questions on Reading Tests Which Measure Higher-Level Processes and Lower-Level Processes

Study 2: Causal Antecedent Inference Generation and Questioning

Study 3: Effects of Inference Question Sequence on Reading Comprehension

Figure 1. Relationships among the four studies reported in this dissertation.
2. Literature Review

2.1 Reading Processes: Componential Models of Reading

The idea that reading processes are comprised of a number of components appears to be strongly supported by current research (Urquhart & Weir, 1998). Grabe (2000) lists the following components:

…orthographic processing, phonological coding, word recognition (lexical access), working memory activation, sentence parsing, propositional integration, propositional text-model formation, comprehension strategy use, inference making, text-model development, and the development of an appropriate situation model (or mental model). (p. 230)

We can recognize that a variety of processes are assumed in this model, which range from the process of a system for spelling words or speech sounds in a language to the process of making a rich mental representation while activating background knowledge.

Ample evidence has been provided that L2 reading ability is related to the following components: “orthographic processing” (e.g., Koda, 1999; Noro, 1999), “phonological coding” (e.g., Koda, 1998; Noro), “word recognition” (e.g., Akamatsu, 2002; Droop & Verhoeven, 2003; Fukkink, Hulstijn, & Simis, 2005; Koda, 1992, 1996), “working memory activation” (e.g., Harrington & Sawyer, 1992; Ikeno, 2002; Yoshida, 2003), “sentence parsing” (e.g., Droop & Verhoeven), “comprehension strategy use” (e.g., Block, 1986, 1992; Carrell, 1989; Hirano, 1998; Iijima, 1998; Phakiti, 2003; Temma et al., 1992; Ushiro, 1999), and inference making (e.g., Hammadou, 1991; Horiba, 1993, 1996; Muramoto, 2000, Yoshida, 2003, 2005; Zwaan & Brown, 1996). Of these, the focus of the present study is on inference making.
Furthermore, in recent years the number of studies which have started to investigate what relationships exist among these components has increased (e.g., Horiba, 1996; Ikeno, 2002; Nassaji, 2003; Noro, 1999; Shiotsu, 2003). Most of these studies discuss the notions of “lower-level processes” and “higher-level processes.” A popular research design has been to compare some “lower-level processes” with some “higher-level processes.”

Let us refer to Grabe’s (2000) description of “low(er)-level processes” and “high(er)-level processes.” The following concerns “low(er)-level processes.”

Low-level processing can be discussed in terms of three sub-component processes: The recognition of orthographic structure (recognizing line forms, letter shapes, letter group patterns), the recognition of morpheme structure, and the processing of phonemic information. (p. 231).

We can see that to process some basic units of a language such as sounds or letters belongs at this level. In addition, Grabe appears to consider syntactic parsing and propositional integration to be examples of “low(er)-level processing.” “High(er)-level processing” is described as “working with larger units of information and information contributed by the reader” (Grabe, p. 233), and refers to the construction of both a text model of comprehension and a situation model. Through their relationship to these, making appropriate inferences and using reading comprehension strategies are thought to substantially contribute to building such mental representations.

It should be noted here that my intention has not been to suggest that all processes can be divided into either lower-level or higher-level processes; rather, each process should be placed along a continuum ranging from the extremely “lower-level processes” to the extremely “higher-level processes.” However, it can be acknowledged that a binary division
provides us with a promising approach, which can reasonably explain how readers use their
cognitive resources or working memory capacity (Miyake & Shah, 1999) during the course
of reading. Therefore, the present study, by adopting this notion, will label the processes that
have been achieved by remarkable consensus in previous studies as “lower-level process”
and “higher-level process.” I will follow Grabe’s (2000) categorization, which does not
seem to sharply deviate from a common understanding on these levels of reading processes.

The mental representations (Kintsch, 1988, 1994, 1998; van Dijk & Kintsch, 1983; van
Oostendorp & Goldman, 1999; van Oostendorp, 2002) should also be mentioned. The
following three different levels of mental representations have been empirically verified: (a)
With regard to “surface” memory, “the words and phrases themselves are encoded as are the
linguistic relations between them,” (b) “textbase” is “the semantic and rhetorical structure of
the text,” and finally, (c) “situation model” “corresponds to a deeper level of
understanding,” and “In the situation model the information provided by the text is
elaborated from prior knowledge and is integrated with it” (Kintsch, 1994, p. 294). That
which is considered “surface” memory is not concerned with any meaning, whereas
“propositional text base” and “situation model” largely concern meaning.

Let us take examples of the last two representations from Kintsch (1994; see Figure
2.1), which are related to propositional or textual meaning. As is shown, the current text
contains two sentences. On the one hand, “textbase” as derived from the text consists of
three propositions; a proposition is defined as the smallest idea unit which can be judged to
be true or false. Moreover, some connections are found between propositions. However, any
information which is not retrieved from readers’ prior knowledge; that is, only the textual
information, is represented as the “textbase.” On the other hand, the “situation model”
represents the information retrieved from reader’ prior knowledge as well as the information
elicited from the text. For example, the information about “red blood” is not seen in the text at all, but the reader’s prior knowledge of blood must be utilized. This situation model indicates how blood circulates around the body through the heart and lungs in the case of septal defect.

Text:
When a baby has a septal defect, the blood cannot get rid of enough carbon dioxide through the lungs. Therefore, it looks purple.

Textbase:

Situation Model:

Figure 2.1. A two-sentence text fragment with its textbase and situation model (Kintsch, 1994, p. 295).

It should be noted that because readers construct their “textbase,” and their “situation model” individually, the representations shown in Figure 2.1 is just one of a possible model.
However, we can reasonably expect that good readers would make almost the same representations as these.

All kinds of knowledge are used in constructing a “situation model”; however, inferences, in particular, play one of the most critical roles (Kinstch, 1998, p. 106). I would like to add more explanations through analyzing some more examples.

(a) Jane could not find the vegetable and the fruit she was looking for. She became upset. (Kinstch, p. 106)

Although there are no explicit markers that connect these sentences, readers draw inferences such as “because Jane could not find those perishable foods, she became upset” and so on. This inference is generated in order to make the text coherent. Take a look at another example.

(b) Jack missed his class because he went to play golf. He told his teacher he was sick. (Kinstch, p. 106)

By reading the sentences in (b), “a situation model may be formed by the elaboration Jack lied” (Kintsch, p. 106). The first example is concerned with what we call “bridging inferences,” described as inferences “required for the coherence of the text” (van Dijk & Kintsch, 1983, p. 49); and the second example is concerned with “elaborative inferences,” described as inferences to “occur when the reader uses his or her knowledge about the topic under discussion to fill in additional detail not mentioned in the text, or to establish connections between what is being read and related items of knowledge” (van Dijk & Kintsch, p. 51). Thus, when we consider situation models constructed by readers, inference
making is one of the requisite processes. Theories and previous surveys which have dealt with inference generation processes in reading will be reviewed in the following section.

2.2 Inference Generation Processes

2.2.1 Reading Models and L2 Reading Research

In this section, I will review five reading comprehension models that have been used to intensively study inference generation processes or construction processes of situation models in the field of cognitive psychology; the constructionist theory (Graesser & Kreuz, 1993; Graesser et al., 1994), the minimalist hypothesis (McKoon & Ratcliff, 1992), the resonance model (Myers & O’Brien, 1998; O’Brien & Myers, 1999), van den Broek et al.’s (1993) framework, and the event indexing model (Zwaan, Langston, & Graesser, 1995; Zwaan, Magliano, & Graesser, 1995; Magliano, Zwaan, & Graesser, 1999). The main purpose of this section is to evaluate these models from the following perspective: whether a model makes it possible for us to obtain a fuller picture of Japanese EFL learners’ comprehension processes, while faithfully reflecting the development of L2 learners’ reading proficiency. It derives from the consideration that some models may fail to be applied to L2 reading studies, but that other models may have some potential to be applied vigorously to the L2 reading studies in the present study as well as those in the future.

This section will be divided into three parts. Firstly, I will review some models such as the constructionist theory and the minimalist hypothesis. It should be noted again that they have all been developed in the field of cognitive psychology, accumulating much empirical data of L1 reading research. Secondly, some L2 reading studies based on these reading comprehension models will be introduced. However, perhaps as a result of the innate difficulties of L2 reading research, which includes a great number of complex factors, we
can find up to the present only a small number of studies on the inference generation processes of L2 readers. However, it is expected that all should offer some insights into the inference generation processes of Japanese EFL learners as well. Therefore, I will not only review the studies that targeted EFL learners, but also the L2 studies administered to American learners of Japanese as a foreign language (JFL). Thirdly, the evaluation of the comprehension models will be attempted in terms of L2 reading research. The above three issues will be discussed below by model, one by one. The following models are not arranged in chronological order because the development of most of them was concentrated in the early 1990s.

1) Constructionist Theory

The constructionist theory developed the search-after-meaning principle, which has the following three assumptions: (a) the reader goal assumption, (b) the coherence assumption, and (c) the explanation assumption (Graesser & Kreuz, 1993; Graesser et al., 1994). The reader goal assumption states that readers attempt to construct meaning representations in accordance with their goals of reading; the coherence assumption states that readers attempt to construct meaning representations which are locally and globally coherent; and the explanation assumption states that readers attempt to construct meaning representations which explain why actions, events, and states are mentioned by the author in the text.

On the basis of these assumptions, the constructionist theory determines which inference should be an “on-line” inference and which should be an “off-line” inference, although the advocates certainly acknowledge some exceptional cases. An “on-line” inference is defined as an inference which is generated “during the course of comprehension,” whereas an “off-line” inference is defined as an inference which is
generated “during a later retrieval task but not during comprehension” (Graesser et al., 1994, p. 371). First of all, the following 13 types have been identified as “knowledge-based inferences”: (1) “referential,” (2) “case structure role assignment,” (3) “causal antecedent,” (4) “superordinate goal,” (5) “thematic,” (6) “character emotional reaction,” (7) “causal consequence,” (8) “instantiation of noun category,” (9) “instrument,” (10) “subordinate goal-action,” (11) “state,” (12) “emotion of reader,” and (13) “author’s intent.” The brief descriptions and examples are shown in Table 2.1.

Table 2.1

Brief Descriptions and Examples of Knowledge-Based Inferences

<table>
<thead>
<tr>
<th>Type of Inference</th>
<th>Brief Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1: Referential</td>
<td>A word or phrase is referentially tied to a previous element or constituent in the text (explicit or inferred).</td>
<td>In reading “…on removing the fork the eye came with it,” a reader processes the part as “fork is the referent for it.”</td>
</tr>
<tr>
<td>Class 2: Case Structure Role Assignment</td>
<td>An explicit noun phrase is assigned to a particular case structure role (e.g., agent, recipient, object, location, time).</td>
<td>In reading “the man leaned his newspaper against the sugarbowl,” a reader processes the part as “against the sugarbowl is assigned to a location role.”</td>
</tr>
<tr>
<td>Class 3: Causal Antecedent</td>
<td>The inference is on a causal chain (bridge) between the current explicit action, event, or state and the previous passage context.</td>
<td>In reading “In his haste and abstraction he stuck a pickle fork into his right eye…” a reader processes the part as “the man was careless and mis-aimed his fork.”</td>
</tr>
<tr>
<td>Class 4: Superordinate Goal</td>
<td>The inference is a goal that motivates an agent’s intentional action.</td>
<td>In reading “A Man to Whom Time was Money, and who was bolting his breakfast in order to catch a train…,” a reader processes the part as “the man wanted to get to work and earn money.”</td>
</tr>
<tr>
<td>Class 5: Thematic</td>
<td>This is a main point or moral of the text.</td>
<td>A reader processes the entire passage as “haste makes waste.”</td>
</tr>
<tr>
<td>Type of Inference</td>
<td>Brief Description</td>
<td>Example</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Class 6: Character Emotional Reaction</td>
<td>The inference is an emotion experienced by a character, caused by or in response to an event or action.</td>
<td>In reading “…the needless outlay reduced him to poverty,” a reader processes the part as “the man became sad.”</td>
</tr>
<tr>
<td>Class 7: Causal Consequence</td>
<td>The inference is on a forecasted causal chain, including physical events and new plans of agents. These inferences do not include the character emotions in class 6.</td>
<td>In reading “…on removing the fork the eye came with it,” a reader processes the part as “the man became blind in his right eye.”</td>
</tr>
<tr>
<td>Class 8: Instantiation of Noun Category</td>
<td>The inference is a subcategory or a particular exemplar that instantiates an explicit noun or an implicit case role that is required by the verb.</td>
<td>In reading “…breakfast…,” a reader processes the part as “bacon and eggs.”</td>
</tr>
<tr>
<td>Class 9: Instrument</td>
<td>The inference is an object, part of the body, or resource used when an agent executes an intentional action.</td>
<td>In reading “…the Man to Whom Time was Money had to sustain life by fishing from the end of a wharf,” a reader processes the part as “the man used a rod and reel (to fish).”</td>
</tr>
<tr>
<td>Class 10: Subordinate Goal-Action</td>
<td>The inference is a goal, plan, or action that specifies how an agent’s action is achieved.</td>
<td>In reading “…who was bolting his breakfast,” a reader processes the part as “the man grasped his fork and moved it toward his mouth.”</td>
</tr>
<tr>
<td>Class 11: State</td>
<td>The inference is an ongoing state, from the time frame of the text, that is not causally related to the story plot. The states include an agent’s traits, knowledge, and beliefs; the properties of objects and concepts; and the spatial location of entities.</td>
<td>In reading “…the Man to Whom Time was Money had to sustain life by fishing from the end of a wharf,” a reader processes the part as “fishermen are poor; the city has a wharf.”</td>
</tr>
<tr>
<td>Class 12: Reader Emotion</td>
<td>The inference is the emotion that the reader experiences when reading a text.</td>
<td>In reading “…on removing the fork the eye came with it,” a reader processes the part as “the reader is disgusted.”</td>
</tr>
<tr>
<td>Class 13: Author Intent</td>
<td>The inference is the author’s attitude or motive in writing.</td>
<td>A reader processes the entire passage as “Bierce [the author of the passage] wants to lambaste workaholics”</td>
</tr>
</tbody>
</table>

*Note. From Graesser et al. (1994, p. 375) with minor revision.*
You may recognize that various types of inferences are included in “knowledge-based inferences,” and that even that which is closely related to grammar, for example, “referential” and “case structure role assignment,” is also covered. Among these, it is predicted that the first six inferences are on-line inferences, and the following five inferences are off-line inferences. No decision is made on the last two inferences (see Table 2.2).

Table 2.2

*Predictions of On-Line Inference Processing* (Graesser et al., 1994, p. 384)

<table>
<thead>
<tr>
<th>Type of Inference</th>
<th>Constructionist Theory</th>
<th>Minimalist Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1: Referential</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Class 2: Case Structure Role Assignment</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Class 3: Causal Antecedent</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Class 4: Superordinate Goal</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Class 5: Thematic</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Class 6: Character Emotional Reaction</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Class 7: Causal Consequence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class 8: Instantiation of Noun Category</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class 9: Instrument</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class 10: Subordinate Goal-Action</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class 11 State</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class 12: Reader Emotion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class 13: Author Intent</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* X = on-line prediction.

As far as I know, there are three studies, Collins and Tajika (1996), Muramoto (2000), and Yoshida (2003), who targeted L2 learners, and used some of these inference types. Collins and Tajika were interested in instrumental inferences (Class 9 in Table 2.2), whereas Muramoto was interested in four types of inferences among them, that is, superordinate goal (Class 4), subordinate goal-action (Class 10), character emotional reaction (Class 6), and
state (Class 11). In contrast, Yoshida targeted eight inference types at first, but she combined and divided them into bridging inferences and elaborative inferences. These results will be reviewed in more detail in the following paragraphs.

Collins and Tajika (1996) examined whether different instructions would have different effects on making inferences about instruments among L1-Japanese learners of English. Their materials were 16 short sentences such as “Yasuko stirred the coffee.” The experimental group was given the instruction—to generate an instrument; and the control group was given the instruction—to read and memorize. The results of the naming task (see section 2.2.2) revealed that there was no significant difference in the scores between the experimental group and the control group. In other words, the instruction of inferring instruments while reading the sentences did not specifically activate learners’ inference generation processes. However, the recall test was in favor of the control group rather than the experimental group. The investigators concluded that instructions to L2 learners designed to enhance elaborative inference making might be an interference with a certain level of their reading comprehension.

Muramoto (2000) also targeted Japanese EFL learners, and compared the amount of inferences constructed by learners with higher proficiency to that of inferences constructed by learners with lower proficiency. Four types of inferences, superordinate goal, subordinate goal-action, character emotional reaction, and state, were selected because they appeared to frequently be generated in narrative readings (Muramoto, p. 44). The sentence verification method (Royer, 2001; see section 2.2.2) was employed. The results can be summarized into the following: (a) All the L2 learners generated inferences, including the lower L2 proficiency learners as well as the higher L2 proficiency learners, (b) however, the higher L2 proficiency learners generated inferences more frequently than the lower L2 proficiency
learners, and (c) the differences between L2 proficiency groups appeared in all of the four types of inference generation; in other words, the effect of inference types was not found in this study.

Correlation analyses among L2 working memory, L2 reading proficiency, and inferences (bridging inferences and elaborative inferences) were conducted by Yoshida (2003). The data collected by a think-aloud procedure was examined in terms of eight inference types if we follow the framework proposed by the constructionist theory, that is, referential (Class 1 in Table 2.2), causal antecedent (Class 3), subordinate goal-action (Class 10), superordinate goal (Class 4), causal consequence (Class 7), state (Class 11), thematic (Class 5), and author intent (Class 13). However, the inferences were eventually integrated into one of two categories for analysis, bridging inferences or elaborative inferences. The results showed moderate correlations between the following pairs: (a) between L2 working memory and total inferences, (b) between L2 working memory and bridging inferences, (c) between L2 working memory and elaborative inferences, and (d) between L2 reading proficiency and elaborative inferences. In turn, there was no correlation between L2 working memory and L2 reading proficiency, and between L2 reading proficiency and bridging inferences. As Yoshida admits, some of these results showing no correlation between L2 working memory and L2 reading proficiency contradict the findings of previous studies, which may be partly due to the lower validity of her test to measure the participants’ L2 reading proficiency.

When these previous studies and the inference types proposed by the constructionist theory are taken into consideration, we can conclude that this theory may provide us with a useful framework for L2 reading research in order to learn how rich mental representations are constructed by different levels of L2 proficiency learners. At the same time, it is
probable that one inference or a few inferences must be selected for one experiment because it would be difficult to deal with a number of types of inferences within one experiment. For example, Collins and Tajika (1996), and Muramoto (2000) focused on one and four inference types, respectively. In contrast, if we were to combine these inferences together without considering types, as Yoshida (2003) did, it may diminish the originality of the constructionist theory.

In sum, this theory and the inference types proposed by this theory must yield important insights into L2 reading research, especially when L2 learners’ mental representations are the critical interests of researchers. However, we also need to be patiently accumulating data if we desire to apply this theory to L2 reading comprehension.

2) Minimalist Hypothesis

The minimalist hypothesis (McKoon & Ratcliff, 1992) constitutes another major branch of existing inference generation studies. This hypothesis offers an alternative view to the constructionist on-line inferences and off-line inferences theory. The proponents of the minimalist hypothesis claim that only two types of inferences are encoded during the course of reading, that is, “those based on easily available information and those required for local coherence” (p. 441). Moreover, these inferences are called minimal inferences, or more formally, “automatic inferences,” rather than “strategic inferences.” It is assumed that these minimal inferences or “automatic inferences” provide the information that can be used for “strategic inferences.”

The relationships between the minimalist hypothesis and the constructionist theory are shown in Table 2.2, which was adopted from Graesser et al.’s (1994). “Automatic inferences” of the minimalist hypothesis generally include “referential” (Class 1), “case
structure role assignment” (Class 2), and “causal antecedent” (Class 3) of the constructionist theory, whereas “strategic inferences” of the minimalist hypothesis generally include “superordinate goal” (Class 4), “thematic” (Class 5), “character emotional reaction” (Class 6), “causal consequence” (Class 7), “instantiation of noun category” (Class 8), “instrument” (Class 9), “subordinate goal-action” (Class 10), “state” (Class 11), “emotion of reader” (Class 12) and “author’s intent” (Class 13) of the constructionist theory. Table 2.2 also indicates how the minimalist hypothesis differs from the constructionist theory in expectations of on-line inferences and off-line inferences. In the minimalist hypothesis it is predicted that three out of 13 types of “knowledge-based inferences” are generated on-line, but in the constructionist theory it is predicted that six inferences are generated on-line. Such predictions are essentially different.

To the author’s knowledge, there has been no L2 research conducted from this viewpoint. It may be due to the inherent difficulty in determining what is easily available information for general L2 learners, because there are large differences in L2 reading proficiency among EFL learners. However, this fact should not downgrade the value of what the minimal hypothesis proposes. The hypothesis makes a clear distinction between local inferences and global inferences; that is, local inferences are automatically generated, whereas global inferences are not. Furthermore, from the minimalist point of view, global inferences do not usually require local coherence; instead, if a text has a locally coherent break at some point, global inferences are made to establish local coherence. These accompanying discussions about global inferences and local inferences would surely be helpful when considering L2 reading comprehension processes. Therefore, this minimalist view will be examined again in terms of van den Broek et al.’s (1993) framework later.
3) Resonance Model

A resonance model (Myers & O’Brien, 1998; O’Brien & Myers, 1999) explains how concepts obtained by reading the earlier parts of a text or concepts within readers’ prior knowledge are activated. It assumes these concepts resonate in accordance with the degree to which they match the input; and semantic and contextual features determine how well concepts are matched against units of inputs, and vice versa. This resonance process is a passive one.

It seems clear that this model can be applied to L2 reading research. For example, we can prepare two sets of reading materials which are different in the degree of semantic and contextual overlaps within the text. We can then compare the response time to certain probes obtained after reading texts which overlap considerably with the response time to certain probes obtained after reading texts which overlap a little. This model is strongly based on readers’ bottom-up processing. However, the main purpose of this section is to find a comprehension model which enables us to observe a fuller picture of L2 reading comprehension, with a focus on inference generation. Therefore, it was determined that this model will not be employed for the present study.


Van den Broek et al.’s (1993) inference types provide us with a useful framework for explaining the mechanism of inference generation during reading. They promise to integrate the previous findings of a number of inference generation studies, which have been accumulated during L1 reading research. The inferences are categorized according to the primary source of information from which they are drawn (e.g., the text or readers’ prior knowledge) and the relational direction to the focal statement (e.g., backward or forward;
see Figure 2.2). As a result, the following six types of inferences are determined: (a) “reinstatements,” (b) “connecting inferences,” (c) “backward elaborations,” (d) “forward elaborations,” (e) “orthogonal elaborations” and (f) “associative inferences.”

![Diagram of inferences](Image)

*Figure 2.2. Types of inferences in reading (van den Broek et al., 1993, p. 171).*

These definitions, the primary sources of information, and the relational directions to the focal statement are compiled into Table 2.3. Furthermore, an example of each inference type is shown in Table 2.4, which has been partly taken from Horiba (1993, 1996). We need more explanations about these inferences, because some inferences have sources and relational direction to the focal statement that are also common to the others. It should be explained which factor it is that differentiates between “reinstatements” and “connecting inferences,” and between “forward elaborations,” “orthogonal elaborations” and
“associative inferences.”

Firstly, both “reinstatements” and “connecting inferences” are backward inferences, which connect the focal statement with the events or states that appeared earlier in the text. It is in the distance between the connected pieces of information in texts that the difference lies. “Reinstatements” connect the focal statement to information from the text that is reinstated from long-term memory,” whereas “connecting inferences” connect the focal statement to “information that was processed recently and, hence, still is attended or in short-term memory” (van den Broek et al., 1993, p. 170).

Table 2.3

**Definitions of van den Broek et al.’s (1993) Inference Types**

<table>
<thead>
<tr>
<th>Type</th>
<th>Definition</th>
<th>Source</th>
<th>Direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reinstatements</td>
<td>inferences which connect the focal statement to information from the text that is reinstated from long-term memory</td>
<td>text</td>
<td>backward</td>
</tr>
<tr>
<td>Connecting Inferences</td>
<td>inferences which connect the focal statement most directly to information that was processed recently and, hence, is still attended to or in the short-term memory</td>
<td>text</td>
<td>backward</td>
</tr>
<tr>
<td>Backward Elaborations</td>
<td>inferences which draw heavily on the reader’s general background knowledge in connecting the focal statement</td>
<td>knowledge</td>
<td>backward</td>
</tr>
<tr>
<td>Forward Elaborations</td>
<td>inferences which anticipate information that is yet to be described in the text</td>
<td>knowledge</td>
<td>forward</td>
</tr>
<tr>
<td>Orthogonal Elaborations</td>
<td>inferences which concern the activation of information that is implied by and coexistent with the information in the focal statement</td>
<td>knowledge</td>
<td>forward</td>
</tr>
<tr>
<td>Associative Inferences</td>
<td>inferences which activate information that is associated with a focal statement</td>
<td>knowledge</td>
<td>forward</td>
</tr>
</tbody>
</table>

*Note.* This table has been summarized by the author, based on van den Broek et al. (1993). Directions = relational directions to the focal statement. Knowledge = readers’ prior knowledge.
Note, however, that van den Broek et al. do not discuss what kind of units (e.g., words or clauses) or how many units (e.g., 7 units, or 1 or 2 sentences) may be contained in either the long-term or the short-term memory. Thus, since their model is theoretically based, researchers who use this framework should themselves determine these issues. Statements related to the distinction between “reinstatements” and “connecting inferences” can be seen in McKoon and Ratcliff (1992), and Kosaka (2000), which will be referred to in the definitions later.

Table 2.4

<table>
<thead>
<tr>
<th>Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reinstatements</td>
<td>Oh, that’s the bamboo thing the thief had in his mouth. (at the sentence, “there was something like a stick standing straight and swaying…” (Horiba, 1993, p. 60)</td>
</tr>
<tr>
<td>Connecting Inferences</td>
<td>Perhaps the baby saw the face that the thief made. (at the sentence, “Then the baby smiled.”) (Horiba, 1996, p. 471)</td>
</tr>
<tr>
<td>Backward Elaborations</td>
<td>Perhaps, the face looked funny. (at the sentence, “Then the baby smiled.”) (Horiba, 1996, p. 471)</td>
</tr>
<tr>
<td>Forward Elaborations</td>
<td>He’s going to go down there. (at the sentence, “The man stuck out his tongue.”) (Horiba, 1996, p. 471)</td>
</tr>
<tr>
<td>Orthogonal Elaborations</td>
<td>Brightening of night, I suppose, means dawn. (at the sentence, “Oh oh, the day is breaking.”) (Horiba, 1996, p. 471)</td>
</tr>
<tr>
<td>Associative Inferences</td>
<td>This reminds me of a ninja. (at the sentence, “Once upon a time, a thief sneaked into the attic of a house.”) (Horiba, 1996, p. 471)</td>
</tr>
</tbody>
</table>

Note. This table has been summarized by the author, based on the protocols obtained by Horiba’s (1993, 1996) think-aloud technique.

1. The 10th sentence before this sentence was “The thief put a piece of bamboo into his mouth…”
2. Before the sentence, the following sentence was written: “The man stuck out his tongue.”

Secondly, “forward elaborations,” “orthogonal elaborations” and “associative inferences” have information sources and relational directions to the focal statement in
common; that is, the information is retrieved from readers’ prior knowledge, and forward is the indicated direction of these inferences. The differences appear to be found in what kind of information would be elicited from readers’ prior knowledge as follows: (a) “Forward elaborations” concern predicting developments in the plots of a story or predicting the events which would occur in a text, (b) “orthogonal elaborations” concern the information coexistent with the focal statement, and are restricted not only to verbal forms but also include visual forms such as spatial information (van den Broek et al., 1993, pp. 170-171), and lastly, (c) “associative inferences” are obtained through an activation of the connection networks; therefore, they need not be concerned with a story’s plot (for an example, see Figure 2.3).

![Figure 2.3. A fragment of the associative net for BANK (Kintsch, 1988, p. 165).](image)

The participants of Horiba (1993, 1996), who used van den Broek et al.’s (1993) categorization, were both American JFL students. The effects of text coherence, language proficiency and different narrative stories on backward inference generation were studied by Horiba (1993). A backward inference was the concept integrating van den Broek et al.’s “reinstatements” and “connecting inferences.” As a result, interaction between text
coherence and language proficiency was found. Native speakers of Japanese generated more backward inferences than American students of Japanese, during the reading of Japanese texts. In addition, only the results of native speakers of English revealed a difference in the amount of backward inferences between reading a coherent text and incoherent text written in English. The other JFL groups did not show such a tendency.

In contrast, Horiba (1996) faithfully adopted van den Broek et al.’s (1993) category of inference types. How American JFL learners allocate their cognitive resources to multiple processes was the focus of her study. The comments of her participants, who were engaged in a think-aloud procedure, were divided into the following nine processes: (a) “graphomorphemic/graphophonemic analysis,” (b) “word recognition,” (c) “syntactic/semantic analysis of sentence,” (d) “backward inferences,” (e) “elaborative inferences,” (f) “predictive inferences,” (g) “general knowledge and associations,” (h) “comments on text structure,” and (i) “comments on own behavior.” Thus, the processes labelled as (d), (e), (f) and (g) were inferences. The differences between Horiba’s category and van den Broek et al.’s category were as follows: In Horiba, (1) van den Broek et al.’s “reinstatements” and “connecting inferences” were combined into “backward inferences”; (2) van den Broek et al.’s “backward elaborations” and “orthogonal elaborations” were combined into “elaborative inferences”; (3) van den Broek et al.’s “forward elaborations” were the same as “predictive inferences”; and (4) van den Broek et al.’s “associative inferences” were the same as “general knowledge and associations” (see Figure 2.2). The results were as follows: (a) Firstly, the learners of higher L2 proficiency generated as many “backward inferences” and “predictive inferences” as L1 readers; in contrast, the learners of lower L2 proficiency hardly generated any of these inferences, and (b) secondly, with regard to “elaborative inferences” and “general knowledge and associations,” there was no
difference in the amounts produced by each group.

In sum, Horiba (1993, 1996) revealed by using van den Broek et al.’s framework that there are different inference generation processes at work among different levels of L2 proficiency groups. Although a difference can be found between Horiba’s studies and the present study in that the target languages of the learners is JFL rather than EFL, this framework should function well for Japanese EFL learners as well. Therefore, with reference to some previous studies, let us more conscientiously examine the inference types proposed by this model.

As was mentioned above, the difference between “reinstatements” and “connecting inferences” is the distance from the focal point to the previously processed point. For “reinstatements,” van den Broek et al. (1993) state only that the previously processed point remains in the long-term memory, whereas with “connecting inferences” the previously processed point remains in the short-term memory. Therefore, the criteria for dividing these inferences should be determined by researchers. The first instance is McKoon and Ratcliff (1992), who assumed “that a set of two or three sentences is locally coherent if it makes sense on its own or in combination with easily available general knowledge” (p. 444). We could easily infer from these statements that a set of four or more sentences is globally coherent if it makes sense on its own or in combination with easily available general knowledge. The second instance is Kosaka (2000), who investigated the relationship between preschoolers’ L1 working memory capacity and L1 text comprehension. As one of the text comprehension tests, two types of inferential questions were prepared, one type to assess global inferences, and the other to assess local inferences. Three to seven sentences were inserted between the connected sentences for the former type of questions, whereas one or no sentence was inserted between the connected sentences for the latter type of
questions.

We should also recognize that some of the inferences proposed by this framework can be measured by using what is called an on-line measure, but cannot be measured by using what is called an off-line measure. A supplementary explanation of this will be included in section 2.2.2.

5) Event Indexing Model

The event indexing model (Zwaan, Langston et al., 1995; Zwaan, Magliano et al., 1995; Magliano et al., 1999) considers multiple levels of dimensions of situation models such as “characters and objects,” “temporality,” “spatiality,” “causality,” and “intentionality.” This model also attempts to explain how these situational dimensions are monitored and updated by readers. To be more specific, it assumes the monitoring of “what characters are involved in a story, what is happening to them, what their goals are, and how they are achieving those goals, all within a narrative time and space” (Magliano et al., p. 220); and that if changes in these dimensions are to occur, the situation models are urged to be changed alongside.

Regarding an update of a situation model, the following assumptions are made:

Breaks in one or two dimensions may not require that readers abandon the model of the current situation. However, when there are breaks in several dimensions, as is the case when a new episode starts, readers must construct a model for the new situation. (Magliano et al., 1999, pp. 224-225)

Furthermore the relationship between situational continuities and inference generation is indicated in the literature. According to the model, if a situation in a story appears to be
continuous it gives readers the cue to use the prior context to comprehend a text; in turn, if a
situation in a story appears to have broken down, it gives readers the cue to not use the prior
context but to use their prior knowledge to make the text coherent.

The second experiment in Zwaan and Brown (1996), in speculating about how L2
learners (American students of French as a foreign language) construct situation models,
employed a verb-clustering task (see section 2.2.2) explored by Zwaan, Langston et al.
(1995). It was reported that a shift in “temporality,” “causality” and “characters and objects”
in a story affected L2 learners’ comprehension; however a shift in “spatiality” and
“intentionality” did not affect L2 learners’ comprehension. To the author’s knowledge, this
is the only research which has examined L2 learners’ comprehension under the assumption
of the event indexing model. If we target narrative comprehension of L2 learners only, and
if a suitable reading material for this experiment is found, this model seems to be a
promising one for future research in this area.

The aforementioned models are all summarized in Table 2.5. You will have noticed
that there have been only a few studies conducted with the aim of explaining L2 learners’
inference generation processes. Moreover, if we consider only the research on Japanese
EFL learners’ inference generation processes, there are only Collins and Tajika (1996),
Muramoto (2000), and Yoshida (2003); the number of previous studies is drastically
reduced.

However, if I summarize the results of these previous studies, the following points
should be mentioned. First, learners with higher L2 proficiency tended to infer more often
than learners with lower L2 proficiency. This was reported by Horiba (1993, 1996),
Muramoto (2000), and Yoshida (2003). Second, when I concentrate my attention on the
inference types of the previous studies—to be specific, in what inference types the difference
can be seen between learners of different levels of L2 proficiency—the difficulty in compiling these results becomes evident. This may be in part because they differed considerably in inference types and the factors in which they were interested.

Table 2.5

**Summary of Reading Models**

<table>
<thead>
<tr>
<th>Reading Model</th>
<th>Feature</th>
<th>Previous Study in L2 Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Constructionist Theory</td>
<td>As many as 13 “knowledge-based inferences” are identified.</td>
<td>Collins &amp; Tajika (1996)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Muramoto (2000)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yoshida (2003)</td>
</tr>
<tr>
<td>2. Minimalist Hypothesis</td>
<td>Automatic inferences and strategic inferences are differentiated.</td>
<td></td>
</tr>
<tr>
<td>3. Resonance Model</td>
<td>Is interested in explaining how concepts obtained by reading or concepts within readers’ prior knowledge are activated.</td>
<td></td>
</tr>
<tr>
<td>4. Van den Broek et al.’s Model</td>
<td>Inferences are categorized according to the primary source of information from which they are drawn and the relational direction to the focal statement.</td>
<td>Horiba (1993, 1996)</td>
</tr>
<tr>
<td>5. Event Indexing Model</td>
<td>Multiple levels of dimensions (i.e., “characters and objects,” “temporality,” “spatiality,” “causality,” and “intentionality”) are identified.</td>
<td>Zwaan &amp; Brown (1996)</td>
</tr>
</tbody>
</table>

*Note.* Underlined studies targeted Japanese EFL learners.

For example, Collins and Tajika (1996) indicated that the instruction to infer instruments implicitly stated in sentences did not have any positive effect on L2 readers’ comprehension; and Muramoto indicated that his L2 learners with higher L2 proficiency generated more inferences concerning superordinate goal, subordinate goal-action, character emotional reaction, and state. Collins and Tajika’s inference type, that is, instrument, is different from any other types of Muramoto. Moreover, if we separate these five different inference types into either “a bridging inference” or “an elaborative inference,” two types (i.e.,
superordinate goal and character emotional reaction) can be categorized as “a bridging inference” and three types (i.e., instrument, subordinate goal-action, and state) can be categorized as “an elaborative inference.” That is, even if I attempt to summarize Collins and Tajika’s results and Muramoto’s results no consistent feature can be shown. Moreover, Zwaan and Brown (1996) employed Zwaan and his colleagues’ original frameworks such as “characters and objects,” “temporality,” and so forth. These dimensions cannot be interpreted from inference types used by Collins and Tajika, and Muramoto. Furthermore, Yoshida (2003), and Horiba (1993, 1996) showed somewhat contradictory results; that is, Yoshida did not reveal any meaningful relationship between her learners’ L2 proficiency and the amount of “bridging inferences,” whereas Horiba (1993, 1996) did. All of these facts do not provide us with an entire picture of L2 learners’ mental representations other than the fact that L2 learners of higher L2 proficiency tend to infer much more often than those of lower L2 proficiency. In sum, taking these mixed results of previous studies into consideration, we need to carefully decide the inference types to be studied, and carefully consider how we should control some factors such as instructions beforehand as in Collins and Tajika. Also, we need to extend this kind of research. Following this, one of the main purposes of the present study was set; that is, to investigate the inference types which differentiate Japanese EFL learners of higher L2 proficiency from those of lower L2 proficiency.

Furthermore, among these models van den Broek et al.’s (1993) framework will be used for the present study. This is because this framework appears to be more comprehensible than the other models, with regard both to inferences whose sources are textual information and inferences whose sources are readers’ background knowledge, and furthermore, targets both narrative comprehension and expository comprehension. In
addition, their categorization of the inferences is likely to be suitable for research, as shown by Horiba’s (1993, 1996) studies.

Thus, Study 1 in this paper aims to examine the inference generation processes of Japanese EFL learners with reference to van den Broek et al.’s (1993) study. In Study 2, a particular inference will be determined among “knowledge-based inferences” proposed by the constructionist theory in accordance with the results of Study 1, and the mechanisms will be delved into more deeply. The target inference type in Study 2 will be mentioned in more detail in section 4.1.

2.2.2 Data Collection Methods for Inferences

Inference generation processing is, like other reading processes, an internal activity which occurs within an individual. This makes the process difficult to be observed. Therefore, data collection methods for inferences have been extensively explored in the domain of cognitive psychology. The following are some popular data collection methods: (a) a written recall task, (b) a think-aloud task, (c) a sentence verification task (Royer, 2001), (d) a verb-clustering task (Zwaan, Langston et al., 1995), (e) a naming task, and (f) a lexical decision task. A brief description is in Table 2.6. Furthermore, the terms, “on-line measures” and “off-line measures” are frequently used with regard to the research whose focus is on inference generation. You may remember that on-line inferences and off-line inferences were discussed in the previous section. “On-line measures” and “off-line measures” are applied to these inferences; that is, the interest of the former measure is on-line inferences, whereas the interest of the latter measure is off-line inferences.

In this section, I will focus on a recall procedure and a think-aloud procedure, because both of these will be employed in the present research. Their potentiality will be discussed
as well as their limitations. It should also be noted that the discussion about a recall procedure and a think-aloud procedure in this section is limited to the measures which assess inference generations; therefore, for example, a written recall task as an alternative testing method to multiple-choice questions, which is for assessing examinees’ propositional understanding, will not be mentioned in the following paragraphs.

Table 2.6

*Tasks for Data Collection*

<table>
<thead>
<tr>
<th>Task</th>
<th>Overview</th>
</tr>
</thead>
<tbody>
<tr>
<td>Written Recall Task</td>
<td>After reading a passage or passages, participants write down whatever they can remember without reference to the passage(s).</td>
</tr>
<tr>
<td>Think-Aloud Task</td>
<td>While participants are reading a passage, they speak aloud about whatever they have in their mind.</td>
</tr>
<tr>
<td>Sentence Verification Task</td>
<td>After reading a passage, participants are presented four types of sentences, and required to judge whether they have appeared in the passage without reference to the passage. The results are interpretable from the perspective of mental representations (Kintsch, 1994, 1998; van Dijk &amp; Kintsch, 1983).</td>
</tr>
<tr>
<td>Verb-Clustering Task</td>
<td>After participants read a passage, several verbs are given, which have appeared in the previously read passage. Participants are required to cluster them into some groups, without reference to the passage.</td>
</tr>
<tr>
<td>Naming Task</td>
<td>After reading a passage, participants are required to read aloud a word on a computer screen. The time from which a word is presented on a screen and to the time at which participants start their utterance is measured.</td>
</tr>
<tr>
<td>Lexical Decision Task</td>
<td>After participants read a passage, a word or a nonword is presented on a computer screen. Participants are required to judge whether it is a word or nonword as quickly, and as accurately as possible.</td>
</tr>
</tbody>
</table>

*Note.* It should be noted that this is not a comprehensive list.

A written recall procedure has been extensively used in the L2 reading studies (e.g., Barry & Lazarte, 1998, 2000; Hammadou, 1991; Horiba, 1993, 1996) as well as L1 reading
studies. One of the advantages of this procedure is that whereas most of the methods are assumed to be administered on an individual basis with a personal computer used, it is possible to conduct a written recall procedure by means of a paper and pencil task. This makes it possible for researchers to collect valuable data from a number of participants at one time. Another advantage is that there is no limit to types of inferences, and no limit to the amount of inferences produced by the participants. In other words, the number and type of inferences made is totally reliant on the readers. Therefore, this technique seems to be suitable for exploratory study rather than confirmatory study (Hosama, 1993).

As the other methods, a written recall procedure has some disadvantages or limitations. If this procedure is compared to the other methods such as multiple-choice questions, in which the participants are allowed to answer with reference to texts, we can recognize that this method is related to the participants’ memory in part. Another disadvantage is a parallel to the factor previously mentioned. The fact that the types of inferences and the amount of inferences are not determined in advance requires the researcher to use some effective frameworks to analyze the protocols. As is often stated about a think-aloud procedure as well, although a vast amount of rich data can be obtained by this procedure, how the data is connected to a certain hypothesis testing largely depends on the fine sense of the researchers. However, confidence in its value can still be felt because we are blessed with plentiful theories and some previous studies regarding inference generation (see section 2.2.1).

When we use a written recall procedure as one of the data collection methods and van den Broek et al.’s (1993) framework for an analysis of the protocols, there are some issues which need to be kept in mind. Van den Broek et al. proposed the following six inference classes, with a focus on their sources and the relational directions to the focal statement: (a)
“reinstatements,” (b) “connecting inferences,” (c) “backward elaborations,” (d) “forward elaborations,” (e) “orthogonal elaborations” and (f) “associative inferences” (see Figure 2.2). However, it can be assumed that in the context of recall, judgments of elaborative inferences may be much more difficult than some on-line measurements such as think-aloud procedures. For example, a written recall task is conducted after participants complete reading; therefore, even if they may have been predicting the events later in the plot (that is, they may have generated forward elaborations) while reading, those predictions would have been already resolved by the time when they are engaged in the recall task. This indicates one of the limitations that are inherent in off-line measures.

Thus, the present study created a new category named “elaborative inferences” as well as “reinstatements” and “connecting inferences.” “Backward elaborations,” “forward elaborations,” “orthogonal elaborations” and “associative inferences” which appeared in van den Broek et al. (1993) are all integrated into “elaborative inferences.” Furthermore, adopting the terms used by McKoon and Ratcliff (1992), I will consistently use “global connecting inferences” instead of “reinstatements,” and “local connecting inferences” instead of “connecting inferences” in the following sections. Frequency of inferences found in protocols was counted by type.

There are some studies which focused on L2 readers’ inference generation during reading, and used a think-aloud procedure (e.g., Horiba, 1993, 1996; Yoshida, 2003, 2005; Zwaan & Brown, 1996). Although there are still some discussions about this topic, a think-aloud protocol has been considered as one of the best on-line methods by a number of researchers. The theoretical development of this methodology largely depends on Ericsson and Simon’s (1984) volume, in which they describe the mechanisms of think-aloud production. According to that study, what is elicited from the participants by this procedure
is “the contents of short-term memory (i.e., to the information that is in the focus of attention)” (Ericsson & Simon, p. 221). Van Someren, Barnard, and Sandberg (1994) basically follow this idea, but assume “working memory” as an alternative to Ericsson and Simon’s “short-term memory.” Figure 2.4 was cited from van Someren et al.’s memory model of a think-aloud procedure. Therefore, the participants are not expected to interpret their behavior, nor does the procedure change the process of the orienting tasks—“reading” in the present context—at all.

![Memory model](van Someren et al., 1994, p. 19).

There are advantages and disadvantages to using a think-aloud procedure. Some are the same as with a written recall procedure, because they both adopt protocol analyses. In contrast, others are completely different between these procedures, because a think-aloud is an on-line measure whereas a written recall is an off-line measure. The first advantage is that a think-aloud task does not impose a memory load on the participants, which is what constitutes one of its superiorities to a written recall task. As was mentioned above, the participants who are thinking aloud during reading are only required to verbalize what is in
their short-term memory or working memory. The second advantage is that a think-aloud procedure makes it possible for researchers to observe the change in participants’ processing. In other words, we are able to compare their processing before a particular operation starts with their processing after a particular operation starts by using this method. This was the decisive advantage when selecting a data collection method for my second piece of research, Study 2. The second research was conducted to investigate how an inference question affects L2 learners’ inference generation processes; therefore, my interest was to examine the learners’ processes before and after the question was given. Consequently, this think-aloud procedure enabled me to observe these changes.

At the same time, a think-aloud procedure assigns exacting tasks to the researchers. These are summarized in the following:

1. A think-aloud procedure is carried out on an individual basis, and it makes it impossible to collect a lot of data at once.

2. During these procedures, the researchers are required to carefully observe the participants’ thinking aloud procedures in order to offer them help when they have trouble. However, it is important to create an ideal environment in which participants feel as relaxed as if he or she were in the room alone (Harada, 1993).

3. Careful preparations are prerequisite to this procedure since most of the participants are not familiar with this method. Ericsson and Simon (1984) warn the researchers not to ask their participants the following points: (a) a generalized description of their processing across trials, (b) explanations of their processing, and (c) descriptions about the reasons why they are carrying out the process. Prior to the data collection, it would be preferable that a researcher does a demonstration in front of a participant, and a
participant practices thinking aloud with some materials.

4. In general, it is necessary to transcribe the participants’ verbalized protocols prior to scoring. This is for accurate judgments by scores. Although some instruments have been developed for transcription such as a transcriber (e.g., SONY B1-85), this task still consumes an enormous amount of time. When more than one language is mixed in protocols such as English and Japanese, it takes more time.

5. As with the other protocol analyses, a researcher who conducts a think-aloud procedure needs theories or frameworks within which to analyze the data.

These reveal some of the difficulties in employing this method for research. Despite these, however, a think-aloud procedure has crucial advantages: namely that it can investigate what an off-line measure cannot.

2.3 Effects of Questions on Inference Generations

In the context of the L2 reading classroom, questions are frequently used (Koiso, 1996; Shimada, 1992; Tanabe, 2000; Temma, 1989). Some of the English language textbooks adopted for Japanese senior high school students of English in Japan contain one or more reading questions on the bottom of the pages (e.g., Ando et al., 2000; Okuma et al., 1998; Yamamoto et al., 2000), and many of the EFL or ESL (English as a second language) reading textbooks include a number of reading questions (e.g., Hill, 1985a, 1985b, 1985c; Mikulecky & Jeffries, 1998). Moreover, questions are sometimes generated by students themselves (e.g., Akita, 1998; Andre & Anderson, 1978-79; Frase & Schwartz, 1975; Gillespie, 1990; Satoh, 2005; Singer & Donlan, 1982). Thus, to study how a reading question facilitates learners’ comprehension is practical and would be beneficial for future
L2 reading instruction. Question types typically observed as part of L2 reading instruction are summarized by Ikeno (2000; see Table 2.7). He listed six conditions: context, timing, formats, language, modes, and question makers. Following this categorization, my focus is limited to inferential questions (with regard to context) which are made by a teacher (with regard to question makers). In the case of experimental research, a researcher (instead of a teacher) would pose a question to a participant (instead of a student).

Table 2.7

<table>
<thead>
<tr>
<th>Condition</th>
<th>Sample of Choices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Context</td>
<td>factual, inferential, or evaluative</td>
</tr>
<tr>
<td>Timing</td>
<td>before, during, or after reading</td>
</tr>
<tr>
<td>Formats</td>
<td>True/False, matching, or open-ended</td>
</tr>
<tr>
<td>Language</td>
<td>L1 for both questions and answers, L1 for questions and L2 for answers, L2 for</td>
</tr>
<tr>
<td></td>
<td>questions and L1 for answers, or L2 for both questions and answers</td>
</tr>
<tr>
<td>Modes</td>
<td>oral or written</td>
</tr>
<tr>
<td>Question Makers</td>
<td>a teacher or a student</td>
</tr>
</tbody>
</table>

*Note.* This table has been summarized by the author, based on Ikeno (2000; translation mine).

The aim of this section is to review some of the previous studies which have inspected the effects of inference questions on reading comprehension by way of experiment, and to find the possibility of changing L2 learners’ comprehension processes by using inference questions. The number of previous studies is severely limited because the area of my interest is considerably narrow. In other words, I only selected studies administered to improve readers’ comprehension (not to assess students’ reading ability), and studies which examined the effects of inference questions on reading comprehension. Therefore, including
L1 reading research, Friedman and Rickards (1981), van den Broek, Tzeng, Risden, Trabasso, and Basche (2001), and Ikeno (1996) will be reviewed in more detail in this section. The former two studies were conducted in the L1 reading context, whereas the latter was conducted in the L2 reading context. Furthermore, Ikeno’s focus was on readers’ text structure prediction, an interest which was slightly different from that of the present study. However, considering that both target L2 learners’ reading comprehension and higher-level reading processes (see section 2.1), it was determined that this research would also be mentioned in this section.

1) Friedman and Rickards’ (1981) Study

Friedman and Rickards (1981) conducted a study of 210 college students taking a psychology course at university to clarify the effects of questions on L1 reading comprehension. Their interests were on (a) what types of questions among “verbatim” questions, “paraphrase” questions, and “inference” questions would improve reading comprehension most significantly, (b) when questions should be given to make them function most effectively, and (c) whether providing answer keys after reading helps reading comprehension. The dependent variable was scores in a delayed recognition test with 48 multiple-choice questions, and the independent variables were the above-mentioned three factors. Their results are mentioned in the following paragraph one by one, although my interest lies in the first factor, viz., question types.

Reading questions used in that study were multiple-choice questions of three options. Each option was constituted of one sentence. A participant was given either one of the following three types of questions consistently: (a) a “verbatim” question, with which a participant was required to choose the identical sentence to that in the passage, and the stem
was “Which of the following is an exact duplication of one sentence in the paragraph you just read?,” (b) a “paraphrase” question, with which a participant was required to choose the sentence whose meaning was the same as the content of the passage, and the stem was “Which of the following has the same meaning as a sentence in the paragraph you just read?,” and (c) “inference” questions, with which a participant was required to choose the sentence whose content could be appropriately inferred from the passage, and the stem was “Which of the following is a correct inference based on a sentence in the paragraph you just read? An inference requires you to extend your thinking beyond the information contained in the passage.” The target parts which the questions asked for were exactly same among these three conditions.

The results revealed an impressive dimension. In the following assessment phase, the group given “inference” questions received the highest score; the group given “paraphrase” questions received the second highest score; and the group given “verbatim” questions received the lowest score on average. Furthermore, with regard to even the unfocused parts of questions, both of the groups given “inference” questions or “paraphrase” questions received a higher score than the other group. It can be concluded from this that indirect as well as direct effects were observed. With regard to the other factors, no consistent effect of the timing of questions was seen, however, the effectiveness of making available the correct answers after completing a series of questions was found.

2) Van den Broek et al.’s (2001) Study

The other L1 study, van den Broek et al. (2001) focused on inference questions only, and studied the timing (i.e., during-reading and after-reading) and readers’ school grades (i.e., 4th-grade, 7th-grade, 10th-grade, and a college level). Inference questions were
constructed to tap the contents: (a) “why the character performed an action,” (b) “what he [the character] did to attain the goal,” and (c) “how he [the character] attained a successful outcome” (van den Broek et al., p. 523). For assessment, a written recall method was employed.

It was revealed that the effectiveness of inferential questions was found in the group of college students, but not in the group of 4th-, 7th-, and 10th-grade students. To be more specific, college students given questions during the reading of a passage obtained higher scores than those given questions after reading a passage; they also obtained higher scores than those who were not given any questions. In contrast, with regard to the results of 4th graders, 7th graders, and 10th graders, there was no difference in recall scores between the group given questions and the group given no questions; indeed, the group without questions recorded a higher score than the group with questions on some occasions. Furthermore, timing did not reveal an interaction with students’ grade levels. That is, within the same time frame, the college students recalled the most, the 10th-grade students the second, the 7th-grade students the third, and the 4th-grade students the least.

We should note one more important result attained by that study. The above-mentioned results are concerned with the effects of questions on the information contained in the questions and answers, that is, “targeted” information. In contrast, the effects of the questions on the information which was not inquired of by them, that is, “untargeted” information, offered different results. The effects of questions were not found on the 10th-graders’ and college students’ comprehension of the “untargeted” information; and what is more, somewhat negative effects of the questions were found on the 7th-graders’ and 4th-graders’ comprehension of the “untargeted” information. These results seem to contradict Friedman and Rickards’ (1981) report because Friedman and Rickards observed
indirect effects of inference questions in their research. In conclusion, we have mixed results in terms of the effects of inference questions on the “untargeted” information.

3) Ikeno’s (1996) Study

Ikeno’s (1996) target was text-structure-guiding questions, and he investigated how these questions would affect L2 learners’ reading comprehension. A total of 152 Japanese college students of English participated in the research. Five questions were attached to each passage under the experimental conditions, whereas there was no question under the control conditions.

It was reported that the group that was given text-structure-guiding questions performed better on a recall task than the group that was not given any questions. Thus, his first hypothesis that text-structure-guiding questions will help L2 readers improve their comprehension, was confirmed. However, his second hypothesis that the degree of improvement is greater for a linguistically difficult text than for a linguistically easy text was rejected; that is, no difference was found in the effects between the two texts which had different readabilities.

These previous studies imply important suggestions for the present research. First, considering Friedman and Rickards’ (1981) results, “inference” questions appear to function more effectively than the other types of questions such as “paraphrase” questions and “verbatim” questions. It is possible to interpret these results from the point of view of the mental representations constructed by readers (see section 2.1). Their “verbatim” questions correspond to “surface” memory; “paraphrase” questions correspond to “propositional text base”; and “inference” questions correspond to “situation model.” It can be concluded that,
among their questions, those which were designed to tap the readers’ deepest level of understanding brought about the most desirable effects on reading comprehension. The effects of these “inference” questions were found not only in Friedman and Rickards (1981), but also in van den Broek et al. (2001). Van den Broek et al.’s study did not attempt a comparison between “inference” questions and the other types of questions; however, they compared the comprehension when “inference” questions were given with the comprehension when “inference” questions were not given. Although an interaction between the effects and readers’ grades was revealed, this study also indicates the potentiality of “inference” questions.

The second point suggested by the previous studies is that whereas the above two studies were based on L1 reading research, Ikeno (1996) indicated the questions in relation to the higher-level processing, that is, text structure prediction, had a positive effect on L2 learners’ reading comprehension. It can be inferred from his study that “inference” questions might function effectively for L2 readers as well as L1 readers.

Third, as was mentioned in the last paragraph of the review of van den Broek et al. (2001), whether “inference” questions affect the comprehension of the “untargeted” information is still unknown. The mixed results may depend on the different degrees of proximity to the questions. To be more specific, these results still appear to indicate the possibility that if we focus on the information not overtly targeted by the questions but slightly related to the questions, there may be positive effects on “untargeted” information as well as “targeted” information.

In sum, many important issues still remain to be investigated regarding inference questions. Let us look at inference questions developed in the field of L1 reading research. In the development of question-answering methodology, which was explored by Graesser
and his colleagues (e.g., Graesser & Clark, 1985; Graesser & Goodman, 1985; Graesser et al., 1994; Long, Golding, Graesser, & Clark, 1990), the following question types have been identified as effective questions for observing readers’ inference generation: (a) “why” questions, (b) “how” questions, and (c) “what-happened-next” questions.

These questions proposed by question-answering methodology have some connections with some of the “knowledge-based inferences” of the constructionist theory (see Table 2.1). First, “why” questions are designed to elicit inferences of “superordinate goals” and “causal antecedents”; that is, regarding an aim which a character in a text wants to achieve, or causes of a certain event in a text. Second, “how” questions are designed to elicit inferences of “subordinate goals-actions” or “causal antecedents events,” that is, regarding the procedures through which a character attains a goal, or causes of a certain event in a text. Third, “what-happened-next” questions are designed to elicit inferences of “causal consequences” that is, regarding the following story plots. The present study will employ one of these questions, which have been developed in Graesser and his colleagues’ question-answering methodology, in accordance with the inference type targeted in Study 2. The details will be described in section 4.1.

2.4 Present Study

As was reviewed in the previous sections, there have been some studies of Japanese EFL learners’ inference generation processes such as Collins and Tajika (1996), Muramoto (2000) and Yoshida (2003). They revealed a part of Japanese EFL learners’ inference generation processes, as well as offering some insights into those processes. For example, it was revealed that there is a difference in the amount of Graesser et al.’s (1994) several types of inferences between higher L2 proficiency learners and lower L2 proficiency learners, and
that there are reliable correlations among a reader’s working memory span, L2 reading proficiency, and some inference types. However, it is still unclear what mental representations are constructed by Japanese EFL learners, particularly when we differentiate between their “local connecting inferences” and “global connecting inferences.” Also, until now, how well inference questions can change Japanese EFL learners’ inference generation processes has not been investigated by means of experiments, although it must be acknowledged that we also found only a few in the field of L1 reading.

Considering these points, I have determined the following issues to be investigated and addressed in the present studies:

1. What mental representations are constructed by Japanese EFL learners? To be more specific, if we follow the van den Broek et al.’s (1993) categorization of inferences, what types and how many inferences do Japanese EFL learners generate while reading?
2. Are there differences or similarities in inference generation in reading between Japanese EFL learners of higher and lower L2 proficiency?
3. Do “inference” questions improve Japanese EFL learners’ inference generation?
4. Do “inference” questions also affect the propositional understanding of Japanese EFL learners?

It should be noted that these questions are merely a start. As we proceed with the present studies, some questions will be developed or updated. Therefore, research hypotheses will be presented before beginning each part of the research, which is placed in the first sections of Chapters 3, 4, 5, and 6.
Endnotes

1. An earlier version of a part of this section has been published in Shimizu (2003b).

2. Ericsson and Simon (1984) seem to assume the two-component model consisting of short-term memory and long-term memory. However, they discuss the control of attention at the same time.
3. Study 1: Mental Representations of EFL Learners of Different L2 Proficiency, With a Focus on Inferences

3.1 Purposes and Research Hypotheses

In the first study, I will investigate the mental representations constructed by Japanese EFL learners, with a particular focus on their inference generation. This research aims to provide fundamental data for the area of EFL reading research which pursues sophisticated reading models of EFL learners. As is mentioned in section 2.1, three different levels of mental representations are assumed in reading, “surface” memory, “textbase,” and “situation model.” The relationship between reading comprehension and mental representations is as follows: To have comprehended an entire text is to have constructed a coherent “situation model” successfully in our memory. It should also be noted that inference making is largely related to the construction of a “situation model.”

Study 1 basically employs van den Broek et al.’s (1993) inference categorization, and intends to determine the types and numbers of inferences that Japanese EFL learners would make. Some modifications were made to van den Broek et al.’s (1993) categorization because of methodological limitations (see section 2.2.2). The inferences which would be targeted in this study were then finally determined. The definitions of these inferences are listed below.

(a) “Local connecting inferences” are defined as inferences that connect two pieces of information written within one or two adjacent sentences in a text.

(b) “Global connecting inferences” are defined as inferences that connect two pieces of separate information across three or more sentences in a text.

(c) “Elaborative inferences” are defined as inferences that elaborate information in the text
via the use of readers’ prior knowledge. Elaboration, for example, takes the form of why, when, where, and how a certain character in a story performed a given action, although this was not explicitly stated in the passage.

The difference between “local connecting inferences” and “global connecting inferences” lies in the distance between the focal point and the related point which is to be connected by the inferences. The line for differentiating these inferences was set with reference to McKoon and Ratcliff’s (1992), and Kosaka’s (2000) criterion (see section 2.2.1). However, you should note that this criterion was only set for practical reasons. Moreover, the inferences “backward elaborations,” “forward elaborations,” “orthogonal elaborations” and “associative inferences” proposed by van den Broek et al. (1993) are all integrated into “elaborative inferences” in the present study (see section 2.2.2).

Considering these, the following hypotheses have been formulated. These are based on the assumption that learners of higher L2 proficiency will be able to make more coherent and richer “situation models” than learners of lower L2 proficiency.

Hypothesis 1A. The amount of “local connecting inferences” generated by higher L2 proficiency learners is greater than that generated by lower L2 proficiency learners.

Hypothesis 1B. The amount of “global connecting inferences” generated by higher L2 proficiency learners is greater than that generated by lower L2 proficiency learners.

Hypothesis 1C. The amount of “elaborative inferences” generated by higher L2 proficiency learners is greater than that generated by lower L2 proficiency learners.
3.2 Method

3.2.1 Participants

The participants in this study were 101 first-year and second-year students enrolled in freshman level EFL courses at a national university in Japan. Most had received formal instruction in English for six and a half years in Japan. However, eight students’ data was removed; some were absent from the class at which a L2 proficiency test was administered, and others had several years’ experience living in English speaking countries.

Furthermore, the data finally analyzed was that of only 64 students of the initial 101 participants. This was because the present study was a comparative study of the students who were in the top one third and the students who were in the bottom one third of the group according to L2 proficiency scores. Thus, the remainder of the data was not entered for analysis (for more information, see section 3.2.3). The 64 students included 31 biology majors, and 33 engineering majors.

3.2.2 Materials

Two types of materials were prepared for this study; one was to assess the participants’ L2 (English in the present case) proficiency, and the other was for written recall protocols in which participants’ inference generation was studied. Let us explain the L2 proficiency test first, and then the reading passages for the recall task.

With regard to the L2 proficiency test, all of the questions were adopted from the previously administered Society for Testing English Proficiency (STEP) 2nd grade tests. The number of items and test methods were almost the same as those of the original STEP 2nd grade test. A total of 75 items were included. The L2 proficiency test had four sections, vocabulary, grammar, composition, and reading comprehension, but it did not contain a
listening comprehension section. All of the items were multiple-choice questions. The items in the vocabulary section, the grammar section, and the reading comprehension section had four options, whereas the items on the composition section had seven options. An example of each section is shown in Table 3.1.

Table 3.1

*Examples of L2 Proficiency Test Used in Study 1*

<table>
<thead>
<tr>
<th>Section</th>
<th>Number of Items</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Vocabulary</td>
<td>20</td>
<td>The light was too (   ) for me to read the magazine.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(l) vague (2) steady (3) thin (4) dim*</td>
</tr>
<tr>
<td>2. Grammar</td>
<td>20</td>
<td>May I (   ) this jacket to see if it fits?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(l) try on* (2) take on (3) pull on (4) wear on</td>
</tr>
<tr>
<td>3. Composition</td>
<td>20</td>
<td>Imoto no seken shirazu niwa yoku odorokasaremashu.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I am often (1) of (2) surprised (3) my sister (4) how (5) the (6) knows*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>how (5) the (6) knows (7) little* world.</td>
</tr>
<tr>
<td>4. Reading Comprehension</td>
<td>15</td>
<td>(see Appendix 3-A)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>75</strong></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Choices with asterisks in a row of “Example” indicate answer keys. As to the composition section, the examinees were expected to arrange seven words or phrases to make an appropriate sentence which had the same meaning as a Japanese sentence, and to answer the 3rd and the 5th words or phrases from the top.

*3 = answer key for the 3rd word or phrase. *5 = answer key for the 5th word or phrase.

The expository prose passages were selected from textbooks for EFL and ESL learners for a recall task (see Appendix 3-B). The following points were considered when choosing these materials: (a) whether the topics were attractive to students with math and science majors, (b) whether the texts contained a number of concrete events and facts appropriate for the recall analysis, (c) whether the writing styles would be familiar to students who were
learning English as a foreign language, and (d) whether almost all of the students would probably be unfamiliar with the content of the stories. Because the aim of the present study was to investigate what types of inferences and how many inferences Japanese EFL learners would make during or after reading a general passage, the materials which met these conditions were used for data collection.

The participants read three passages with two different readability levels, that is, around 4.0 and around 9.0 on the Flesch-Kincaid readability scale. Thus, Passage 1 was predicted to be much easier than Passage 2 and Passage 3. Passage 1 and Passage 2 were approximately equal in length, and Passage 3 was longer than the other two passages (see Table 3.2).

Table 3.2

*Passages Used in Study 1*

<table>
<thead>
<tr>
<th></th>
<th>Passage 1</th>
<th>Passage 2</th>
<th>Passage 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
<td>Do Pigeons Take the Train?</td>
<td>Research on the Effects of Color and Appearance</td>
<td>Do Animals Sleep?</td>
</tr>
<tr>
<td><strong>Word</strong></td>
<td>195</td>
<td>223</td>
<td>389</td>
</tr>
<tr>
<td><strong>Flesch Reading Ease</strong></td>
<td>79.8</td>
<td>59.7</td>
<td>59.1</td>
</tr>
<tr>
<td><strong>Flesch-Kincaid Grade</strong></td>
<td>4.3</td>
<td>8.8</td>
<td>9.2</td>
</tr>
</tbody>
</table>

*Note.* *These indices were provided by Microsoft Word 2000’s readability measurement tools. When readability was measured, titles were tentatively excluded.*

Titles were originally attached to Passage 1 and Passage 3, but Passage 2 had no title; therefore, to equalize the conditions of these three materials, a new title was created by the
author following discussion with a native speaker of English who had been teaching English as a foreign language at a Japanese university.

### 3.2.3 Data Collection Procedures

Prior to the main study, a pilot study was conducted in order to determine the appropriate timeframe for reading and recalling. Five second-year students who were non-English majors participated in the pilot study. All were volunteers. They were enrolled at the same university as the participants in the main study, but they were not the participants of the main study. For the purpose of this pilot study, the participants were instructed to stop reading whenever they were ready to move on to the recall task, and to stop writing whenever they completed recalling. Time was measured by the researcher.

#### Table 3.3

<table>
<thead>
<tr>
<th>Participant</th>
<th>Passage 1</th>
<th>Passage 2</th>
<th>Passage 3</th>
<th>Qualification of Language Ability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reading</td>
<td>Recall</td>
<td>Reading</td>
<td>Recall</td>
</tr>
<tr>
<td>Student 1</td>
<td>2.0</td>
<td>6.0</td>
<td>2.0</td>
<td>8.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>STEP 2nd grade holder</td>
</tr>
<tr>
<td>Student 2</td>
<td>2.0</td>
<td>5.0</td>
<td>4.0</td>
<td>9.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Not specified)</td>
</tr>
<tr>
<td>Student 3</td>
<td>8.0</td>
<td>6.0</td>
<td>9.0</td>
<td>6.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>STEP 3rd grade holder</td>
</tr>
<tr>
<td>Student 4</td>
<td>6.0</td>
<td>6.0</td>
<td>8.0</td>
<td>5.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>14.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>STEP pre-2nd grade holder</td>
</tr>
<tr>
<td>Student 5</td>
<td>8.0</td>
<td>10.0</td>
<td>9.0</td>
<td>12.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>12.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Not specified)</td>
</tr>
<tr>
<td>Average</td>
<td>5.2</td>
<td>6.6</td>
<td>6.4</td>
<td>8.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9.6</td>
</tr>
</tbody>
</table>

Results of this pilot study are shown in Table 3.3. We can see large individual differences. For example, Student 1 read Passage 3 for 3.0 minutes, whereas Student 3 read it for 15.0 minutes, viz., 5 times as long as that of Student 1; and Student 4 recalled Passage 2 for 5.0 minutes, whereas Student 5 recalled it for 12.0 minutes.
Firstly, times for recall were decided so that almost all of the students would be able to finish writing whatever they remembered about a text. Therefore, 10.0 minutes were given for recalling a passage. Secondly, times for reading were determined more carefully than by simply taking the average of the pilot study or by taking the maximum time. This was because giving too much time may let students use the strategy of rehearsal for rote learning when the present focus was their reading comprehension. Taking all of these factors into consideration, it was determined that 3.0 minutes would each be given to read Passage 1 and Passage 2, and 6.0 minutes would be given to read Passage 3 for the main study (see Table 3.4). As a result, the reading speed of 64-74 words per minute was set. Although these reading rates were slower than an ideal speed for Japanese university learners of English (e.g., Ushiro, 2000), the present study intended to measure learners’ inference generation during and after reading, and let them take more time than in speed reading; therefore, they were considered to be appropriate times.

<table>
<thead>
<tr>
<th>Allotted Times for a Recall Task in Study 1 (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time to Read</td>
</tr>
<tr>
<td>Time to Recall</td>
</tr>
</tbody>
</table>

Moreover, after a series of recall tasks, the students who took part in the preliminary study had a short interview with the researcher, and were asked whether they had known the plots to the stories; and it was shown that no student had known them. This preliminary investigation was conducted on September 18th and 20th, 2001.

The procedures of the main study will be explained hereafter. The L2 proficiency test
and a recall test were administered in regular English classes on different days; the former
was conducted on April 10th and 12th, whereas the latter was conducted on October 2nd
and 4th, 2001. Considering that some students may have successfully developed their
English language ability during that term, but others not, the present study did not employ
the research design of the correlational analysis. Instead, a comparison of reading
comprehension between the top third students and the bottom third students on L2
proficiency test was made. The data of the students with the middle level of L2 proficiency
was not entered for analysis.

With regard to the L2 proficiency test, the students were given 75 multiple-choice
items and told to write the answers on answer sheets. To answer these questions, 60 minutes
were allotted.

Participants also worked on a recall task by simply following the researcher’s
instructions. The participants received a booklet in which there were three English passages
and three answer sheets with lines for a recall task; an answer sheet was followed
immediately by a passage. Firstly, the researcher read a rubric printed on the first page³
aloud to them. As is shown in Appendix 3-C, the participants were told in advance that they
were supposed to read three English passages in total; and after they finished reading each
passage, they were going to start writing in their L1 (i.e., Japanese) whatever they could
remember as to the passage without looking at the passage. They were also instructed to
follow the researcher’s instructions for the start and finish times.

3.2.4 Scoring

The L2 proficiency test was an objective measure, and only the researcher worked on
the scoring of this. However, the recall protocols were scored by the researcher, and a
Japanese graduate student, who was studying English language education in a university doctoral course. Any information written in participants’ protocols but not seen in the original passage was basically considered to have been produced through inference generation. The raters counted how many instances of inferences were found in each protocol.

Inferences were then classified into three categories, “local connecting inferences,” “global connecting inferences” or “elaborative inferences,” as discussed in section 3.1. They were judged by the main sources of information from which inferences were drawn, and by their function. If two separate pieces of information in a reading passage were recalled together with some coordinate or subordinate conjunctions in a protocol, this counted as a connecting inference. In addition, if the connected pieces of information in a protocol originally appeared within two adjacent sentences in a reading passage, it was “a local connecting inference.” If the pieces of information connected by inferences did not appear within two adjacent sentences in a reading passage, the inference was labelled “a global connecting inference.” In short, a connecting inference was categorized as either “a local connecting inference” or “a global connecting inference” based on the distance between the two pieces of information in the original text. “Elaborative inferences,” which are different from both kinds of “connecting inferences,” make part of a passage more informative by adding such information as “why,” “when,” “who,” “where” and “how” through the activation of readers’ prior knowledge.

It should also be noted that the main purpose of this study was to find whether mental representations of higher L2 proficiency learners were different from those of lower L2 proficiency learners, and if so, how different they were; therefore, only appropriate inferences were targeted. Inferences incompatible with the content of a text were excluded.
from this investigation.

Moreover, although it was not the main focus of the present study, the numbers of idea units (IUs) correctly produced in the learners’ protocols were counted. Judgment on IUs proceeded as follows: (a) Reading passages were divided into IUs basically following the criteria of Muramoto (1998, pp. 104-105), (b) numbers were attached to each IU in passages, (c) two raters determined independently whether each IU was shown without errors by looking at each participant’s recall protocol and reading passages divided into IUs, and finally, (d) the numbers of IUs recalled correctly were counted. One point was given for an each correct recall of an IU.

3.2.5 Data Analyses

The real numbers of inferences were entered by type for analysis; in other words, the proportion of inferences by type to the total number of inferences was not computed at the present time. This was because the main purpose of the present study was to investigate how rich mental representations were constructed by L2 proficiency, with a focus on their inference generation processes; thus, the real number of inferences were judged to reflect their mental representations themselves rather than the proportion.  

For hypothesis testing, a three-way analysis of variance (ANOVA; by L2 proficiency, type of inference, and passage) was performed in order to examine interactions between each pair of variables (e.g., the L2 proficiency \( \times \) type of inference interaction), and also between all of the variables (i.e., the L2 proficiency \( \times \) type of inference \( \times \) passage interaction). The type of inference and passage were within-subject factors, while L2 proficiency was a between-subject factor. As the in-depth analysis, a \( t \)-test and a one-way ANOVA were conducted. At the same time, the effect size was computed for hypothesis testing.
A *t*-test was also conducted to investigate the difference in participants’ scores on the sum of inferences due to their L2 proficiency levels. In addition, the Pearson’s product-moment correlation was employed to examine inter-rater reliability, whereas Cronbach’s alpha correlation and a *t*-test were employed to examine the validity of the test material.

The statistical software package SPSS for Windows (Version 10.0E) was used for all of the data entry and analysis.

### 3.3 Results

#### 3.3.1 Validity and Descriptive Statistics of L2 Proficiency Test

I will start by discussing the validity of the L2 proficiency test used in the present study. According to Messick (1989), validity is defined as “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment” (pp. 116-117). In the present study, the validity of interpretations and uses based on the L2 proficiency test was examined from the structural aspect and the external aspect. There were two reasons why these two aspects were targeted among the six aspects (the content aspect, the substantive aspect, the structural aspect, the generalizability aspect, the external aspect, and the consequential aspect): (a) I followed a number of previous reading studies reporting internal consistency of the test such as Cronbach’s alpha and the strength of relationship with some external criterion measure, and (b) I needed to limit some of the aspects because test validation was not the main focus of the present study. To elaborate more about the structural aspect and the external aspect, the former concerns “the internal structure of the test by assessing the extent to which observed dimensionality of
response data is consistent with the hypothesized dimensionality of a construct” (Chapelle, 1999, p. 261), and the latter concerns “relationships of test scores with other tests and behaviors” (Chapelle, p. 262). The results of the structural aspect (i.e., Cronbach’s alpha) are shown in the next paragraph, and those of the external aspect (i.e., a t-test) are shown at the beginning of the next section (see section 3.3.2).

Regarding validation from the structural aspect, Cronbach’s alpha was reported to be $\alpha = .833$ as to the L2 reading proficiency test, which had 75 items. It was a high coefficient. Thus, we can say that positive evidence was found for the validity of the L2 proficiency test from the structural aspect. Furthermore, point biserial correlation coefficients were computed by item (see Appendix 3-D). Of 75 items, 38 items met the Henning’s (1987) criteria (i.e., .250 and above), and the average for a total of 75 items was .234, which were slightly below .250. The top third participants were labelled “Upper,” the bottom third were labelled “Lower,” and the remainder were labelled “Middle.” Due to there being more than one student at the cut-off scores, the numbers of participants in the three groups were not exactly the same.

Table 3.5

<table>
<thead>
<tr>
<th>L2 Proficiency Group</th>
<th>n</th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper</td>
<td>32</td>
<td>55.219</td>
<td>3.875</td>
<td>50.000</td>
<td>64.000</td>
</tr>
<tr>
<td>Middle</td>
<td>29</td>
<td>44.138</td>
<td>2.863</td>
<td>40.000</td>
<td>49.000</td>
</tr>
<tr>
<td>Lower</td>
<td>32</td>
<td>34.438</td>
<td>4.024</td>
<td>23.000</td>
<td>39.000</td>
</tr>
<tr>
<td>Total</td>
<td>93</td>
<td>44.613</td>
<td>9.392</td>
<td>23.000</td>
<td>64.000</td>
</tr>
</tbody>
</table>

Note. Full score = 75.000.

The means and standard deviations (SDs) of the test are shown in Table 3.5. The mean
scores of “Upper,” “Middle,” and “Lower” were 55.219 ($SD = 3.875$), 44.138 ($SD = 2.863$), 34.438 ($SD = 4.024$), respectively. Note that the mean score of “Upper” accounted for 73.6% of a full score; and that even the student with lowest score in “Upper” obtained 66.7% of a full score. Considering that around 60% was a pass/fail line (Society for Testing English Proficiency, n.d.), all of the students grouped into “Upper” in this study had almost the same English language ability as STEP 2nd grade holders.

A $t$-test established the difference in English proficiency between “Upper” and “Lower,” $t (62) = -21.204$, $p < .001$. In addition, the effect size indicated by $g^6$ was extremely large, $g = 5.261$, according to Cohen’s (1988) criteria. In the following sections, a comparison of inference generation was made between these groups.

### 3.3.2 Results of a Recall Test, With a Focus on Inferences

Due to the fact that two raters were involved in the scoring procedures of a recall test, the Pearson’s product-moment correlation coefficients between the scores were computed for inter-rater reliabilities (see Table 3.6).

<table>
<thead>
<tr>
<th></th>
<th>Passage 1</th>
<th>Passage 2</th>
<th>Passage 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r$</td>
<td>.939**</td>
<td>.911**</td>
<td>.913**</td>
</tr>
</tbody>
</table>

*Note. $r$ = Pearson’s product-moment correlation coefficients.

**$p < .01$.**

In terms of scoring IUs, very high coefficients were obtained, ranging from .911 ($p < .01$; Passage 2) to .939 ($p < .01$; Passage 1). Following this, the average of the two scores was
computed one by one for the following analysis.

The mean scores are shown in Table 3.7. A full score was the same as the total number of IUs in each passage. Out of 77 points, “Upper” obtained 30.453 ($SD = 8.339$) on average, whereas “Lower” obtained 22.078 ($SD = 7.459$) on average. The means of the two different and independent groups were determined by the $t$-test ($t (62) = 4.235$, $p < .001$) to be significantly different. That is, “Upper” group scored significantly higher on the recall test which focused on their comprehension of propositional meaning than “Lower” did. This was supported by the measure of magnitude, which reported the large effect, $g = 1.059$.

Table 3.7

**Descriptive Statistics of Idea Units Produced by Written Recall Protocols**

<table>
<thead>
<tr>
<th>L2 Proficiency Group</th>
<th>Passage 1 (Full score = 28.000)</th>
<th>Passage 2 (Full score = 19.000)</th>
<th>Passage 3 (Full score = 30.000)</th>
<th>Total (Full score = 77.000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper ($n = 32$)</td>
<td>Mean (Mean (%)) 11.406 (40.7)</td>
<td>8.422 (44.3)</td>
<td>10.625 (35.4)</td>
<td>30.453 (39.5)</td>
</tr>
<tr>
<td></td>
<td>SD 4.049</td>
<td>2.643</td>
<td>4.423</td>
<td>8.339</td>
</tr>
<tr>
<td></td>
<td>Minimum 3.000</td>
<td>4.500</td>
<td>1.000</td>
<td>17.500</td>
</tr>
<tr>
<td></td>
<td>Maximum 19.500</td>
<td>14.000</td>
<td>18.500</td>
<td>49.000</td>
</tr>
<tr>
<td>Lower ($n = 32$)</td>
<td>Mean 8.359 (29.9)</td>
<td>6.156 (32.4)</td>
<td>7.563 (25.2)</td>
<td>22.078 (28.7)</td>
</tr>
<tr>
<td></td>
<td>SD 3.984</td>
<td>2.179</td>
<td>3.240</td>
<td>7.459</td>
</tr>
<tr>
<td></td>
<td>Minimum 2.000</td>
<td>2.000</td>
<td>2.000</td>
<td>6.500</td>
</tr>
<tr>
<td></td>
<td>Maximum 16.500</td>
<td>12.500</td>
<td>15.000</td>
<td>41.500</td>
</tr>
<tr>
<td>Total ($N = 64$)</td>
<td>Mean 9.883 (35.3)</td>
<td>7.289 (38.4)</td>
<td>9.094 (30.3)</td>
<td>26.266 (34.1)</td>
</tr>
<tr>
<td></td>
<td>SD 4.270</td>
<td>2.660</td>
<td>4.144</td>
<td>8.911</td>
</tr>
<tr>
<td></td>
<td>Minimum 2.000</td>
<td>2.000</td>
<td>1.000</td>
<td>6.500</td>
</tr>
<tr>
<td></td>
<td>Maximum 19.500</td>
<td>14.000</td>
<td>18.500</td>
<td>49.000</td>
</tr>
</tbody>
</table>

*Note.* Values in parentheses next to means were computed through dividing a mean by a full score.

A two-way ANOVA (by L2 proficiency and passage) was run to investigate whether the difference in means of IUs between “Upper” and “Lower” was not affected by the choice
of passage. There was no interaction between L2 proficiency and the passage \((F(2, 124) = .411, n.s.)\); and furthermore, there was the main effect for L2 proficiency \((F(1, 62) = 17.932, p < .001)\). Overall, there was no significant effect of text factors such as text length and readability on the difference in the amount of recalled IUs between “Upper” and “Lower.” Moreover, standardized mean differences offered almost the same results as follows: the effect size for the interaction was small, \(\eta^2 = .011\), whereas the effect size for the main effect for L2 proficiency was large, \(\eta^2 = .616\) (Cohen, 1988, pp. 284-287). Thus, another positive piece of evidence was provided for the validity of the L2 proficiency test.

In the remainder of this section, inter-rater reliabilities as to three types of inferences, and the results of statistical analyses are provided.

Table 3.8

Inter-Rater Reliability in the Scoring of Inferences by Passage

<table>
<thead>
<tr>
<th>Passage 1</th>
<th>Passage 2</th>
<th>Passage 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCIs</td>
<td>GCI s</td>
<td>ELs</td>
</tr>
<tr>
<td>(r)</td>
<td>(.806^{**})</td>
<td>(.796^{**})</td>
</tr>
</tbody>
</table>

Note. \(r\) = Pearson’s product-moment correlation coefficients.

**\(p < .01\).

LCIs = local connecting inferences; GCI s = global connecting inferences; ELs = elaborative inferences.

As Table 3.8 indicates, high inter-rater reliabilities were found in eight of the nine ratings, which ranged between \(r = .771\) \((p < .01; \text{“elaborative inferences” in Passage 3})\) and \(r = .906\) \((p < .01; \text{“elaborative inferences” in Passage 2})\); in turn, the rating of “elaborative inferences” in protocols of Passage 1 produced a moderate reliability, \(r = .610\) \((p < .01)\). These results may be because (a) rating “elaborative inferences” was more difficult than rating the other two inferences because raters should have judged whether learners used their background
knowledge by looking at their written protocols, and (b) in addition to the difficulty of the task, because rating was started from Passage 1 through Passages 2 and 3, both raters were not yet familiar with rating during counting “elaborative inferences” as to Passage 1. Therefore, basically, the average of the counts by the two raters was calculated for final scores, whereas for “elaborative inferences” for Passage 1, all disagreements were resolved through discussion by the raters to produce final scores.

The mean number of inferences and standard deviations are shown in Table 3.11 by inference type and passage. In terms of the average number of inferences, “Lower” generated 4.453 (SD = 3.236) for Passage 1, 4.781 (SD = 2.703) for Passage 2, and 3.203 (SD = 2.856) for Passage 3, while “Upper” generated 7.156 (SD = 3.796) for Passage 1, 7.234 (SD = 2.954) for Passage 2, and 5.922 (SD = 2.913) for Passage 3.

A three-way repeated-measures ANOVA was performed to assess the significance of the observed difference in means between L2 proficiency, passage, inference type, and the interactions between them. As is shown in Table 3.9 and Table 3.10, main effects for passage ($F(2, 124) = 5.583, p < .01$), for inference type ($F(2, 124) = 17.058, p < .001$), and for L2 proficiency ($F(1, 62) = 22.434, p < .001$) were significant. In addition, two interaction effects were found between inference type and L2 proficiency ($F(2, 186) = 6.355, p < .01$; see Figure 3.1), and between passage and inference type ($F(4, 248) = 12.489, p < .001$; see Figure 3.2). The former interaction indicated that L2 proficiency levels did differently affect the average numbers of three inference types; and the latter interaction indicated that the passages did differently affect the average numbers of three inference types. No other interaction was statistically justified.

Furthermore, the effect size index of $\eta^2$ was computed, and the following results were obtained: (a) The large effect sizes were found in L2 proficiency ($\eta^2 = 0.365$), in the
interaction between passage and inference type \( (\eta^2 = 0.232) \), and in the main effect for inference type \( (\eta^2 = 0.218) \), (b) the medium effect sizes were found in the main effect for passage \( (\eta^2 = 0.086) \), and in the interaction between inference type and L2 proficiency \( (\eta^2 = 0.081) \), (c) the small effect size was found in the interaction among the three variables, passage, L2 proficiency, and inference type \( (\eta^2 = 0.018) \), and (d) there was no effect in the interaction between passage and L2 proficiency \( (\eta^2 = 0.001) \). The results of significance testing did not contradict the results of effect size estimation.

Table 3.9

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>( \eta^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passage</td>
<td>2</td>
<td>26.121</td>
<td>13.060</td>
<td>5.583**</td>
<td>0.086</td>
</tr>
<tr>
<td>Passage × L2 Proficiency</td>
<td>2</td>
<td>0.237</td>
<td>0.118</td>
<td>0.051</td>
<td>0.001</td>
</tr>
<tr>
<td>Error (Passage)</td>
<td>124</td>
<td>290.087</td>
<td>2.339</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inference Type</td>
<td>2</td>
<td>65.795</td>
<td>32.898</td>
<td>17.058***</td>
<td>0.218</td>
</tr>
<tr>
<td>Inference Type × L2 Proficiency</td>
<td>2</td>
<td>24.510</td>
<td>12.255</td>
<td>6.355**</td>
<td>0.081</td>
</tr>
<tr>
<td>Error (Inference Type)</td>
<td>124</td>
<td>239.139</td>
<td>1.929</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Passage × Inference Type</td>
<td>4</td>
<td>70.095</td>
<td>17.524</td>
<td>12.489***</td>
<td>0.232</td>
</tr>
<tr>
<td>Passage × Inference Type × L2 Proficiency</td>
<td>4</td>
<td>5.307</td>
<td>1.327</td>
<td>0.946</td>
<td>0.018</td>
</tr>
<tr>
<td>Error (Passage × Inference Type)</td>
<td>248</td>
<td>347.986</td>
<td>1.403</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note. **p < .01, ***p < .001.*

\( \eta^2 \) = eta squared (effect size).

0.010 \leq \eta^2 (small) < 0.059; 0.059 \leq \eta^2 (medium) < 0.138; 0.138 \leq \eta^2 (large; Cohen, 1988, pp. 284-287). Effect size of 0.009 and below is regarded as having no effect.
Figure 3.1. A line graph describing an interaction between inference type and L2 proficiency.

Figure 3.2. A line graph describing an interaction between passage and inference type.

To summarize these results, firstly, it was indicated that for all of the passages “Upper”
generated significantly more inferences than “Lower.” That is, the priority of “Upper” as to inference generation to “Lower” was not interfered with the passage factor.

Table 3.10

Results of a Repeated-Measures ANOVA: Tests of Between-Subjects Effects

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>$\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2 Proficiency</td>
<td>1</td>
<td>110.250</td>
<td>110.250</td>
<td>22.434***</td>
<td>0.365</td>
</tr>
<tr>
<td>Error</td>
<td>62</td>
<td>304.694</td>
<td>4.914</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.*** $p < .001$.

$\eta^2 = \text{eta squared (effect size)}$.

0.010 $\leq \eta^2$ (small) $< 0.059$; 0.059 $\leq \eta^2$ (medium) $< 0.138$; 0.138 $\leq \eta^2$ (large; Cohen, 1988, pp. 284-287). Effect size of 0.009 and below is regarded as having no effect.

Secondly, in order to conduct an in-depth analysis of the interaction between inference type and L2 proficiency, *t*-tests were performed to test differences in the mean number of inferences between groups by passage and inference type. The present study did not employ the one-way ANOVA through computing the sum of inferences by type. Instead, results which appeared consistently across three passages were interpreted.

The results are presented below in the order of Passage 1, Passage 2, and Passage 3. Regarding “local connecting inferences” and “global connecting inferences” generated for Passage 1, the *t*-tests indicated that the number of inferences generated by “Upper” was significantly greater than that of “Lower” at the .05 level ($t (62) = 3.907, p < .001$; $t (62) = 2.026, p < .05$, respectively). On the other hand, there was no statistical difference between the groups in terms of “elaborative inferences” for Passage 1 ($t (62) = 1.029, \text{n.s.}$).

The results for Passage 2 showed similar results to those for Passage 1: As for “local connecting inferences” and “global connecting inferences,” significant differences existed between the groups ($t (56.273) = 2.855, p < .01$; $t (62) = 2.218, p < .05$), but with
“elaborative inferences” no significant difference was found ($t (62) = 1.624, \text{n.s.}$). For Passage 3, the results indicated that the difference between the two proficiency groups was statistically significant in “local connecting inferences” ($t (62) = 4.570, p < .001$) and “elaborative inferences” ($t (62) = 2.378, p < .05$), but not in “global connecting inferences” for the same passage ($t (62) = 1.008, \text{n.s.}$). Thus, the results of Passage 3 were different from those of Passages 1 and 2 in part (see Table 3.11).

Table 3.11

| T-Tests for Average Number of Inferences of Two Groups |
|-------------------------------------|-----|-----|-----|-----|-----|-----|-----|
| Passage | Inference Type | Lower ($n = 32$) | | | | | |
| | | Mean | $SD$ | Mean | $SD$ | $df$ | $t$ | $p$ | $g$ |
| 1 | LCI | 1.781 | 1.367 | 3.313 | 1.745 | 62 | 3.907 | .000 | 0.977 |
| | GCI | 1.172 | 1.268 | 1.938 | 1.722 | 62 | 2.026 | .047 | 0.507 |
| | EI | 1.500 | 1.626 | 1.906 | 1.532 | 62 | 1.029 | .308 | 0.257 |
| | Total | 4.453 | 3.236 | 7.156 | 3.796 | 62 | 3.065 | .003 | 0.766 |
| 2 | LCI | 1.297 | 1.3 | 2.422 | 1.81 | 56.273 | 2.855 | .006 | 0.714 |
| | GCI | 1.375 | 1.218 | 2.016 | 1.089 | 62 | 2.218 | .030 | 0.555 |
| | EI | 2.109 | 1.523 | 2.797 | 1.849 | 62 | 1.624 | .109 | 0.406 |
| | Total | 4.781 | 2.703 | 7.234 | 2.954 | 62 | 3.47 | .001 | 0.866 |
| 3 | LCI | 1.563 | 1.343 | 3.281 | 1.651 | 62 | 4.570 | .000 | 1.142 |
| | GCI | 1.000 | 1.320 | 1.313 | 1.155 | 62 | 1.008 | .317 | 0.252 |
| | EI | 0.641 | 1.064 | 1.328 | 1.242 | 62 | 2.378 | .021 | 0.594 |
| | Total | 3.203 | 2.856 | 5.922 | 2.913 | 62 | 3.770 | .000 | 0.943 |

Note. LCI = local connecting inference; GCI = global connecting inference; EI = elaborative inference.

$g =$ effect size. $0.200 \leq g \text{ (small)} < 0.500;$ $0.500 \leq g \text{ (medium)} < 0.800;$ $0.800 \leq g \text{ (large; Cohen, 1988). Effect size of 0.199 and below is regarded as having no effect.}$
Figure 3.3. Means of inferences by type and group in Passage 1. LCI = local connecting inference; GCI = global connecting inference; EI = elaborative inference.

Figure 3.4. Means of inferences by type and group in Passage 2. LCI = local connecting inference; GCI = global connecting inference; EI = elaborative inference.
A summary of the results of $t$-tests is as follows: (a) The difference in the number of “local connecting inferences” between “Upper” and “Lower” was revealed in all three passages, (b) the difference in the number of “global connecting inferences” between “Upper” and “Lower” was revealed in two passages, and (c) the difference in the number of “elaborative inferences” between “Upper” and “Lower” was revealed in one passage.

In addition to $t$ values as mentioned above, effect sizes (Cohen, 1988) were computed (see Table 3.11). The results showed the same tendency as the results of the $t$-tests. To sum up, with regard to “local connecting inferences,” large effects were found on two (i.e., Passage 1 and Passage 3) of the three passages, and had a medium effect on the other one (i.e., Passage 2). With regard to “global connecting inferences,” medium effects were found on two passages (i.e., Passage 1 and Passage 2), and had a small effect on the other one (i.e., Passage 2). As to “elaborative inferences,” a medium effect was found on one passage (i.e.,...
Passage 3), and small effects were found on the other two passages (i.e., Passage 1 and Passage 2). These results are on the same line as the results of the significance testing.

Table 3.12

Repeated-Measures One-Way ANOVAs for Effects of L2 Proficiency on Inference Types

<table>
<thead>
<tr>
<th>Variable and Source</th>
<th>(df)</th>
<th>(SS)</th>
<th>(MS)</th>
<th>(F)</th>
<th>(\eta^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Passage 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within groups</td>
<td>2</td>
<td>5.953</td>
<td>2.977</td>
<td>2.263</td>
<td>0.068</td>
</tr>
<tr>
<td>Between groups</td>
<td>1</td>
<td>211.523</td>
<td>211.523</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Passage 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within groups</td>
<td>2</td>
<td>12.859</td>
<td>6.430</td>
<td>4.205*</td>
<td>0.119</td>
</tr>
<tr>
<td>Between groups</td>
<td>1</td>
<td>243.844</td>
<td>243.844</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Passage 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within groups</td>
<td>2</td>
<td>13.818</td>
<td>6.909</td>
<td>7.059**</td>
<td>0.185</td>
</tr>
<tr>
<td>Between groups</td>
<td>1</td>
<td>109.440</td>
<td>109.440</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Upper</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Passage 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within groups</td>
<td>2</td>
<td>41.271</td>
<td>20.635</td>
<td>11.624***</td>
<td>0.273</td>
</tr>
<tr>
<td>Between groups</td>
<td>1</td>
<td>546.260</td>
<td>546.260</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Passage 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within groups</td>
<td>2</td>
<td>9.771</td>
<td>4.885</td>
<td>1.966</td>
<td>0.060</td>
</tr>
<tr>
<td>Between groups</td>
<td>1</td>
<td>2.250</td>
<td>2.250</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Passage 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within groups</td>
<td>2</td>
<td>82.036</td>
<td>41.018</td>
<td>29.584***</td>
<td>0.488</td>
</tr>
<tr>
<td>Between groups</td>
<td>1</td>
<td>61.035</td>
<td>61.035</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. \(\eta^2 = \) eta squared (effect size).

0.010 \(\leq \eta^2\) (small) < 0.059; 0.059 \(\leq \eta^2\) (medium) < 0.138; 0.138 \(\leq \eta^2\) (large; Cohen, 1988, pp. 284-287).

Effect size of 0.009 and below is regarded as having no effect.

* \(p < .05\). ** \(p < .01\). *** \(p < .001\).

Furthermore, as to interaction between inference type and L2 proficiency, one-way, repeated measures ANOVAs were calculated by group and passage (see Table 3.12). The results of four ANOVAs (i.e., Passage 2 of “Lower,” Passage 3 of “Lower,” Passage 1 of
“Upper,” and Passage 3 of “Upper”) out of six revealed a statistically significant main effect for inference type. We can interpret these results as follows: To cite an example of Passage 2 of “Lower,” a difference was shown in the amount among three types of inferences generated by “Lower” as to Passage 2. Effect sizes were medium or large. Further analyses to find the pairs of inference types that have a difference in the amount are reported in the following paragraph.

With regard to the four cases, whose $p$ values were less than .05, comparisons between all of the possible pairwise combinations of the three means of inferences were made by the Tukey’s Honestly Significant Difference (HSD) procedure. Table 3.13 indicates the results of the multiple comparisons. As a result, “local connecting inferences” tended to be more than the other two inference types; in addition, this tendency was more powerful in “Upper.” Only “Passage 2” of “Lower” offered a different result, that is, the amount of “elaborative inferences” was meaningfully larger than that of “local connecting inferences.”

Table 3.13

Results of Tukey’s HSD Test (1)

<table>
<thead>
<tr>
<th>Group and Passage</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Passage 1</td>
<td>Passage 2</td>
</tr>
<tr>
<td>Lower</td>
<td>--</td>
<td>EI &gt; LCI*</td>
</tr>
<tr>
<td>Upper</td>
<td>LCI &gt; GCI*</td>
<td>LCI &gt; EI*</td>
</tr>
</tbody>
</table>

Note. *$p < .05$.

LCI = local connecting inference; GCI = global connecting inference; EI = elaborative inference.
The third point which should be noted about the results of the three-way repeated-measures ANOVA is the interaction between passage and inference type. The post-hoc analyses were conducted by using ANOVAs. The results of the effects of inference types on their amounts by passage, and then, the results of the effects of passages on the amounts of inferences by inference type are shown in the following paragraphs. The participants’ L2 proficiency was not considered in this case.

Table 3.14

Descriptive Statistics of Three Types of Inferences by Passage

<table>
<thead>
<tr>
<th>Inference Type</th>
<th>Passage 1</th>
<th>Passage 2</th>
<th>Passage 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local Connecting</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>2.547</td>
<td>1.859</td>
<td>2.422</td>
</tr>
<tr>
<td>SD</td>
<td>1.736</td>
<td>1.663</td>
<td>1.726</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Maximum</td>
<td>7.000</td>
<td>6.500</td>
<td>6.000</td>
</tr>
<tr>
<td>Global Connecting</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>1.555</td>
<td>1.695</td>
<td>1.156</td>
</tr>
<tr>
<td>SD</td>
<td>1.548</td>
<td>1.191</td>
<td>1.240</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Maximum</td>
<td>6.500</td>
<td>4.000</td>
<td>6.000</td>
</tr>
<tr>
<td>Elaborative</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>1.703</td>
<td>2.453</td>
<td>0.984</td>
</tr>
<tr>
<td>SD</td>
<td>1.580</td>
<td>1.715</td>
<td>1.198</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Maximum</td>
<td>7.000</td>
<td>9.500</td>
<td>4.500</td>
</tr>
</tbody>
</table>

Note. *N = 64.*

Means and standard deviations by inference type and passage are shown in the above table. As to “local connecting inference,” 2.547 (√D = 1.736) for Passage 1, 1.859 (√D = 1.663) for Passage 2, and 2.422 (√D = 1.726) for Passage 3 were generated on average; as to “global connecting inference,” 1.555 (√D = 1.548) for Passage 1, 1.695 (√D = 1.191) for
Passage 2, 1.156 ($SD = 1.240$) for Passage 3 on average; and as to “elaborative inference,”
1.703 ($SD = 1.580$) for Passage 1, 2.453 ($SD = 1.715$) for Passage 2, 0.984 ($SD = 1.198$) for Passage 3 on average. To sum up, we can reasonably state that these means were not so large because they were all within the range from 0.984 for “elaborative inferences” of Passage 3 to 2.547 for “local connecting inferences” of Passage 1.

As is indicated in Table 3.15, a significant main effect of inference type was found in all three passages. Effect sizes were medium or large.

Table 3.15

Repeated-Measures One-Way ANOVAs for Effects of Passage on Inference Types (1)

<table>
<thead>
<tr>
<th>Variable and Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>$\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passage 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within groups</td>
<td>2</td>
<td>36.659</td>
<td>18.329</td>
<td>11.423***</td>
<td>0.153</td>
</tr>
<tr>
<td>Between groups</td>
<td>1</td>
<td>22.781</td>
<td>22.781</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Passage 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within groups</td>
<td>2</td>
<td>20.346</td>
<td>10.173</td>
<td>5.104**</td>
<td>0.075</td>
</tr>
<tr>
<td>Between groups</td>
<td>1</td>
<td>11.281</td>
<td>11.281</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Passage 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within groups</td>
<td>2</td>
<td>78.885</td>
<td>39.443</td>
<td>30.375***</td>
<td>0.325</td>
</tr>
<tr>
<td>Between groups</td>
<td>1</td>
<td>66.125</td>
<td>66.125</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. **$p < .01$. ***$p < .001$.
0.010 $\leq \eta^2$ (small) $< 0.059$; 0.059 $\leq \eta^2$ (medium) $< 0.138$; 0.138 $\leq \eta^2$ (large; Cohen, 1988, pp. 284-287). Effect size of 0.009 and below is regarded as having no effect.

Post hoc comparisons, Tukey’s HSD tests revealed that “local connecting inferences” were larger than “global connecting inferences,” and “elaborative inferences” in Passage 1 and Passage 3, whereas “elaborative inferences” were larger than “global connecting inferences” in Passage 2 (see Table 3.16).
Table 3.16

Results of Tukey’s HSD Test (2)

<table>
<thead>
<tr>
<th>Passage</th>
<th>Pair Which Revealed Significant Differences in Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LCI &gt; GCI* LCI &gt; EI*</td>
</tr>
<tr>
<td>2</td>
<td>EI &gt; GCI*</td>
</tr>
<tr>
<td>3</td>
<td>LCI &gt; GCI* LCI &gt; EI*</td>
</tr>
</tbody>
</table>

Note. *p < .05.

LCI = local connecting inference; GCI = global connecting inference; EI = elaborative inference.

To investigate the effects of passages on the amounts of generated inferences, repeated-measures one-way ANOVAs were employed by inference type (i.e., “local connecting inferences,” “global connecting inferences,” and “elaborative inferences”). Significant effects were found in all inference types; $F = 4.596$ ($p < .05$) for “local connecting inferences,” $F = 3.111$ ($p < .05$) for “global connecting inferences,” and $F = 21.153$ ($p < .001$) for “elaborative inferences” (see Table 3.17).

Table 3.17

Repeated-Measures One-Way ANOVAs for Effects of Passage on Inference Types (2)

<table>
<thead>
<tr>
<th>Variable and Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>$\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local Connecting Inferences</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within groups</td>
<td>2</td>
<td>17.167</td>
<td>8.583</td>
<td>4.596*</td>
<td>0.017</td>
</tr>
<tr>
<td>Between groups</td>
<td>1</td>
<td>994.630</td>
<td>994.630</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global Connecting Inferences</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within groups</td>
<td>2</td>
<td>10.008</td>
<td>5.004</td>
<td>3.111*</td>
<td>0.024</td>
</tr>
<tr>
<td>Between groups</td>
<td>1</td>
<td>414.188</td>
<td>414.188</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elaborative Inferences</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within groups</td>
<td>2</td>
<td>69.042</td>
<td>34.521</td>
<td>21.153***</td>
<td>0.109</td>
</tr>
<tr>
<td>Between groups</td>
<td>1</td>
<td>563.755</td>
<td>563.755</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. *$p < .05$. ***$p < .001$.

$0.010 \leq \eta^2$ (small) $< 0.059$; $0.059 \leq \eta^2$ (medium) $< 0.138$; $0.138 \leq \eta^2$ (large; Cohen, 1988, pp. 284-287). Effect size of 0.009 and below is regarded as having no effect.
However, multiple comparisons (Tukey's HSD procedures) revealed that significant differences were found only between pairs of “elaborative inferences”; to be more specific, this type of inference was generated the most in Passage 2, the second most in Passage 1, and the least in Passage 3 (see Table 3.18). In terms of “local connecting inferences” and “global connecting inferences,” no significant difference was suggested. Effect sizes measured by \( \eta^2 \) were computed at the same time.

Small effect sizes were found for “local connecting inferences” and “global connecting inferences,” and a medium effect size for “elaborative inferences.” As you may have noticed, the results obtained from significance testing and the results obtained from strength of association did not match perfectly in the current case; however, since these are not the main focus of the present study, I will not discuss this point further.

Table 3.18

Results of Tukey’s HSD Test (3)

<table>
<thead>
<tr>
<th>Inference</th>
<th>Pair Which Revealed Significant Differences in Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local Connecting</td>
<td>--</td>
</tr>
<tr>
<td>Global Connecting</td>
<td>--</td>
</tr>
<tr>
<td>Elaborative</td>
<td>Passage 2 &gt; Passage 1*</td>
</tr>
<tr>
<td></td>
<td>Passage 2 &gt; Passage 3*</td>
</tr>
<tr>
<td></td>
<td>Passage 1 &gt; Passage 3*</td>
</tr>
</tbody>
</table>

*\( p < .05 \).

3.4 Discussion

The focus of this research was to compare the reading processes of lower L2 proficiency learners with the reading processes of higher L2 proficiency learners, focusing especially on their inference generation. It has been considered that making inferences pertains to how rich the mental representations are that readers are able to obtain after reading (van Dijk & Kintsch, 1983). “Connecting inferences” serve the function of making
a text coherent, whereas “elaborative inferences” serve the function of making a text more
detailed with something unwritten complemented by readers’ background knowledge. Thus,
Hypotheses 1A, 1B and 1C were proposed on the assumption that the learners’ reading
processes should be examined more deeply by classifying inferences into types. As a result,
the following pieces of evidence were obtained. Hypothesis 1A that the amount of “local
connecting inferences” generated by higher L2 proficiency learners is greater than that
generated by lower L2 proficiency learners, was supported. The results of \( t \)-tests comparing
the two means of different L2 proficiency groups revealed in all three passages that “Upper”
constructed more “local connecting inferences” than “Lower.” In contrast, the results on
“global connecting inferences” and “elaborative inferences” were mixed, with two of the
three \( t \)-tests which compared the means of “global connecting inferences” showing
significance, and one of the three \( t \)-tests which compared the means of “elaborative
inferences” showing significance. Therefore, it would be safe to say that Hypotheses 1B and
1C were not supported, or that only Hypothesis 1B manifested such a tendency.

Furthermore, the following discussion can be derived from the function of inference
types. The distinct difference in reading processes between these L2 proficiency learners
centered on whether they could build up coherent mental representations by finding certain
relationships between pieces of textual information which were within a few sentences’
distance of each other. This process included identifying referents of pronouns,
understanding substitutive verbs such as “do,” and implicit causal relations. Let us look at
some examples from Passage 1 (see Appendix 3-B). The fourth sentence was divided into
two IUs, “Scientists are not sure how pigeons know their way” and “but they do.” If a
participant’s protocol indicated that “they” in the second IU referred to “pigeons,” and “do”
in the same unit could be replaced by “know (their way),” they were thought to have
engaged in making “local connecting inferences.” Also, please refer to the example of IUs labelled (22) and (23) in the appendix. Although the relationships between these two units were not marked by any words, readers were expected to recognize the implicit causal relation of these units in order to understand these meanings; “because London is a big city, the pigeons get tired of flying,” as one example. Thus, the present study concluded that higher L2 proficiency learners were better at finding these connections than lower L2 proficiency learners.

The reasons why there were no noteworthy differences between the different L2 proficiency groups in the amount of “an elaborative inference” could be interpreted from its function as well. The purpose of the present study was to study Japanese EFL learners’ inference generation in general; therefore, types of “elaborative inferences” or concrete examples of background knowledge, which should be used, were not limited beforehand (for an example of readers’ background knowledge, see Ushiro, Koizumi et al., 2004; Ushiro, Shimizu et al., 2004). For example, an association such as *The Lion King* after reading the sentence of “the lion and leopard” (Zwaan & Brown, 1996) can be also considered as a member of “elaborative inferences,” and counted as one. In sum, how far apart the appeared background knowledge was from what a passage originally expected to be utilized was not a concern of the present study. It followed the previous studies that employed protocol analyses (e.g., Coté & Goldman, 1999; Horiba, 1993, 1996; Miyaura, 1998; Tajika, 1999; Yoshida, 2003, 2005). These properties inherent in “elaborative inferences” may have diminished differences in reading processes between higher and lower L2 proficiency learners.

When one or more sentences appeared between the two pieces of information connected by readers, the inference was labelled “a global connecting inference” in the
current study (see section 3.1). The distance made it different from “local connecting inferences.” The passage in which a t-test did not show the significant difference in the amount of “global connecting inferences” between groups was Passage 3. Passage 3 was much longer than Passage 1 and Passage 2, and much more difficult than Passage 1. One possible interpretation was that textual factors such as text length and text difficulty measured by the readability index might have affected the results. That is, when the reading material is longer, even higher L2 proficiency learners may not be able to construct global coherence effectively.

Let us begin with a discussion of the other results hereafter. Since these are not the main focus of the study nor are concerned with the hypotheses, I will limit my discussions to some main findings. As is shown in Table 3.13, a certain pattern of “Upper” learners’ inference generation emerged in two passages; their number of “local connecting inferences” was significantly greater than “global connecting inferences” and “elaborative inferences.” However, no fixed pattern was found in “Lower” learners; the number of “elaborative inferences” was significantly greater than “local connecting inferences” in Passage 2, and “local connecting inferences” were greater than “elaborative inferences” in Passage 3. The magnitude correlations were totally reversed by passage. On some occasions, “Lower” may adopt some reading strategy such as using their background knowledge to compensate for their impaired understanding.

Furthermore, the significant interaction between inference type and passage indicates that the type and the number of inferences generated by the participants relied on reading texts. In particular, “elaborative inferences” marked this tendency. As is mentioned in section 3.2.2, the present study selected passages from EFL or ESL reading textbooks whose text structures and genres are familiar to those learners. However, we can derive the
conclusion from the results that we should concern ourselves with selection of passages, particularly when we study readers’ reading processes.

Finally, although the current study could reveal entire pictures of L2 learners’ mental representations constructed during the course of or after reading general passages, there are still some issues to be addressed for further study in this area. First, I believe that an on-line measure offers a particularly promising approach to conduct research with a focus on readers’ inference generation processes; second, how inferences are related to identifying propositional meanings of text is still not known; and third, to find an effective way of activating learners’ inference generation processes is requisite when we consider the context of English language teaching. These issues are succeeded by the remaining studies in this dissertation.

3.5 Summary

Study 1 investigated the types and number of inferences generated by Japanese EFL learners during the course of or after reading. It was considered that the information which has not been explicitly written in texts but has appeared on the participants’ protocols originated from their inferences. Based on van den Broek et al. (1993), three categories were established: “local connecting inferences,” “global connecting inferences,” and “elaborative inferences.” The main findings were as follows: (a) There was a relationship between readers’ L2 proficiency and their inference generation processes, (b) the higher L2 proficiency learners generated more “local connecting inferences” than the lower L2 proficiency learners, and (c) the number of “elaborative inferences,” on the contrary, was hardly related to the learners’ L2 proficiency.

“A connecting inference,” the number of which differed between the L2 proficiency
groups, is one which connects some pieces of information within a few sentences in a text. This type of inference contributes to making a text coherent. “An elaborative inference” is, in contrast, defined as that which adds more information to a text, such as why, when, where and how a certain character in a story performed a given action, to a text by using readers’ own prior knowledge. There was no difference in the number of these between the different L2 proficiency levels.

To summarize the results of Study 1, the mental representations constructed by the higher L2 proficiency learners differed from those constructed by the lower L2 proficiency learners. These results imply that higher L2 proficiency learners are able to construct a more coherent situation model than lower L2 proficiency learners.

Endnotes

1. Earlier versions of this chapter have been published in Shimizu (2002, 2003a).
2. Appreciation is expressed to Professor Akihiko Mochizuki for permission to use the test.
3. Instructions and examples concerning a written recall were mostly adopted from Kimura and Yoshida (2001).
4. However, to use the proportion for a data entry must be useful to consider individual differences in the total number of inferences. I believe we should also encourage further research using this method.
5. Z transformation was applied for the computation of means.
6. The following equation for Hedges’s $g$ is for the two independent samples (Kline, 2004, pp. 101-104):

$$g = (M_1 - M_2)/s_p,$$

79
where $M_1$ and $M_2$ are, respectively, the population means expressed in raw (original measurement) unit, and

$$s_1^2 = \frac{SS_w}{df_W} = \frac{((n_1-1)s_1^2 + (n_2-1)s_2^2)}{n_1 + n_2 - 2},$$

where $SS_w$ and $df_W$ are, respectively, the total within-groups sums of squares and degree of freedom, $n_1$ and $n_2$ are the group sizes, and $s_1^2$ and $s_2^2$ are the group variances.

Cohen’s (1988) interpretive guidelines for this standardized mean difference are as follows: $0.200 < g$ (small) $< 0.500$; $0.500 < g$ (medium) $< 0.800$; $0.800 < g$ (large).

7. The present study used $\eta^2$ as an effect size measure in ANOVA, which is calculated as the ratio of the effect variance ($SS_{\text{effect}}$) to the total variance ($SS_{\text{total}}$): $\eta^2 = SS_{\text{effect}}/SS_{\text{total}}$. 


4. Study 2: Causal Antecedent Inference Generation and Questioning

4.1 Purposes and Research Hypotheses

Study 1 reported that, in terms of inference generation processes, what differentiated lower L2 proficiency learners from higher L2 proficiency learners was the amount of “local connecting inferences.” “Global connecting inferences” and “elaborative inferences” followed these. In response to these results, one type of inference which belongs to “connecting inferences” (or “bridging inferences”), “a causal antecedent inference” (Graesser & Kreuz, 1993; Graesser et al., 1994; see section 2.2.1), will be examined more carefully in Study 2.

The first aim of this study is to develop further the attempt to clarify the inference generation processes of EFL learners. Another reason that I targeted this inference type is that this inference particularly is known to play an important role in narrative comprehension. In accordance with Graesser et al. (1994), the inference is defined as follows: “The inference is on a causal chain (bridge) between the current explicit action, event, or state and the previous passage context” (p. 375). Take a look at one of the reading materials which will be used in the present study (see Passage 3 in Appendix 4-A).

Once there was an old fisherman who had a pet trout named Henry that he kept in the bathtub. The man was tired of changing the water of the tub quite often. So he began to take Henry out of the tub for a few minutes at a time. Soon Henry could stay out of the tub for a long time. One day, the fisherman started for town as usual with Henry coming along behind him. As the man crossed a bridge, he noticed that Henry had suddenly disappeared. The man could see through a hole in a bridge Henry was dead in the water below. He’d fallen through the hole into the stream and drowned.

(Maruyama, 1993a, pp. 38-39)
In this context, some “causal antecedent inferences” are expected to be made at the final sentence, such as something like, “because the owner made his pet get used to staying out of the tub, the pet became unable to swim.” These inferences are “causal antecedent inferences,” and are relevant to the third and the fourth sentences.

The second aim of this study is to empirically examine the effects of questions on learners’ inference generation processes. In particular, “why” questions used within a question-answering methodology (see section 2.3) were employed. “Why” questions have been proposed as questions which can elicit readers’ understanding of “causal antecedents,” an inference focused on in the present study.

Before conducting the research, five hypotheses, Hypotheses 2, 3, 4, 5A, and 5B, were formulated. In the remainder of this section, these are introduced one by one. Hypothesis 2 and Hypothesis 3 concern the inspection of the effects of the questions.

Hypothesis 2. The number of generated causal antecedent inferences is increased by “why” questions.

Hypothesis 3. The number of idea units whose propositional meanings are successfully constructed during a second reading with “why” questions is larger than that during a first reading without a question.

Furthermore, in terms of Hypothesis 3, it is important to note what was discussed in section 2.3. I assume that those “why” questions will have positive effects, not only on making inferences, but also on understanding the information which is closely related to the inferences. This expected effect is an indirect outcome because the target of the question is, above all, inference making.
The other hypotheses concern correlation analyses. Considering the inference’s proximity to propositions in particular parts of a text, Hypothesis 4 was formed. In other words, it is predicted that there would be a high correlation between generating causal antecedent inferences and understanding propositional meaning in a text which are to be connected by the inferences

Hypothesis 4. The number of generated causal antecedent inferences highly correlates with the number of idea units whose propositional meanings are successfully constructed.

Hypothesis 5A and Hypothesis 5B were constructed following Study 1, which reported a strong relationship between readers’ L2 proficiency and their generation of connecting inferences. However, the relations are not still well known as to whether strong relations can be seen between readers’ L2 proficiency and the inferences generated on-line, or only off-line. This is because Study 1 employed an off-line measure, that is, a written recall protocol (see section 2.2.2). Therefore, the current study employs an on-line measure, that is, a think-aloud procedure. This is a promising approach for observing EFL learners’ on-line inference generation processes.

Hypothesis 5A. L2 reading proficiency moderately correlates with the number of causal antecedent inferences generated before “why” questions.

Hypothesis 5B. L2 reading proficiency moderately correlates with the number of causal antecedent inferences generated after “why” questions.

The foci of Hypotheses 5A and 5B are as follows: (a) whether there is a moderate
correlation between readers’ L2 reading proficiency and inference generation before “why” questions, indicating the relation with inferences generated on-line, or (b) whether there is not a correlation between readers’ L2 reading proficiency and inference generation before “why” questions, but is a moderate correlation after “why” questions, indicating the relation with inferences generated off-line. As was stated in section 2.2.1, the present study defines “on-line” as “during the course of comprehension,” and “off-line” as “during a later retrieval task but not during comprehension” (Graesser et al., 1994, p. 371). Furthermore, in the present study questioning is considered to be “a later retrieval task.”

Moreover, with regard to the degree to which Hypotheses 5A and 5B correlate, the criterion of “moderate” was formulated due to the results of Study 1. Study 1 showed a strong relationship between learners’ L2 proficiency and connecting inference generation. In later sections, I will add a supplementary explanation, citing Kiyokawa (1990), of the interpretation of the correlation coefficients. Also, in order to narrow the target, Study 2 measures readers’ L2 reading proficiency rather than their overall L2 proficiency.

4.2 Method

4.2.1 Participants

Forty-eight undergraduate students from two national universities in Japan voluntarily participated in this study. Their university year level ranged from the first to the fourth, and their majors were English language education, international studies, agriculture, and other fields, not biased in favor of one particular academic field (see Table 4.1). They had received a formal English language education for six or more years.
Table 4.1

Academic Fields of the Participants in Study 2

<table>
<thead>
<tr>
<th>Academic Field</th>
<th>Number of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. English Language Education</td>
<td>20</td>
</tr>
<tr>
<td>2. International Studies</td>
<td>12</td>
</tr>
<tr>
<td>3. Agriculture</td>
<td>6</td>
</tr>
<tr>
<td>4. Natural Science</td>
<td>3</td>
</tr>
<tr>
<td>5. Engineering</td>
<td>2</td>
</tr>
<tr>
<td>6. Comparative Culture</td>
<td>3</td>
</tr>
<tr>
<td>7. Library Science</td>
<td>1</td>
</tr>
<tr>
<td>8. Biology</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>48</strong></td>
</tr>
</tbody>
</table>

*Note.* These majors are arranged from the largest group to the smallest group.

The data from only 37 of these 48 students were analyzed, with data from 11 students being dropped for the following reasons: (a) First, it was noticed by the researcher that at least three students seemed to have had an uneasy time of the think-aloud task; and then, it was revealed by a questionnaire that all of these students had felt the task to be *absolutely difficult*, (b) second, considering the individual differences, it was determined that all of the data of the participants who marked to be *absolutely difficult* on the question which inquired about the task (see Appendix 4-B), should be excluded. As a result of (a) and (b), a total of five students were excluded from all of the following protocol analyses, (c) third, the data of a student who had been living outside Japan for about 10 years was also dropped from all the following protocol analyses because her English reading skill seemed to be automatized according to the researcher’s observation, (d) fourth, the data of the four students who had already read either of the passages used for analysis were not entered, and (e) lastly, one student was not able to do thinking aloud smoothly for the first passage she started at; therefore, this data was excluded. Thus, 37 students’ protocols remained for analysis.
4.2.2 Materials

The following three types of materials were used: (a) parts of the reading comprehension sections of Test of English as a Foreign Language (TOEFL) practice test (Educational Testing Service, 1999), (b) five narratives for a think-aloud task, and (c) a questionnaire. In the following paragraphs, some explanations will be offered as to these materials.

First, three passages and 30 multiple-choice questions with four options were adopted from TOEFL reading comprehension sections (Educational Testing Service, 1999; pp. 40-41, pp. 142-144, pp. 146-147). TOEFL has been developed to make an assessment of examinees’ English language ability in the academic field, and has often been required of EFL candidates for colleges and graduate schools in North America, Canada, and other English speaking countries (Educational Testing Service, n.d.). Since the students in the current study were EFL university students, this test was considered to be appropriate for the L2 reading proficiency test. The topics of the texts were arts, biology and economics; and these were predicted to be relatively unfamiliar to all of the participants in advance of the test.

Second, five humorous narrative stories (see Appendix 4-A) were selected from EFL or ESL reading textbooks. They were used to investigate the participants’ reading processes, or more specifically, to study how they process causal-effect relationships in passages. The reason why humorous stories were chosen was that this genre had often been seen in previous reading research which focused on readers’ inference generation such as Horiba (1993, 1996), Miyaura (1998), Tajika (1999) and Yoshida (2003). Among the five passages in the present study, the text was frequently organized that, in order to understand the punch lines, readers should have constructed causal antecedent inferences immediately after the
Material selection was made through a series of discussions between a psychology major graduate student, who is Japanese and had been studying inference generation during reading in the first language (L1), and the author.

Table 4.2

<table>
<thead>
<tr>
<th></th>
<th>Passage 1</th>
<th>Passage 2</th>
<th>Passage 3</th>
<th>Passage 4</th>
<th>Passage 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Word</td>
<td>90</td>
<td>116</td>
<td>105</td>
<td>112</td>
<td>98</td>
</tr>
<tr>
<td>Flesch Reading Ease&lt;sup&gt;1&lt;/sup&gt;</td>
<td>81.9</td>
<td>85.6</td>
<td>79.3</td>
<td>90.6</td>
<td>83.8</td>
</tr>
<tr>
<td>Flesch-Kincaid Grade Level&lt;sup&gt;1&lt;/sup&gt;</td>
<td>5.0</td>
<td>4.9</td>
<td>4.8</td>
<td>4.0</td>
<td>5.0</td>
</tr>
<tr>
<td>Number of IUs Sandwiched Between Connected Parts</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Topic</td>
<td>A wine festival</td>
<td>Jim and a cleaning woman, Mrs. Roper</td>
<td>An old man who started to keep his pet on the ground</td>
<td>A wife’s revenge</td>
<td>Mrs. Kim’s affection toward her daughter</td>
</tr>
</tbody>
</table>

Note. 1. These indices were provided by Microsoft Word 2000’s readability measurement tools.  
2. This passage was originally in Hill (1977, as cited in Tajika), and was modified by Tajika in order to “leave some room for inferences” (Tajika, p. 135).  
3. Some modifications were made by the researcher to adjust readability and passage length.

Materials were modified through a pilot study, and a discussion with a native speaker of English who had been studying sociology at a graduate school in Japan. The main modifications were as follows: (a) Passage lengths were shortened, and (b) some literary words and infant words were replaced by general words; for example, “brook” was replaced by “stream,” and “tummy” was replaced by “stomach.”
The stories were around 100 words in length each, and ranged from 4.0 to 5.0 in readability measured by the Flesch-Kincaid formula (see Table 4.2). However, it should be noted that the data elicited from four passages only, Passages 1, 2, 3, and 5, was analyzed, with Passage 4 excluded. The reason was as follows: The number of students who did a think-aloud task properly in all five passages was only 32 because many participants performed poorly on producing protocols on Passage 4. Therefore, it was determined that the data of 37 students should be entered for the following analyses, after the data of Passage 4 was excluded.

The third piece of material used in this study was a questionnaire for participants made by the researcher (see Appendix 4-B). It inquired familiarity with the passages used for the think-aloud task (from question 1 to question 5), the degree of difficulty in thinking aloud during reading (question 6), the reasons why they chose their answer to question 6 (question 7), and their background such as experience in living in foreign countries and some certificates of English language tests. The first five questions were something: “Did you know the contents written in “A Wine Festival,” the first passage, before you read it?” (translation mine); and the participants were required to respond with yes or no. Question 6 was a 5-point Likert-type scale, ranging from scale 1 (absolutely easy) to scale 5 (absolutely difficult). Participants were required to choose either one. Question 7 was an open-ended, short answer question. They were able to write freely as much as they wanted to about a think-aloud task. This questionnaire was prepared, firstly, to exclude the data of students who had read the passage before, and, secondly, to study how a think-aloud assessment functioned in the present study.
4.2.3 Data Collection Procedures

Before the main study, a preliminary study was conducted in order to determine the clarification of the explanation for a think-aloud task, appropriateness of reading passages, and the length of time for a series of steps. Three students in the first year of undergraduate school participated in the pilot study. They were in the same national university that some of the participants of the main study also belonged to. After the pilot study, the words in rubrics were partially revised for the main study.

Let us return to the procedures of the main study. This research was undertaken on an individual basis. First of all, the L2 reading proficiency test was administered for 33 minutes. After that, the participants underwent think-aloud tasks for about 60 minutes. Lastly, a questionnaire was answered for a few minutes. A short break was given between the L2 reading proficiency test and the think-aloud task, if a participant wanted it.

Since all of the participants were thought to be unfamiliar with a think-aloud protocol, the procedures were accompanied by some steps as in the previous studies such as Kaiho and Harada (1993), Horiba (1996), and Coté and Goldman (1999). Prior to the task, it was explained why they would be requested to be engaged in thinking aloud, and how this would proceed (see Appendix 4-C); and then, a demonstration of the technique was given by the researcher by using a short passage (see Appendix 4-D); moreover, the participants had some practice with another short passage (see Appendix 4-E).

During thinking aloud period, sentences were presented on the computer screen using Microsoft Office PowerPoint (Version 2000), and participants were able to read them at their own pace by pressing a key. In other words, their reading speed was neither measured nor controlled. The unit for presentation was basically a sentence. A new sentence was presented in bold face in addition to old sentences in regular face, thus allowing the readers
to go back to the old sentences at anytime. As is shown in Table 4.3, a question was given orally by the researcher when the participants finished the first reading. One of the interests of the present study was to investigate how EFL learners react to “why” questions. After the question, they started reading the same passage again. These procedures were repeated five times with the order of text presentation counterbalanced. All of the protocols including answers to “why” questions were recorded on cassettes.

Table 4.3

<table>
<thead>
<tr>
<th>Phase</th>
<th>Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The participants started to speak aloud their thoughts while looking at a passage on screen while pressing a key.</td>
</tr>
<tr>
<td>2</td>
<td>When they had finished the whole passage, they told the researcher that they had finished.</td>
</tr>
<tr>
<td>3</td>
<td>The researcher orally gave a “why” question, “why do you think the event in the last sentence occurred?” (translation mine) in Japanese (i.e., L1).</td>
</tr>
<tr>
<td>4</td>
<td>The participants answered the question orally in Japanese. If they had no idea, they were able to skip the question.</td>
</tr>
<tr>
<td>5</td>
<td>The participants reread the same passage, thinking aloud.</td>
</tr>
</tbody>
</table>

Note. These procedures were repeated for each passage.

After this task, the questionnaire was filled in. Some of the participants also had an informal interview about their responses to the questionnaire. During all of these procedures, the researcher stayed in the same room with the participants. This study was conducted over a 4-month period, from early July 2003 until late October 2003.

4.2.4 Transcribing and Scoring

With regard to the L2 reading proficiency test, all of the items were multiple-choice
questions, and one point was given for the correct answer. They were scored by the author. Item discrimination of the test was also examined.

As to the protocol data, firstly, what the participants verbalized was transcribed by the researcher by listening to the audiotapes and using a transcriber (SONY B1-85). Transcription was limited to the parts which would fall under scoring for the following procedures. Following this, the protocols were independently scored by two raters (a non-native speaker of English who was in an undergraduate course in the department of English language education, and the author) with reference to a scoring guide (see Appendix 4-F). The raters concentrated their attention on the following points: (a) whether a causal antecedent inference was made at the target point before a question, (b) whether a causal antecedent inference was made at the target point after a question, (c) whether each target IU was correctly processed during the first reading before a question, and (d) whether each target IU was correctly processed during the second reading after a question. The protocols collected in “phase 1” in Table 4.3 were used for the judgment on (a) and (c); the protocols collected in “phase 4” were used for the judgment on (b); and the protocols collected in “phase 5” were used for the judgment on (d).

The judgments on (a) and (b) were formed as zero or one, and each passage had one target inference; therefore, a full score was 4.000. As to (c) and (d), on the other hand, 19 IUs were identified. The criteria for dividing passages into IUs were made by the author through adopting Stein and Glenn’s (1979), and revising Muramoto’s (1998, pp. 104-105) one in order to apply it to an English passage (see Appendix 4-G). Each IU was dichotomously scored; a full score was 19.000 for (c) and (d). Regarding scoring of IUs, the scores given by the two raters were averaged for a final score, whereas all discrepancies in inference scores were resolved through discussions by the raters.
Table 4.4

Frequency Levels of Words in Target IUs

<table>
<thead>
<tr>
<th>Frequency Level</th>
<th>Number of Words</th>
<th>Word (Passage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>119</td>
<td>they (P1), decoded (P1), that (P1), each (P1), of (P1), the (P1), best (P1), he (P1), had (P1), into (P1), it (P1), so (P1), that (P1), there (P1), should (P1), be (P1)…</td>
</tr>
<tr>
<td>2000</td>
<td>14</td>
<td>empty (P1), bottle (P1), wine (P1), plenty (P1), pure (P1), whenever (P2), breathe (P2), medicine (P2), stream (P3), found (P4), wake (P4), wake (P4), glad (P5), clever (P5)</td>
</tr>
<tr>
<td>3000</td>
<td>1</td>
<td>drowned (P3)</td>
</tr>
<tr>
<td>4000</td>
<td>2</td>
<td>a.m. (P4), a.m. (P4)</td>
</tr>
<tr>
<td>5000</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>6000</td>
<td>4</td>
<td>feast (P1), coughing (P2), dusty (P2), cough (P2)</td>
</tr>
<tr>
<td>7000</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>8000</td>
<td>2</td>
<td>tub (P3), tub (P3)</td>
</tr>
<tr>
<td>0</td>
<td>5</td>
<td>Henry (P3), Henry (P3), Kate (P5), Kim (P5), Joan (P5)</td>
</tr>
<tr>
<td>Total</td>
<td>147</td>
<td>--</td>
</tr>
</tbody>
</table>

Note. Frequency levels are based on JACET List of 8000 Basic Words (JACET Basic Words Revision Committee, 2003).
P = Passage.
Because the number of the words at the 1000 frequency level is quite large, only some of them are shown in this table.
When the same word appeared twice, I counted it as two.

Furthermore, the frequency levels of all of the words in the 19 IUs were investigated with reference to the JACET List of 8000 Basic Words (JACET Basic Words Revision Committee, 2003). More than 90% of the total were within the 3000 frequency levels, about 5% of the total were between the 4000 to the 8000 frequency levels (see Table 4.4), and the remainder were proper nouns. Although some seemed to be relatively difficult for the participants, such as feast, cough, and tub, the purpose of the current research was to inspect how much the participants’ ability to catch propositional meaning of IUs would be improved
by a question; therefore, those words of some difficulty were left without revision as well as the other words.

The first draft of the scoring guide was made by the author, and underwent several revisions through discussions with three non-native speakers of English. Two of them had specialized in English language education, and the other had specialized in L1 reading in the field of psychology. All were graduate students. Target IUs, defined as those that should have been integrated by inferences, were also determined when making this scoring guide.

### 4.2.5 Data Analyses

Concerning the L2 reading proficiency test, item analysis was conducted to identify and eliminate nonfunctional items with low item-discrimination indexes. Then, two nonparametric tests, the Kruskal-Wallis test, which corresponds to a one-way ANOVA, and the Mann-Whitney U test, which corresponds to a $t$-test, were used to verify that there were significant differences in the scores among the three groups divided by the L2 reading proficiency test.

Moreover, a chi-square ($\chi^2$) test was employed to examine the association between the participants’ attitudes toward a think-aloud task seen on a questionnaire and their L2 reading proficiency. This procedure was performed to confirm that good performers were not disproportionately grouped into the higher level of L2 reading proficiency group. In other words, it was to avoid some explanations in advance, such as some positive results from the protocol data must have been due to the students’ good performance of the task, not having higher L2 reading proficiency.

I employed a Wilcoxon matched-pairs signed-rank test to test Hypotheses 2 and 3. With regards to these analyses, the Monte Carlo $p$ value was adopted when it was available.
This is because there are several major advantages to using the Monte Carlo method as opposed to using the asymptotic $p$ value for inference such as “unbiased” (SPSS, 1996, p. 28). Exact tests were unavailable on the current occasion with the statistical package showing such a limitation. Moreover, it should be noted that nonparametric approaches were consistently undertaken in all of the procedures for analysis considering the number of participants.

The inter-rater reliability of scoring procedures, and the answers to Hypotheses 4, 5A, and 5B were determined by the Pearson product-moment correlation. Only coefficients which were statistically significant at the .05 level were discussed in this study. We know that the criteria for interpreting correlation coefficients largely depend on the academic fields which researchers work in. Thus, among many other criteria (e.g., Cohen, 1992), the present study will use those of Kiyokawa (1990). According to this, the coefficients $|0.700|$ and greater are regarded as “high,” the coefficients between $|0.400|$ and $|0.699|$ are regarded as “moderate,” and the coefficients between $|0.200|$ and $|0.399|$ are regarded as “low.” In addition, coefficients $|0.199|$ and below are regarded as having almost no correlations.

Most of the data was analyzed using SPSS program for Windows (Version 10.0E), and only the $\chi^2$ test was performed by using Java Script-STAR, the free downloaded software (Version 3.6.4J and 3.6.6J; Tanaka, n.d.).

4.3 Results

4.3.1 Participants’ Responses to a Questionnaire

This section will deal with the responses of all 48 students who took part in the present study. This was to reveal how this task was conducted by L2 learners in general, and how
the five students whose data was finally dropped (see section 4.2.1) responded to this questionnaire. The original questionnaire sheet is presented in Appendix 4-B.

Answers concerning familiarity with each reading text are summarized in Table 4.5.

Table 4.5

<table>
<thead>
<tr>
<th>Question</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (as to Passage 1)</td>
<td>0</td>
<td>48</td>
</tr>
<tr>
<td>2 (as to Passage 2)</td>
<td>0</td>
<td>48</td>
</tr>
<tr>
<td>3 (as to Passage 3)</td>
<td>4</td>
<td>44</td>
</tr>
<tr>
<td>4 (as to Passage 4)</td>
<td>2</td>
<td>46</td>
</tr>
<tr>
<td>5 (as to Passage 5)</td>
<td>0</td>
<td>48</td>
</tr>
</tbody>
</table>

*Note. N = 48.*

Except for Passage 3 and Passage 4, none had read the passages. In contrast, four students for Passage 3 and two students for Passage 4 answered that they were already familiar with the passages. Following interviews with those students revealed that they might have read them during high school classes. The data of the students who answered *yes* to these questions was excluded for statistical analysis because they were reasonably considered to have different amounts of background knowledge from the others.

The participants’ responses to question 6 were dispersed throughout 5 choices. The choice which attracted the largest number of the students (i.e., 22 out of 48 students) was *quite difficult*, and the second largest number of the students answered *quite easy* (i.e., 13 out of 48 students). Four students responded with *absolutely easy*, and five students responded with *absolutely difficult*. The remainder of the four students answered *either way*. I can summarize the results as follows: (a) for the present participants, thinking aloud was more likely to be difficult than not, and (b) in spite of that, we can see an appreciable
individual difference in reactions to this task.

Figure 4.1. Participants’ attitudes toward think-aloud protocols: Question 6. The numbers around the graph indicate the number of respondents. N = 48.

As to question 7 (free writing pertaining to this task), only outstanding remarks will be mentioned (for more information, see Appendix 4-H). Five students responded, absolutely difficult to thinking aloud. Three of them stated, “Because I usually do not do so.” (Students AR, AS and AT; translation mine). And the other two found difficulty in doing two different tasks simultaneously, that is, reading and thinking aloud (Student AU and Student AV). As was mentioned in section 4.2.1, the protocol data of these five students were not analyzed because they were not able to reveal their natural processing.

Among the students who selected 4 (i.e., quite difficult), seven gave the following reason, “Because I usually do not do so.” (Students W, X, Z, AA, AB, AJ and AO). Four students pointed out that some comprehension was represented beyond words (Students V,
AG, AL and AN), which has been described as a limitation of this measurement. The other six students (Students Y, AD, AH, AK, AM and AN) gave us some insightful comments from the perspective of their developed metacognitive awareness of reading processes; that is, simultaneous parallel processing and automaticity of processing.

Four students responded with either way. One of them offered the following reason, “Because I usually do so.” (Student S). The others referred to the benefit from practice (Student R), and expressed a positive opinion such as “I could recognize my thoughts by thinking aloud.” (Student T).

A total of 13 students judged the task to be quite easy. Five of them answered, “Because I usually do so.” (Students H, I, N, O and Q), and two answered, “Because what I needed to do was only to say what I was thinking about.” (Students G and K), reflecting what this task expected the participants to do (Harada, 1993, p. 83). In addition, “Because I could do it in a relaxed manner.” was found in a student’s (Student E) answer. To create a relaxed atmosphere is one of the prerequisites for administering think-aloud research (Harada, p. 84). Some answers, furthermore, showed positive opinions such as “I could concentrate on reading by doing this task.” (Student L), and benefits from the practice phase such as “I am now familiar with the task.” (Student P).

The answers of the students who marked absolutely easy were similar to those mentioned above. Among the four respondents, three wrote, “Because I usually do so.” (Students A, C and D). Moreover, two students answered, “Because what I needed to do was only to say what I was thinking about.” (Student B and Student C).

These responses to question 7 can be summarized as follows:

1. Whether they often think aloud or not in everyday life may be significantly related to whether participants feel a think-aloud task is easy or difficult. Eight students out of 17
who marked *absolutely easy* or *quite easy* pointed out that it was related to their habits, whereas 10 students out of 27 who marked *absolutely difficult* or *quite difficult* answered that they did not do so in their daily lives.

2. Ideal conditions and some limitations of this measurement which have often been discussed in previous studies were found in the participants’ comments as well. It seems to reflect students’ higher metacognitive awareness toward reading processing.

### 4.3.2 Validity and Descriptive Statistics of L2 Reading Proficiency Test

The participants’ scores on the L2 reading proficiency test were employed as an instrument for dividing 37 students into three different L2 reading proficiency groups. To begin with, the corrected item-total point biserial correlation coefficients were computed; and then, the values of Cronbach’s alpha for the entire scale (with each item deleted in turn) were computed for the construct validation from the structural aspect (see section 3.3.1). The validation from the structural aspect was adopted in this case because the data for the analysis were available. As is shown in Appendix 4-I, out of 30 items, five items (items 3, 5, 12, 13, and 21) resulted in point biserial correlation coefficients below .250 (Henning, 1987). That is, 83% of the total items were over Henning’s criterion of acceptability. Moreover, the value of Cronbach’s alpha was high, \( \alpha = .848 \). Thus, the positive evidence for the validity of the L2 reading proficiency test was presented by these statistics.

To assess the participants’ L2 reading proficiency, the 25 items, whose point biserial correlation coefficients were over .250, were left remaining. The newly computed Cronbach’s alpha increased by .017 to .865. In addition, the following well-known scheme (e.g., Matsuno & Negishi, 1998) was used for an interpretation of scores on a paper and pencil TOEFL: (a) The upper group (“Upper”) consisted of the students whose L2 reading
proficiency was approximately above the prerequisite for undergraduate students of university in English speaking countries, that is, 550 and more, (b) the middle group (“Middle”) consisted of the students whose L2 reading proficiency was approximately above the prerequisite for community college students in English speaking countries, especially in the U.S. and Canada, with TOEFL scores from 480 to 549, and (c) And lastly, the lower group (“Lower”) consisted of the students whose L2 reading proficiency did not reach the prerequisite for university students in English speaking countries, that is, 479 and less. The participants’ scores on the L2 reading proficiency test were converted into those on a paper and pencil TOEFL test with reference to the transformation table (Educational Testing Service, 1999, p. 204). Participants were assigned to one of the groups.

As Table 4.6 indicates, the mean scores for “Lower,” “Middle,” and “Upper” were 7.563 ($SD = 3.366$), 14.538 ($SD = 1.450$), 20.250 ($SD = 2.188$), respectively.

Table 4.6

<table>
<thead>
<tr>
<th>Group (the paper and pencil TOEFL)</th>
<th>n</th>
<th>Mean</th>
<th>$SD$</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower (310-479)</td>
<td>16</td>
<td>7.563</td>
<td>3.366</td>
<td>1.000</td>
<td>12.000</td>
</tr>
<tr>
<td>Middle (480-549)</td>
<td>13</td>
<td>14.538</td>
<td>1.450</td>
<td>13.000</td>
<td>18.000</td>
</tr>
<tr>
<td>Upper (550-670)</td>
<td>8</td>
<td>20.250</td>
<td>2.188</td>
<td>19.000</td>
<td>25.000</td>
</tr>
<tr>
<td>Total (310-670)</td>
<td>37</td>
<td>12.757</td>
<td>5.654</td>
<td>1.000</td>
<td>25.000</td>
</tr>
</tbody>
</table>

*Note. Full score = 25.000.*

Significant differences in these means were confirmed by the Kruskal-Wallis test ($\chi^2 (2) = 31.406, p = .000$). The following Mann-Whitney U test showed that the means of two different and independent groups were significantly different, viz., (a) between “Upper” and “Middle” ($U = .000, p = .000 < .050/3 = .017$), (b) between “Middle” and “Lower” ($U = .000$, $p = .000 < .050/3 = .017$).
\( p = .000 < .050/3 = .017 \), and (c) between “Upper” and “Lower” \((U = .000, p = .000 < .050/3 = .017)\). As is shown, the Bonferroni approach, which uses a significance level of \( .050/k \) (group numbers) for each individual test, was adopted.

Next, a \( \chi^2 \) test was run to determine the strength of association between readers’ L2 reading proficiency, and their attitudes toward a task of thinking aloud, that is, the responses to question 6 on the questionnaire (see section 4.2.2; absolutely easy = 1, quite easy = 2, either way = 3, quite difficult = 4, and absolutely difficult = 5). This was because the present study analyzed the data elicited by a think-aloud task; therefore, it should have been confirmed that there was no association between readers’ L2 reading proficiency, and their attitudes toward the task beforehand. Students who answered absolutely easy and quite easy were combined and coded as easy, and those who answered absolutely difficult and quite difficult were combined and coded as difficult.\(^3\)

Table 4.7

<table>
<thead>
<tr>
<th>L2 Reading Proficiency Group</th>
<th>Easy</th>
<th>Either Way</th>
<th>Difficult</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower</td>
<td>8</td>
<td>2</td>
<td>6</td>
<td>16</td>
</tr>
<tr>
<td>Middle</td>
<td>5</td>
<td>0</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>Upper</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>8</td>
</tr>
</tbody>
</table>

The numbers of students are shown in Table 4.7. The \( \chi^2 \) test revealed that there was no association between the participants’ L2 reading proficiency and their attitudes toward a think-aloud task, \( \chi^2 (4) = 3.977 \), n.s. Therefore, prerequisites of the analysis were attained.
4.3.3 Results of Think-Aloud Protocols

To get an indication of interrater reliability, the Pearson product-moment correlations were computed. Both the coefficients for IUs and for inferences were very high, \( r = .931 \) (\( p < .001 \)) and \( r = .941 \) (\( p < .001 \)), respectively. To compute these statistics, I did not consider the following points: (a) whether they were the IUs understood during the first reading or the second reading, (b) whether they were the inferences generated before questions or after questions, and (c) in which passage IUs were understood or inferences were made. Thus, 74 pairs (\( = 37 \times 2 \)) of ratings for IUs whose maximum score was 19, and 74 pairs (\( = 37 \times 2 \)) of ratings for inferences whose maximum score was 4 were entered, respectively, into the correlation equation.

The reliability for inferences, measured by Cronbach \( \alpha \), was .553, indicating that they tapped the participants’ inference generations in a moderately reliable manner. Eight items (i.e., the before-question terms of Passages 1, 2, 3, and 5, and the after-question terms of Passage 1, Passage 2, Passage 3 and Passage 5) were included in the computation. It can be considered that the moderate coefficient is partially due to the total number of inferences targeted in the current research, that is, four inferences. Considering the limitations of the present methodology (i.e., think-aloud techniques) and text length (i.e., a longer text in comparison with studies which target inferences in a few sentences), it was determined to proceed without any further modifications.

With regard to IUs, a high coefficient, \( \alpha = .840 \), was obtained, with eight items (i.e., the first readings of Passage 1, Passage 2, Passage 3 and Passage 5, and the second readings of Passage 1, Passage 2, Passage 3 and Passage 5) entered. On that computation, the number of correctly reproduced IUs was taken for analysis, instead of the percentage score. I will describe the results of inferences first, and of IUs in the rest of this section.
When the participants’ L2 reading proficiency was not taken into consideration, an average of 1.568 (SD = 0.867) points were given to the 37 students before questions, whereas an average of 2.216 (SD = 0.976) points were given after questions. A Wilcoxon matched-pairs signed-rank test showed a significant difference between these means (z = -3.816, p < .001).

Table 4.8

Results of Wilcoxon Matched-Pairs Signed-Ranks Tests Which Compared Mean Numbers of Inferences

<table>
<thead>
<tr>
<th>L2 Reading Proficiency Group</th>
<th>Before</th>
<th>After</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>-3.816</td>
<td>.000</td>
</tr>
<tr>
<td>Lower</td>
<td>16</td>
<td>1.500 (0.633)</td>
<td>2.000 (1.033)</td>
<td>-2.126</td>
</tr>
<tr>
<td>Middle</td>
<td>13</td>
<td>1.462 (0.519)</td>
<td>2.000 (0.717)</td>
<td>-2.333</td>
</tr>
<tr>
<td>Upper</td>
<td>8</td>
<td>1.875 (1.553)</td>
<td>3.000 (0.926)</td>
<td>-2.264</td>
</tr>
</tbody>
</table>

Note. Full score = 4.000.
Monte Carlo significance level was employed for these analyses.
Before = before questions; After = after questions.

Mean scores and standard deviations were then computed after the participants had been divided into three different L2 reading proficiency groups. In terms of the statistics before questions, the average scores were 1.500 (SD = 0.633) for “Lower,” 1.462 (SD = 0.519) for “Middle,” and 1.875 (SD = 1.553) for “Upper,” whereas, in terms of the statistics after questions, the average scores were 2.000 (SD = 1.033) for “Lower,” 2.000 (SD = 0.717) for “Middle,” and 3.000 (SD = 0.926) for “Upper.” In order to investigate whether the effects of a question were different in each group, a Wilcoxon matched-pairs signed-rank test was repeated three times. As a result, statistical differences were found in all these three pairs at the level of p < .05 (see Table 4.8).
Furthermore, the Kruskal-Wallis tests and the Pearson product-moment correlations were employed; the former was to examine differences in mean scores across L2 reading proficiency levels, and the latter was to examine relationships between the participants’ scores on the L2 reading proficiency test and the amount of inferences generated during thinking aloud. The results of the Kruskal-Wallis tests were as follows: (a) No significant difference was found in the amount of inferences before “why” questions among the L2 reading proficiency groups, $\chi^2 (2) = 0.080, p = .978$, and (b) a significant difference, on the contrary, was found among the groups in the amount of inferences after “why” questions, $\chi^2 (2) = 5.835, p = .0495$. To follow the results of (b), a multiple comparison (the Mann-Whitney test) of Ryan’s procedure (Kiriki, 1990, pp. 171-172) was undertaken to determine where the significant differences were to be found. The significant difference was found between “Middle” and “Upper” ($U = 22.500, p = .023 < .033 = .05 \times 2/3 = 2$).

Figure 4.2. Inferences generated before “why” questions and those generated after “why” questions. Before = before questions; After = after questions.
\[ \times .05/(3 \times (2-1)), \text{ but none were found between “Lower” and “Upper” (} U = 31.500, p = .042 > .017 = .05 \times 1/3 = 2 \times \times .05/(3 \times (3-1)) \text{ nor between “Lower” and “Middle” (} U = 104.000, p = 1.000 > .033 = .05 \times 2/3 = 2 \times .05/(3 \times (2-1)). \]

The Pearson product-moment correlation coefficient between the scores on the L2 reading proficiency test and the amount of inferences before “why” questions did not reach a statistically significant level \( r = .290, \text{n.s.} \), but the coefficient between the scores on the L2 reading proficiency test and the amount of inferences after “why” questions was a moderate and significant one \( r = .533, p < .001 \). The results are shown in Table 4.9.

Table 4.9

**Pearson Product-Moment Correlations Among Five Measures**

<table>
<thead>
<tr>
<th>Measure</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. L2 Reading Proficiency Test</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Inferences Before Questions</td>
<td>.290</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(-.038 - .561)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Inferences After Questions</td>
<td>.533***</td>
<td>.606***</td>
<td>--</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.253 - .731)</td>
<td>(.351 - .777)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. IUs Before Questions</td>
<td>.604***</td>
<td>.623***</td>
<td>.706***</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.348 - .776)</td>
<td>(.375 - .788)</td>
<td>(.495 - .838)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. IUs After Questions</td>
<td>.606***</td>
<td>.534***</td>
<td>.660***</td>
<td>.892***</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>(.351 - .777)</td>
<td>(.254 - .731)</td>
<td>(.427 - .811)</td>
<td>(.799 - .943)</td>
<td></td>
</tr>
</tbody>
</table>

*Note.***p < .001.*

The values in parentheses indicate the lower and upper boundaries of the 95% confidence interval.

The results of targeted IUs will be described in the remainder of this section. The average numbers of IUs understood by the students during the primary reading without questions and during the second reading with questions are shown in Table 4.10. The
targeted IUs were determined from the following perspective: whether they were expected to be integrated with the last sentence by a hypothetical proficient reader by generating inferences. The total number of the targeted IUs was 19. Thirty-seven participants obtained 13.432 points (SD = 2.542) during the first reading before “why” questions, and 14.473 (SD = 2.705) during the second reading after “why” questions, on average. When I examined these results more carefully by allotting either one of the three L2 reading proficiency levels to the participants, the following descriptive statistics were obtained: the mean score for “Lower” increased by 0.719 to 13.094 (SD = 2.703), that for “Middle” increased by 1.462 to 15.077 (SD = 2.139), and that for “Upper” increased by exactly 1.000 to 16.250 (SD = 2.345).

Table 4.10

<table>
<thead>
<tr>
<th>L2 Reading Proficiency Group</th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>Mean (SD)</td>
</tr>
<tr>
<td>Lower</td>
<td>16</td>
<td>12.375 (2.540)</td>
</tr>
<tr>
<td>Middle</td>
<td>13</td>
<td>13.615 (2.191)</td>
</tr>
<tr>
<td>Upper</td>
<td>8</td>
<td>15.250 (2.188)</td>
</tr>
<tr>
<td>Total</td>
<td>37</td>
<td>13.432 (2.542)</td>
</tr>
</tbody>
</table>

Note. Full score = 19.000.
Monte Carlo significance level was employed for these analyses.
Before = before questions; After = after questions.

The results of the Wilcoxon matched-pairs signed-ranks tests are also shown in Table 4.10. There were significant differences between the pair of “Lower” (z = -2.225, p = .022), of “Middle” (z = -2.661, p = .006) and of “Total” (z = -4.031, p = .000), but not of “Upper” (z = -1.866, p = .078). However, it is important to note that the “Upper” result also showed a
tendency for the “Upper” students to benefit as well as the others from the second reading with “why” questions.

![Figure 4.3. Numbers of IUs understood during first reading without “why” questions and second reading with “why” questions. Before = before questions; After = after questions.](image)

The Kruskal-Wallis test, which was performed to investigate group differences in scores of IUs before “why” questions, revealed a statistically significant main effect for L2 reading proficiency level, $\chi^2 (2) = 7.619, p = .018$. By the Mann-Whitney test of Ryan’s procedure, a significant difference was found between “Lower” and “Upper” ($U = 21.000, p = .006 < .017$), but not between “Middle” and “Upper” ($U = 31.500, p = .133 > .033$), nor between “Lower” and “Middle” ($U = 70.500, p = .139 > .033$).

Similar results were obtained under the other condition (i.e., after “why” questions) as well: (a) A statistically significant main effect for L2 reading proficiency level was found using the Kruskal-Wallis test, $\chi^2 (2) = 9.443, p = .005$, and (b) the Mann-Whitney test of Ryan’s procedure revealed a significant difference between “Lower” and “Upper” ($U =
18.000, \( p = .002 < .017 \), but neither between “Middle” and “Upper” (\( U = 32.500, p = .164 > .033 \)), nor between “Lower” and “Middle” (\( U = 59.500, p = .051 > .033 \)).

The correlation analysis (the Pearson product-moment correlation) of the number of IUs whose meanings were successfully constructed and the number of generated inferences showed moderate correlations; to explain more precisely, \( r = .623 (p < .001) \) between the number of IUs during the primary reading without “why” questions and the number of generated inferences before “why” questions, and \( r = .660 (p < .001) \) between the number of IUs during the second reading with “why” questions and the number of generated inferences after “why” questions (see Table 4.9). Quite equivalent coefficients were obtained in spite of the different conditions (i.e., during the first reading without “why” questions or the second reading with “why” questions).

4.4 Discussion

Hypotheses presented in the first section of this chapter will be discussed one by one. Hypothesis 2 is that the number of generated causal antecedent inferences is increased by “why” questions. It was constructed to learn whether a certain operation, “why” questions in this case, functions effectively to change inference generation processes of EFL learners. Reading questions, in the present context, were not used for measuring learners’ reading ability, but for influencing learners’ reading comprehension.

As is mentioned in section 2.3, “why” questions originated from a question-answering methodology (e.g., Graesser & Clark, 1985; Graesser & Goodman, 1985; Graesser et al., 1994; Long et al., 1990) as well as the other two kinds of questions, “how” questions and “what-happened-next” questions. The present study, aiming to investigate the effectiveness of a question on learners’ causal antecedent inference generation, employed “why”
questions, which are designed to draw readers’ causal antecedent inferences. The Mann-Whitney U test determined that the mean number of inferences generated after “why” questions was significantly larger than that of the number of inferences generated before “why” questions. Therefore, Hypothesis 2 was supported. Moreover, the Mann-Whitney U test was run for each L2 reading proficiency group, and all of the groups showed significant differences in the amount of inferences between the period before and after questions (see Table 4.8).

As is shown in Table 4.10, the results of the Mann-Whitney U test supported Hypothesis 3 (i.e., the number of idea units whose propositional meanings are successfully constructed during a second reading with “why” questions is larger than that during a first reading without the questions). In addition, there was no difference in this result among L2 reading proficiency groups. Although “Upper” only showed such a tendency, we can consider it merely due to the small number of participants who belong to the group, and to the ceiling effect (Brown, 1996). If there had been a greater number of participants and a greater number of target IUs of more difficulty in the passages, it would have reached a statistically significant level. Therefore, the results of “Upper” are included in my discussion as well. These effects regarding this hypothesis can be stated as indirect effects of the question. Why they are indirect will be explained in the following paragraphs. We should also keep in mind that in the present case the effect of the question had contained the effect of second reading as well; therefore, my discussion about effects of “why” questions are effects of “why” questions in the second readings.

The result of Hypothesis 3 can be interpreted based on the connectionist model of narrative text comprehension proposed by Trabasso and his colleagues. Their model has been extensively introduced into previous reading studies (e.g., Horiba, 1993, 1996; Horiba,
van den Broek, & Fletcher, 1993; Miyaura, 1998; Muramoto, 1998; Trabasso & Sperry, 1985); and, according to Langston and Trabasso (1999), the latest model provides “a quantitative account for accessing and making available information from clauses over the course of understanding a text” (p. 30). Furthermore, what the model predicted was proven to be supported by empirical evidence in the Langston and Trabasso’s study. A sample of the product from their discourse analysis is shown in Figure 4.4.

![Network representation of two versions of a story](image)

**Figure 4.4.** Network representation of two versions of a story (Langston & Trabasso, 1999, p. 37).

Each node in that figure represents a clause or a sentence (e.g., a main character, Ivan, is introduced in S1; he learns that giant has been terrorizing and killing villagers in E1; therefore, it gave him a cretin motivation in G1), and an alphabet letter next to a node indicates the function of the node in the story plot (e.g., S = a setting). A connection, on the
other hand, represents a causal relation between the nodes. These connections are determined by the “counter factual reasoning” test or by the criteria of “weak sufficiency.” To be more specific, if event A did not happen, event B would also not have happened; or if event A was included in a text, event B would be included because of this. The former example was derived from the test of “counter factual reasoning,” and the latter example was derived from the criteria of “weak sufficiency.” Thus, the node that has many lines shows that it relates to many other parts of the text.

The network included in Figure 4.4 above represents the story, in which a character at first failed to attain a goal (node O1), but after that, he or she could finally achieve the goal (node O3), whereas the lower network included in the same figure represents the story, in which a character could achieve his or her goal (node O1) without any intervention. As you may know, the upper network in Figure 4.4 has a more complex structure than the lower network, as a result of more nodes.

Although the present study did not directly adopt the method of Trabasso and his colleagues’ discourse analysis—arithmetical computations of the activation value, to which each node was attached, or the connection strength, to which each line was attached (see Figure 4.5), the results obtained in the present study, Study 2 can be explained in parallel with their model. Nodes are applied to what the present study labelled as the target IUs, which were expected to be integrated by the inferences; connections, in contrast, are applied to what the present study labelled as “causal antecedent inferences.” Thus, “why” questions were designed to be used to make the EFL learners recognize the connections between integrated parts. However, the results appear to indicate an interesting relationship between the question and the IUs integrated by inferences. In sum, although the question originally targeted the learners’ connection making (i.e., inferences), it also activated their ability to get
the meanings of the nodes connected with lines (i.e., getting the propositional meanings related to each other in texts). That is why I called this effect more indirect in the earlier part.

Figure 4.5. A: The text representation at time $t$, consisting of six nodes and their connections. B: The text representation at time $t+1$, consisting of the original six nodes and their connections from time $t$, plus a new node and its connections (Langston & Trabasso, 1999, p. 39).

To test the following hypothesis, the Pearson-product moment correlation was employed. Regarding Hypothesis 4, the correlation coefficient was computed between the number of IUs whose propositional meanings were successfully constructed out of 19 IUs in total and the number of generated causal antecedent inferences out of 4 in total. The results indicated $r = .623 \ (p < .001)$ for the before-question condition and $r = .660 \ (p < .001)$.
for the after-question condition (see Table 4.9). Therefore, Hypothesis 4 that the number of generated causal antecedent inferences highly correlates with the number of idea units whose propositional meanings are successfully constructed was rejected. It was assumed that because these variables were closely related to each other theoretically, these coefficients were expected to be over .700; however, that did not turn out to be the case. Some learners may have complemented their deficient comprehension by random guessing as was shown in previous L2 reading studies (e.g., Hammadou, 1991; Tajika, 1999). In addition, the issue concerning interaction between reading processes was discussed by Horiba et al. (1993). Horiba et al’s recall data revealed that L2 readers’ deficiency of lower-level processes appeared to be complemented by their higher-level processing.

Hypothesis 5A that L2 reading proficiency moderately correlates with the number of causal antecedent inferences generated before “why” questions and Hypothesis 5B that L2 reading proficiency moderately correlates with the number of causal antecedent inferences generated after “why” questions will be discussed together, as they are in parallel. To test Hypothesis 5A, the correlation analysis was employed between the scores of the L2 reading proficiency test and the number of causal antecedent inferences correctly generated before “why” questions. The coefficient was \( r = .290 \) (see Table 4.9); however, considering that it was not significant at the .05 level (see section 4.2.5), we can say that there was no relationship between L2 reading proficiency and inference generation before “why” questions. Thus, Hypothesis 5A was rejected. In turn, as to Hypothesis 5B, \( r = .533 \) (\( p < .001 \)) was obtained as a correlation coefficient between the scores of L2 reading proficiency test and the number of causal antecedent inferences correctly generated after “why” questions (see Table 4.9). It fell in the range between \( .400 \) and \( .700 \) (Kiyokawa, 1990); therefore, it was determined that Hypothesis 5B was supported.
Before examining the difference in results between Hypothesis 5A and Hypothesis 5B, the discussion about on-line inferences and off-line inferences, which have been argued about in the last 15 years by the constructionist theory (Graesser & Kreuz, 1993; Graesser et al., 1994) and the minimalist hypothesis (McKoon & Ratcliff, 1992), should again be referred to. On-line inferences are defined as inferences which are generally constructed during the course of comprehension, probably, within 650 milliseconds after recognizing words or sentences. Off-line inferences, on the other hand, are defined as inferences which are generally constructed after comprehension or while undertaking a retrieval tasks. Thus, it is thought that making off-line inferences takes time, probably, in seconds (Graesser et al.). The discussions were carried out between the constructionist theory and the minimalist hypothesis about which inferences are made on-line and which are made off-line. Their predictions contradicted each other. In sum, the constructionist theory predicted many types of inferences to be on-line, including referential, case structure role assignment, causal antecedent, superordinate goal, thematic, and character emotional reaction; but the minimalist hypothesis, as is shown by its name, predicted only limited inferences to be on-line. The minimalist hypothesis considers only what is called automatic inferences--comprising of inferences which contribute to construct locally coherent representations and inferences which are generated by using easily available information (see section 2.2.1)--as on-line inferences. Even so, causal antecedent inferences, the target inferences of the present study, are one of the on-line inferences agreed on by both the constructionist theory and the minimalist theory (see Table 2.2); in other words, the constructionist theory and the minimalist hypothesis do not disagree about this type of inference.

However, it should be noted that these theories (or hypotheses) have been developed
by the outcomes obtained from L1 reading research. Therefore, those predictions may not necessarily be applied to the present study, which targeted L2 learners’ readings. Collins and Tajika’s (1996) study also reported this possibility. Learners’ L2 proficiency continues to develop, and, except for those who are extremely advanced learners, L2 learners have ongoing deficiencies in processing words and sentences. Considering these differences in processing between L1 readers and L2 readers, it should be concluded that some inferences defined as on-line inferences by these theories may not be for L2 learners. Thus, it would be impossible to interpret the present results directly from the viewpoint of the constructionist theory or the minimalist hypothesis, but it would be useful to employ the concepts of on-line inferences and off-line inferences, and interpret the results using these.

My expectation was that generation of causal antecedent inferences would moderately correlate with readers’ L2 reading proficiency. This was made on the basis of Study 1, and I expect that this would also be matched by teachers’ general knowledge. The results were that there was such a correlation between L2 reading proficiency and the inferences measured after questions, but not between L2 reading proficiency and the inferences measured before questions. These appear to indicate how L2 learners develop their ability to infer from a text. That is, L2 learners with L2 reading proficiency to some degree firstly become able to generate inferences in the off-line context; and then, as their L2 reading proficiency develops and becomes closer to L1 readers, they generate the inferences on-line. Although causal antecedent inferences are generally made on-line by L1, they may require much more time to be generated by L2 readers, or L2 readers need the help of retrieval tasks to make inferences.

Lastly, and this is not directly related to testing the hypothesis, the same tendency as in the results of Hypotheses 5A and 5B was found by the Wilcoxon matched-pairs
signed-ranks tests and the Mann-Whitney U test. By these nonparametric analyses, it was revealed that the scores on IUs reflected individual L2 reading proficiency more faithfully than the scores on inferences. In other words, whether the participants were able to understand the target IUs differentiated between “Upper” and “Lower,” which was the case in spite of the different conditions (i.e., both during the first reading with “why” questions and during the second reading without “why” questions); however, whether the participants were able to generate causal antecedent inferences differentiated between “Upper” and “Middle” after “why” questions, but did not differentiate any pairs before “why” questions. These results appear to give more evidence for the finding shown by the correlation analysed; that is, the relationship between readers’ L2 reading proficiency and the amount of causal antecedent inferences becomes stronger when the inferences are measured off-line. Repeatedly, it can be concluded that the process of the inference generation by L2 learners seem to be different in part from hypotheses derived from text comprehension models developed in the field of L1 reading.

In sum, Study 2 showed how inference making was related to the construction of propositional meanings, and how well “why” questions could influence learners’ inference making and the construction of propositional meanings. Particularly, the effectiveness of posing questions on learners’ reading comprehension was empirically shown to have profound pedagogical implications for instructors. However, when we consider the context of reading instruction in English language teaching classroom, generally a series of questions would be used. Study 2, on the other hand, only dealt with one type of question. Therefore, we still need to concern ourselves with how we should present a series of questions in order to make them more effective.
4.5 Summary

Study 2 has focused on causal antecedent inference generation processes (Graesser et al., 1994) of Japanese EFL learners. As is shown in Trabasso and his colleagues’ causal network model (Horiba, 1993, 1996; Horiba et al., 1993; Langston & Trabasso, 1999; Muramoto, 1998; Stein & Glenn, 1979), it is well known that a causal inference is closely related to narrative reading comprehension. Study 2 has also focused on “why” questions. Whether this inference question was able to affect learners’ inference generation processes was another interest of the present study.

The main findings are summarized into the following four points: (a) “Why” questions enabled the learners to generate more causal antecedent inferences, (b) the learners were able to understand better some of the particular parts in a text, which were closely related to the target inferences, in the second reading with “why” questions than in the first reading without any questions, (c) there was a moderate correlation between understanding some of the particular parts in a text, which were closely related to the target inferences, and making inferences, and (d) with regard to the relationship between L2 reading proficiency and the inferences, there was no correlation on the condition of before “why” questions, whereas there was a moderate correlation on the condition of after “why” questions. The above-mentioned (b) and (d) are particularly interesting findings because the finding of (b) implies that “why” questions have an indirect effect on being able to grasp propositional meaning as well as a direct effect on making inferences, and the finding of (d) indicates the stronger relationship between learners’ L2 reading proficiency and causal inferences generated in the off-line context instead of the on-line context.
Endnotes

1. Part of an earlier version of this chapter has been published in Shimizu (2005).

2. As to the wording of question 6, if there was something to be compared with a think-aloud task such as a silent reading task, the content of the question would be more clearly conveyed to the respondents. However, no student seemed to be troubled with the wording of the currently used question.

3. The five students who marked 5 (absolutely difficult) on question 6 had already been excluded.

4. In Brown (1996), with regard to a ceiling effect, the following statement can be seen: “If all the students have scored so high on a measure that the dispersion is depressed, the related statistics may be impossible to interpret.” (p. 139). The present participants in the upper group scored 15.250 on the before condition, and 16.250 on the after condition on average. If we add 1SD to them each, they become 17.438 (= 15.250 + 2.188) and 18.595 (= 16.250 + 2.345), respectively. Considering that the full scores are 19.000, we can say that the means of “Upper” on the both conditions were very high, which indicates a ceiling effect.
5. Study 3: Effects of Inference Question Sequence on Reading Comprehension

5.1 Purposes and Research Hypotheses

Study 2 confirmed experimentally that there are positive effects of an inference question on EFL learners’ reading comprehension, that is, their inference generation and ability to understand propositional meaning in a particular part of a text. To further research this area, Study 3 was conducted. The first aim of Study 3 was to find out how we could more effectively pose inference questions in reading tasks; to be more specific, to learn in what order we should arrange a set of inference questions so that EFL learners’ comprehension would be most improved. I decided to confine my attention to the presentation sequence of questions. In the following paragraphs, I will refer to Fushimi (1992), the programmed instruction by Skinner (1968), and the “Suido Method” by Toyama and Ginbayashi (1975) as the previous studies regarding the presentation sequence. Although these studies have been developed in academic fields other than English language teaching, I believe that their notions and principles can be applied to the context of English language teaching as well. Another aim of the present study was to investigate the interaction between question sequence and L2 reading proficiency. Before making some hypotheses, the above-mentioned previous studies will be reviewed.

Fushimi’s (1992) research has been conducted in the context of learning some new concepts in L1 such as “metals,” and focused on which sequence would be more preferable in learning, (a) the sequence in which a “well-known case” (“seichi jirei,” in Japanese) is followed by a “misunderstood case” (“gochi jirei,” in Japanese), or (b) the sequence in which a “misunderstood case” is followed by a “well-known case.” A “well-known case” was defined as a case that learners acknowledge correctly being a member of the denotation of the concept, such as “gold” in “metals.” In contrast, a “misunderstood case” was defined
as the case that learners understand *incorrectly* not to be a member of the denotation of the concept, such as “calcium” in “metals.” Also, note that his materials were passages written in Japanese, that is, the participants’ L1, which were comprised of about 600 characters or 2,100 characters. Considering this, although Fushimi’s study was not conducted in the field of English language teaching—but in the field of educational psychology, to be more specific, learning of scientific concepts—the results should be mentioned before discussing the hypotheses of Study 3. He showed that the group given the explanation of the “well-known case” first, and then the “misunderstood case” was able to learn better than the other group. What can be derived from Fushimi’s study is when students learn something unfamiliar, the presentation sequence from easier to more difficult may be more effective than the opposite sequence.

The other previous studies (Skinner, 1968; Toyama & Ginbayashi, 1975) come from general teaching principles or teaching methodology. First, the notion of small frames of instruction can be seen in the programmed instruction developed in the middle of the 20th century by Skinner. This is based on the idea that a set of questions are provided brick by brick by teaching machines or programmed sheets, and the change from a question to the next one should be a short step so that anyone can understand each question. Second, the same idea as Skinner can be seen in the “Suido Method,” which was explored by Toyama and Ginbayashi. The origin is explained in detail in their book entitled *Computational Systems by Suido Method*. A brief summary is as follows. By starting to study computational systems in arithmetic textbooks published in Japan and outside of Japan in the late 1950s, they found that most of the textbooks had either one of the following two courses, “from specific to general” or “from general to specific.” An even more interesting finding was that the most reputable textbook in those days had exhaustively adopted the
former one, “from specific to general.” Toyama and his colleagues devised the exact opposite approach from the reputable textbook, and started to reconstruct computational systems in the order of “from general to specific.” This computational system was later named the “Suido Method,” the word “Suido” evoking the image of a canal system whereby water is naturally diverted to households from water resources. Skinner’s programmed instruction targeted general educational contexts, and Toyama and Ginbayashi targeted arithmetical learning, both of which are different from English language teaching; however, in common with the present study, their interests lie in finding the way for their students to learn effectively.

Moreover, what is also common to these three previous studies is that the sequences are determined from easier items to more difficult items. In more concrete terms, Fushimi (1992) found the sequence of a “well-known case” to a “misunderstood case,” Skinner (1968) proposed small steps with difficulty increasing, and Toyama and Ginbayashi (1975) supported the sequence of “general” to “specific.” This sequence in which an easier question is given first and then a more difficult question next, is likely to be effective for poor readers. Thus, the first hypothesis of the present study was formulated as follows:

Hypothesis 6A. The question sequence in which learners answer inferential questions during the early stages of the sequence, far from thematic questions which are presented at the end of sequence, is more beneficial for learners of lower L2 reading proficiency than the opposite sequence.

This sequence is very general in reading comprehension tasks or reading comprehension tests. This hypothesis is also related to the results of Study 1, which showed that in
comparison with the higher L2 proficiency learners the lower L2 proficiency learners were especially poor at making “local connecting inferences,” that is, dealing with a small part of text in detail. Considering this, I can reasonably predict that giving some guides (questions in this case) helpful to comprehending small parts of text would be more effective for poor readers. Thus, Hypothesis 6A was constructed. Hypothesis 6B, on the other hand, may be a relatively adventurous one.

Hypothesis 6B. The question sequence in which learners answer a thematic question during the early stages and answer inferential questions far from a thematic question during the final stages is more beneficial for learners of higher L2 reading proficiency than the opposite sequence.

This hypothesis includes the following assumption: (a) Good readers must have already understood the answers to easier questions by the time they are given a set of questions; therefore, an obvious effect of the sequence from easier question to more difficult question would not be observed, but (b) if good readers, on the contrary, are given a thematic question first of all, it may positively affect their comprehension of following questions; this is because the thematic question will function as a guiding question which reveals an outline of a passage.

Also, note that the present study regards a question that asks for a main theme of a passage as one type of an inference question, following Graesser and Kreuz (1993), and Graeser et al. (1994; see Table 2.1). However in the following, I call it a thematic question, rather than an inference question.
5.2 Pilot Study 1

The focus of the main study was to investigate whether different question sequences would affect learners’ reading comprehension differently according to their L2 reading proficiency. Texts and questions were to be adopted from the 6th section of the fiscal 2003 test and the fiscal 2004 test of the National Center for University Entrance Examinations (Daigaku Nyushi Center: DNC; see section 5.4.1.2). It was judged that the questions in the fiscal 2003 test were well organized, but that in the fiscal 2004 test there was no question which inquired into the theme of the passage. Therefore, this pilot study was conducted to collect data for making efficient distractors for a multiple-choice question, which would be used in the main study.

Firstly, the question stem was made, which was “what is the main theme of this story?” following the pattern of the fiscal 2003 test. Secondly, the author had a discussion about the answer key and possible distractors with a native speaker of English from New Zealand, who was studying sociology at a graduate school in Japan. We agreed that although “competition,” “swimming,” and “attitude” seemed to be the topic at first glance, “friendship” must be the answer to the main theme of this passage. Thirdly, to confirm our expectations or find new efficient distractors, the following pilot study was conducted thanks to the participation of 40 Japanese senior high school students. They were majoring in math and science and were in the third year at a private senior high school in Japan. They had had English language education at school for at least five years.

One passage from the 6th section in the fiscal 2004 test of DNC was followed by one open-ended question that asked for the main theme of the passage, and one yes-no question which attempted to discover whether a student had read the passage. Indeed, the former question was a focus of interest.
Pilot Study 1 was administered in an English class on January 11, 2005. Although the passage was written in English, the participants answered the question in their L1. The question was also written in their L1. It took about 20 minutes for them to finish the task, including reading the passage and writing down their answers.

Most of the participants, 38 out of 40, wrote answers on the line. All of the answers are listed below (translation mine), with similar answers placed side-by-side for instant recognition.

(a) anata wa hitori ja nai [you are not alone]
(b) anata wa hitori dewa nai [you are not alone]
(c) yujo [friendship]
(d) yujo [friendship]
(e) yujo [friendship]
(f) yujo [friendship]
(g) yujo no subarashisa [How wonderful a friendship is!]
(h) yujo no chikara ya taisetsusa [the power or the importance of a friend]
(i) shinrайдekiru tomodachi to iu koto ni tsuite [in a friend we can trust]
(j) suieibu no Angela to Kate no yujo [a friendship between Angela and Kate in the swimming club]
(k) yujo ni yori jibun no tarinai bubun wo oginau koto wa taisetsu da to iu koto [a friend’s help is important to support our own weaknesses]
(l) futari de yujo to suiei no shori dochira mo totta [each of them found a good friend and won the game]
(m) sports wo toshite umareru yujo [friendship built up through sports]
(n) futari no bishojo no hitotsu no race deno yujo monogatari [a story of a friendship between two beautiful girls’ in a race]
(o) Angela to watashi [Angela and I]
(p) Angela to watashi futari ga etamono [what two of them had]
(q) suiei senshu no yujin [a swimmer’s friend]
(r) suiei wo toshite no yujo [a friendship through swimming]
(s) suiei kara hirogaru yujo no wa [a friendship built up through swimming]
(t) suiei to shinyu no futatsu wo doji ni teni haitta koto [swimming and a friendship]
(u) tomodachi wo soncho shi otagaini tasukeau koto ga taisetsu de aru [it is important to respect friends and help each other]
(v) suiei [swimming]
(w) suiei [swimming]
(x) swimming ni tsuite [about swimming]
(y) oyogu koto wa tanoshi [enjoyment in swimming]
(z) suiei no kessho sen [the final race of swimming]
(aa) suiei no race no hanashi [a story of a swimming race]
(ab) swimming school no taikai [a competition at a swimming school]
(ac) kyoso [competition]
(ad) suiei no rival ishiki [keen sense of rivalry in swimming]
(ac) rival no Angela [Angela as her rival]
(af) suiei de kisoi atta furari [two girls who had competed in swimming]
(ag) tomodachi to issho ni kisou koto wa subarashi [it is wonderful to compete with a friend]
(ah) akiramezuni ganbareba ii [do your best without giving up]
(ai) ganbareba mukuwareru [all your effort will be rewarded]
(aj) kyoryoku suru koto wa daiji da [it is important to cooperate harmoniously]
(ak) dochira ga makete mo futari de ganbare ba sorewa futari no shori ni naru [if both of them direct their efforts they can achieve a victory together in spite of whether either one loses a game]
(al) sonkei shiai kyoryoku shiau to yoi kekka ni naru to iu koto [giving your respects and cooperation will lead to desired results]

About a half of the responses included something related to “friends” or “friendship.”
In addition, there were many answers referring to “swimming,” and a few answers referring to “competition” (or “race”) such as (z), (aa), (ab) and (ac), and “a rival” (or “rivalry”) such as (ad) and (ae). As a result, “swimming,” “competition,” and “a rival” were selected as distractors for a thematic question.
Furthermore, to examine whether these words would not be too difficult for the participants of the following main study, word frequency levels were investigated on the basis of *JACET List of 8000 Basic Words* (JACET Basic Words Revision Committee, 2003). Three words were judged to be within the 3000 frequency levels, however, “swimming” was found to be at the 6000-frequency level. However, “swim,” which is a root form of “swimming,” was found to be in the 2000 word frequency level. Thus, in addition to the other words, “swimming” was adopted for the main study (see Table 5.1).

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>friendship</td>
<td>2000</td>
</tr>
<tr>
<td>swimming</td>
<td>6000</td>
</tr>
<tr>
<td>rival</td>
<td>3000</td>
</tr>
<tr>
<td>competition</td>
<td>2000</td>
</tr>
</tbody>
</table>

*Note. *Answer key.

It was somewhat surprising for the author to see that none of the participants in this pilot study had read the passage adopted from the fiscal 2004 test of DNC before. However, this tendency worked in the author’s favor because of the plan to conduct research using the material. Thus, it was expected that few students, at least in the 10th or the 11th grades, would have read or learned the material.

### 5.3 Pilot Study 2

Another pilot study was conducted to set appropriate times for reading passages and answering questions, and to find out whether all of the instructions would be clear to senior high school students.
Three students in the 12th grade at a public school, who were volunteers and not to be participants in the main study, took the same tests as those planned for the main study (see section 5.4.1.2). Tentatively, 15 minutes were provided for the Society for Testing English Proficiency (STEP) test, and about 31 minutes were provided for the DNC tests. The details of the DNC tests were as follows: 14 minutes for reading (7 minutes for reading each passage), and 16 and a half minutes for answering 11 questions (1 and a half minutes for answering each question).

Participants also completed a questionnaire (see Appendix 5-A) for the revision of procedures and wording used in instructions. All of these procedures were undertaken for approximately 60 minutes in total.

All of the tasks were not completed in 50 minutes, which is the usual amount of class time at senior high schools in Japan. Therefore, some redundant explanations were cut off, and time for taking the STEP test was reduced from 15 minutes to 13 minutes because 15 minutes was judged to be relatively long for the participants in the main study. And, one and a half minutes for each question in the DNC tests seemed appropriate.

With regard to the questionnaire, this was completed by two of the three participants (see Appendix 5-A). Concerning question (a-1), that is, “as to English tests of National Center for University Entrance Examinations, how do you feel about reading time?” the students wrote “I couldn’t finish reading, and I read them too slowly” (translation mine), and “it was short” (translation mine). Furthermore, as an overall impression, a student replied, “they were difficult” (translation mine). The two students judged that there were no problems other than those reported above. Reading time and difficulty levels remained unchanged even after this pilot study was administered. This was because a number of L2 learners of a higher proficiency than the students in this pilot study were to be involved in
the main study.

5.4 Main Study

5.4.1 Method

5.4.1.1 Participants

In all, 827 Japanese learners of EFL participated in this study. They were 10th-, 11th-, and 12th-grade students from four different public senior high schools in Japan. As one index of their academic proficiency levels, brief explanations of the courses their alumnae took are summarized in Table 5.2, which was made by the author with reference to school internet home pages and personal communication with their English teachers. Each school was randomly assigned a letter (A, B, C, and D). Differences in the schools will be once more considered in terms of the validation of the L2 reading proficiency test in section 5.4.2.1. The school differences in academic levels were assumed by the courses taken by their alumni--former students of the schools. The assumption was that if the higher percentage of the alumni went to universities or started to prepare for the entrance examinations the school academic levels were thought to be higher. With regard to School A and School B, almost 100% students go to universities or wish to do so. However, the rates were lower at School C and School D.

Before the administration of the test, the author contacted the teachers who were teaching English to the participants and was told that none of the participants had at school learned the materials which would be used in the main study. Thus, it was expected that these materials would be familiar to only a few students, who had found these by themselves or at their cram school.
Table 5.2

*Schools of the Participants*

<table>
<thead>
<tr>
<th>School</th>
<th>Participant</th>
<th>Course of Their Alumni</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>201 10th-grade students, who had completed English I and the first half of English II (Ministry of Education, Science, Sports and Culture, 1999)</td>
<td>Generally speaking, every year about 80% students enter a university, and about 20% students prepare for the following year’s university entrance examinations. A few students may start working immediately after their graduation. Almost all of the 12th-grade students take the DNC test. [an English teacher, personal communication, February 2, 2005; translation mine]</td>
</tr>
<tr>
<td>B</td>
<td>273 10th-grade students, who had completed English I</td>
<td>Generally speaking, every year about 70.0% students enter a university, and about 28.5% students prepare for the following year’s university entrance examinations. About 1.5% students enter a technical college, or foreign schools. No students start working. [an English teacher, personal communication, February 14, 2005; translation mine]</td>
</tr>
<tr>
<td>C</td>
<td>223 11th-grade students</td>
<td>According to data in for the fiscal 2003, 53% of students entered a university or a junior college, and 33% entered a technical college. 12% of students prepared for the following year’s entrance examinations, and 2% of students started working immediately after their graduation. [summary of school home page by the author; translation mine]</td>
</tr>
<tr>
<td>D</td>
<td>130 11th- and 12th-grade students</td>
<td>According to data for the fiscal 2004, 81% of students entered a university or a junior college, and 16% of students entered a technical college. 3% of students started working immediately after their graduation. [summary of school home page by the author; translation mine]</td>
</tr>
</tbody>
</table>

Furthermore, it should be noted that some tables in the following sections show the data of fewer than 827 students. The first reason for this was that two students who submitted DNC test answer sheets without any marks were from the beginning not included for analysis. The second reason was that the data of the students who had read the same passages as was used in this study were excluded from analysis.
5.4.1.2 Materials

Two types of reading tests and one questionnaire were used in this study. The reading test numbered 1, given below, was administered to examine the effects of question sequences on learners’ reading comprehension whereas the reading test numbered 2 was employed to measure the participants’ L2 reading proficiency. A questionnaire (see Appendix 5-D) was prepared to confirm that participants had not previously read the same passages. This was because only the data of the students who had not previously read the passages should be analyzed.

1. Part A of the 6th section in the fiscal 2003 and 2004 DNC English tests: 11 multiple choice questions with 4 choices
2. “3A” and “4B” from the STEP 2nd Grade Test 2 in 1997: 5 gap-filling questions and 5 question-and-answer questions, which had 4 choices
3. Questionnaire

Let us now look more closely at the two reading passages from the DNC tests.

Table 5.3

<table>
<thead>
<tr>
<th></th>
<th>Fiscal 2003 Test</th>
<th>Fiscal 2004 Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>665</td>
<td>693</td>
</tr>
<tr>
<td>Flesch Reading Ease*</td>
<td>81.7</td>
<td>83.8</td>
</tr>
<tr>
<td>Flesch-Kincaid Grade Level*</td>
<td>5.3</td>
<td>5.0</td>
</tr>
</tbody>
</table>

*Note. These indices were provided by Microsoft Word 2003’s readability measurement tools.*
Both the fiscal 2003 and 2004 DNC English tests assessed examinees’ narrative reading (see Appendix 5-B). Table 5.3 indicates the word numbers and readability of each passage. These two texts were almost identical according to these indices.

With regard to the reading comprehension questions, five questions had already been attached to the passages, respectively. The questions for the fiscal 2003 test were roughly organized from a question regarding detailed information to one seeking the main topic of the passage, as shown in the followings: (a) The first three questions--“According to Elena’s parents, why would it be good for Elena to move to Japan?” (item 1), “What problems did Elena have after she entered junior high school?” (item 2), and “Why did Elena feel hurt when she was asked about going back to Brazil?” (item 3)--were found to be inference questions which would elicit from readers something like characters’ feelings or the reasons why a certain event occurred in a text, (b) the fourth question--“What did Natsumi mean when she said Elena is ‘different’?”--was an inference question as well, but was more closely related to the theme of the passage than the previous three questions, and (c) the last question--“What is the main theme of this story?”--was a thematic question, which asked examinees the main theme of the entire passage.

In contrast, the fiscal 2004 DNC English test did not have a thematic question; all of the five questions were inference questions, and only the final question slightly overlapped with the main theme--“How did Kate feel after Angela joined the swimming club?” (item 1), “Why did the coach talk to Kate before the trial races?” (item 2), “Why did Kate want to help Angela?” (item 3), “Why was Kate disappointed immediately after the final race?” (item 4), and “What did Angela mean when she said, ‘You’re not the only one’?” (item 5). As there was no exact question that asked for a main theme, the following thematic question was prepared (see section 5.2): “What is the main theme of this story?” 1. swimming, 2.
friendship, 3. a rival, 4. competition. Thus, the participants were required to answer a total of 11 questions from the two reading texts.

With regard to these DNC tests, two booklets were distributed to each participant. One booklet (Booklet A) included two reading passages, and the other (Booklet B) included 11 multiple-choice questions for the passages. The participants were instructed to choose one choice among four options for each question, and mark the number in Booklet B directly. In other words, a sheet only for marks was not given. In Booklet B, one page was spared for one question, and the participants were not allowed to go back to the earlier questions. The time to start and finish reading a passage, and to start and finish answering each question, was signaled by the administrators who had a stopwatch (for more information, see section 5.4.1.3).

Table 5.4

Readability of Passages in the English Tests of the STEP Tests

<table>
<thead>
<tr>
<th></th>
<th>Passage “3A”</th>
<th>Passage “4B”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>216</td>
<td>288</td>
</tr>
<tr>
<td>Flesch Reading Ease*</td>
<td>41.5</td>
<td>42.8</td>
</tr>
<tr>
<td>Flesch-Kincaid Grade Level*</td>
<td>12.0</td>
<td>12.0</td>
</tr>
</tbody>
</table>

Note. These indices were provided by Microsoft Word 2003’s readability measurement tools. Titles were eliminated at the time of measuring readability. When readability was measured, blanks for a gap-filling test were filled with correct words.

The STEP 2nd grade test was also considered to be appropriate for measurement of the participants’ L2 reading proficiency. It comprised of two expository passages titled “Home Schooling in America” and “A New Kind of Tax,” respectively. It contained five gap-filling items and five question-and-answer items. The latest version of the STEP test was not used.
in this study to avoid passages familiar to many participants. In addition, no considerable
difference between the old version and the latest version of the reading part of the STEP test
could be found (see Appendix 5-C).

5.4.1.3 Data Collection Procedures

The participants were randomly grouped into Group A or Group B. Then, according to the group, each participant received two booklets, in one of which questions were ordered as shown in Table 5.5. All of the participants answered the fiscal 2003 test first, and after that, the fiscal 2004 test. As well as a series of reading questions, information regarding key paragraphs was written on a question sheet (see Appendix 5-B). They were to reduce the advantage of Group A in answering questions in the order of story plots.

Table 5.5

*The Order of Presentation of Questions*

<table>
<thead>
<tr>
<th></th>
<th>Group A (n = 403)</th>
<th>Group B (n = 391)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fiscal 2003 Test</td>
<td>Items 1 → 2 → 3 → 4 → 5</td>
<td>Items 5 → 4 → 3 → 2 → 1</td>
</tr>
<tr>
<td>Fiscal 2004 Test</td>
<td>Items 1 → 2 → 3 → 4 → 5 → 6</td>
<td>Items 6 → 5 → 4 → 3 → 2 → 1</td>
</tr>
</tbody>
</table>

The participants were given 13 minutes for the STEP tests, and 30 and a half minutes for the DNC tests. For the DNC tests, the time for reading each passage (7 minutes each) and for answering each question (1 and a half minutes each) was measured by test administrators, and the participants were obliged to follow their instructions (see Appendix 5-E). After completing a reading test, a questionnaire was given.

These were administered in classes supervised by their English teachers. It was confirmed through their observation that 1 and a half minutes, allotted to each question,
were enough for the participants to turn over a page and respond to a question. This research was conducted from January 27 to February 10, 2005.

5.4.1.4 Data Analyses

A two-way ANOVA was performed to assess the significance of the observed difference in means between the three different L2 reading proficiency groups (i.e., “Lower,” “Middle,” and “Upper”) and the different two groups in question sequence (i.e., Group A and Group B), and the interaction between them. The ANOVA was conducted with the scores of an interval scale, whereas the $2 \times 2$ Fisher’s exact tests were conducted item by item in the analysis of categorical data.

As well as this, hierarchical cluster analyses were applied to investigate the effects of question order on scores. According to SPSS (1999), “…clustering begins by finding the closest pair of objects (cases or variables) according to a distant measure and combines them to form a cluster. The algorithm continues one step at a time, joining pairs of objects, pairs of clusters, or an object with a cluster, until all of the data are in one cluster” (p. 293). The present study used the Ward method, which is designed to optimize the minimum variance within clusters using the within groups sum of squares, and the Squared Euclidean distance to measure intervals. Dendrograms were examined making the following assumptions: (a) If there was no effect of question sequence on readers’ processes the dendrograms of Group A and Group B would match perfectly because both groups had been given the same questions, and (b) if there, on the contrary, was an effect of question sequence on readers’ processes the dendrograms of Group A and Group B would not match.

The value of Cronbach’s alpha was computed and a one-way ANOVA was run for test validity. Moreover, it was decided that if low Cronbach’s alpha or point biserial correlation
coefficients were obtained as to the materials, it would be reasonable to make as many alterations as necessary in order to improve them, or even to avoid proceeding to the next step, viz., statistical analyses.

The statistical software SPSS package for Windows (Version 10.0E, 11.0J and 13.0E) and Java Script-STAR (Version 3.6.4J and Version 3.6.6J; Tanaka, n.d.) were used for data entry and analysis. SPSS 11.0J and 13.0E were run for the Brown Forsythe modifications of ANOVA, which would be preferable to the standard ANOVA when the homogeneity of variance assumption was not tenable (Glass & Hopkins, 1996, pp. 405-406).

5.4.2 Results

5.4.2.1 Validity and Descriptive Statistics

5.4.2.1.1 Society for Testing English Proficiency Test

In this section, I will report the results of examinations of validity of interpretations and uses based on the STEP test adopted to measure the participants’ L2 reading proficiency. To begin with, let us discuss the construct validity of the STEP test from the structural aspect and the external aspect (Messick, 1989, 1996; see section 3.3.1). First, the validation from the structural aspect concerns the internal consistency of the test, and the present study examined it by computing Cronbach’s alpha (for more information, see section 3.3.1). Second, the validation from the external aspect, in contrast, concerns “relationships of test scores with other tests and behaviors” (Chapelle, 1999, p. 262). Then, the academic levels of the participants’ schools were considered. Students from four high schools in Japan participated in this research, and the school academic levels were inferred from the information about the percentages of their alumni who entered university immediately or a few years after graduation. The structural aspect and the external aspect for the test
validation were selected by considering the following issues: (a) whether the data for an analysis were available, and (b) whether the procedures were not so complex because the test validation was not the main focus of the present study. The results are reported in the following paragraphs.

Cronbach’s alpha coefficient of the test, which contained 10 items in total, was .414, that is, a moderate correlation coefficient. Because Cronbach’s alpha coefficient largely relies on the number of items and item discrimination, the reasons why the test produced only a moderate reliability were examined from these aspects.

Table 5.6

<table>
<thead>
<tr>
<th></th>
<th>Passage “3A”</th>
<th></th>
<th>Passage “4B”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item Number</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>(r_{pb})</td>
<td>.164</td>
<td>.201</td>
<td>.113</td>
</tr>
</tbody>
</table>

*Note. \(r_{pb}\) = point biserial correlation coefficients.*

Firstly, regarding the relationship between reliability and test length, there is the Spearman-Brown formula, which predicts the necessary numbers of items to attain a certain level of reliability (Henning, 1987). On this occasion, I would ideally have a test 3.3 times as long as the present one in order to achieve reliability .700 and greater. Spending about 43 minutes (= 13 \(\times\) 3.3) to assess participants’ L2 reading proficiency was preferable, but unfortunately, it was impractical on the present occasion, in which tests were administered to a large number of participants in class. Secondly, point biserial correlation coefficients of the test were computed. As is shown in Table 5.6, they ranged from .164 to .264. The items in 4B showed especially lower coefficients.

Next, the external aspect of construct validity was examined. The focus was whether...
the scores of the STEP test agreed with school differences (see section 5.4.1.1). In other words, it was expected that the schools at which more alumnae entered universities or prepared for the following year’s university entrance examinations every year would attain higher scores in this test. In the present case, the average scores of School A and School B, at which almost 100% of students fell into this category (see Table 5.2), were predicted to be higher than those of School C and School D. With regard to School C and School D, around 60% and 80% of the students go to universities, respectively. Means and standard deviations of these four schools are shown in Table 5.7. The highest mean was that of School A, 3.615 (SD = 1.872), whereas the lowest one was that of School C, 2.667 (SD = 1.384). The difference was almost 1 point out of 10 points.

Table 5.7

Scores of the STEP Test by School

<table>
<thead>
<tr>
<th>School</th>
<th>n</th>
<th>Mean</th>
<th>SD</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>195</td>
<td>3.615</td>
<td>1.872</td>
<td>3.351</td>
<td>3.880</td>
<td>0.000</td>
<td>10.000</td>
</tr>
<tr>
<td>B</td>
<td>272</td>
<td>3.353</td>
<td>1.906</td>
<td>3.125</td>
<td>3.580</td>
<td>0.000</td>
<td>10.000</td>
</tr>
<tr>
<td>C</td>
<td>216</td>
<td>2.667</td>
<td>1.384</td>
<td>2.481</td>
<td>2.852</td>
<td>0.000</td>
<td>8.000</td>
</tr>
<tr>
<td>D</td>
<td>111</td>
<td>2.802</td>
<td>1.656</td>
<td>2.490</td>
<td>3.113</td>
<td>0.000</td>
<td>7.000</td>
</tr>
<tr>
<td>Total</td>
<td>794</td>
<td>3.154</td>
<td>1.774</td>
<td>3.030</td>
<td>3.277</td>
<td>0.000</td>
<td>10.000</td>
</tr>
</tbody>
</table>

*Note.* Full score = 10.000.

Whether differences existed among the four schools’ average scores was determined by using a one-way ANOVA. Because of unequal variances in the population, the Brown-Forsythe statistic (1974a, as cited in Glass & Hopkins, 1996, p. 406) was employed. The results showed the significant difference at the .001 level ($F^* (3, 661.348) = 13.356$). Furthermore, the Tukey’s HSD procedure was employed to compare all of the possible
pairwise combinations of the four means, which revealed significant differences between the following pairs at the .05 level: (a) School A > School C, (b) School A > School D, (c) School B > School C, and (d) School B > School D. Although the effect size for the variable was small ($\eta^2 = 0.047$), these results met my expectations. Please consider the confidence intervals as well, which indicated the same results. The 95% confidence intervals for the means of School A and School B were higher than those of School C and School D. The intervals of School A and School B did not overlap with School C and School D. Thus, it is reasonable to conclude that it was shown that the test validity from the external aspect was high.

To summarize the examination of the validity, we can say that positive evidence was shown from the external aspect, whereas relatively negative evidence was shown from the structural aspect. Therefore, to return to the discussion of the structural aspect, it seemed reasonable to drop some questions with low point biserial correlations from the test. In order to not cut too many items, $r_{pb} = .200$ was set as a limit; and four items were left. After six items were dropped, the reliability of the test declined slightly from .414 to .382 (Cronbach’s $\alpha$), but the average point biserial correlation coefficient greatly increased from .167 to .223. Item discrimination was especially crucial in this case because the purpose of this test was to divide the learners into different L2 reading proficiency groups according to the scores; therefore, to leave four items while excluding six items was considered to be reasonable.

Let us now move to the descriptive statistics. Means and standard deviations are shown in Table 5.8. The means for Group A and Group B were 0.856 ($SD = 0.935$) and 0.844 ($SD = 0.960$), respectively. A $t$-test comparing the means confirmed that there was no significant difference between the mean scores of Group A and Group B ($t (792) = 0.180, p > .05$).
Furthermore, no effect was found by effect size indexes as well, $g = 0.013$. These indicated the homogeneity in L2 reading proficiency of these groups.

Table 5.8

<table>
<thead>
<tr>
<th></th>
<th>Group A ($n = 403$)</th>
<th>Group B ($n = 391$)</th>
<th>df</th>
<th>$t$</th>
<th>$g$</th>
</tr>
</thead>
<tbody>
<tr>
<td>STEP Test</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>df</td>
</tr>
<tr>
<td></td>
<td>0.856</td>
<td>0.935</td>
<td>0.844</td>
<td>0.960</td>
<td>792</td>
</tr>
</tbody>
</table>

*Note.* Full score = 4.000.

$g =$ effect size. $0.200 < g ($small) < 0.500; 0.500 \leq g$ (medium) < 0.800; $0.800 \leq g$ (large; Cohen, 1988). Effect size of 0.199 and below is regarded as having no effect.

Thus, on the basis of their scores of the STEP test after six items with lower point biserial correlation coefficients were excluded, the participants were allotted to either one of the three groups, upper group (“Upper”), middle group (“Middle”), or lower group (“Lower”). Those who received 2 to 4 points were grouped into “Upper,” those who received 1 point were grouped into “Middle,” and those who received 0 points were grouped into “Lower.” “Upper” was comprised of 152 participants (19%), “Middle” was comprised of 302 (38%), and “Lower” was comprised of 340 (43%).

A one-way ANOVA and Tukey’s HSD post-hoc test revealed that there was statistical difference among participants in the average scores of the L2 reading proficiency test ($F(2, 794) = 3746.256, p < .001$). Moreover, the effect size measured by $\eta^2$ was large, 0.501. Therefore, it was confirmed that the “Upper” learners’ L2 reading proficiency was higher than that of the “Middle” learners, and that the “Middle” learners’ L2 reading proficiency was higher than that of the “Lower” learners.
Table 5.9

Means and SDs of STEP Test by Group(1)

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Middle</td>
<td>1.000</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Upper</td>
<td>2.447</td>
<td>0.661</td>
<td>2.461</td>
<td>0.682</td>
<td>2.454</td>
<td>0.670</td>
</tr>
</tbody>
</table>

Note. Full score = 4.000.

Preliminarily considering some critical arguments such as that the L2 reading proficiency of “Lower” was not able to be measured accurately because of “a floor effect” of the test, descriptive statistics of the original 10 items and results of a two-way ANOVA were computed for more information.

Table 5.10

Means and SDs of STEP Test by Group(2)

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower</td>
<td>2.065</td>
<td>0.103</td>
<td>2.047</td>
<td>0.102</td>
<td>2.056</td>
<td>1.197</td>
</tr>
<tr>
<td>Middle</td>
<td>3.377</td>
<td>0.106</td>
<td>3.315</td>
<td>0.112</td>
<td>3.348</td>
<td>1.182</td>
</tr>
<tr>
<td>Upper</td>
<td>5.013</td>
<td>0.154</td>
<td>5.434</td>
<td>0.156</td>
<td>5.224</td>
<td>1.849</td>
</tr>
</tbody>
</table>

Note. Full score = 10.000.

Means and standard deviations are reported in Table 5.10. The ANOVA revealed the main effect for L2 reading proficiency \( (F (2, 794) = 298.350, p < .000) \); but, neither the main effect for the question sequence group (i.e., Group A or Group B; \( F (1, 794) = 1.248, \text{n.s.} \)) nor the interaction effect \( (F (2, 794) = 1.834, \text{n.s.}) \) was significant. The differences in means between all of the possible pairs among “Lower,” “Middle,” and “Upper” were determined at the .001 level by the Tukey’s HSD post-hoc test. Thus, it was concluded that the
participants’ L2 reading proficiency levels determined by the four items were corroborated by the data from 10 items.

5.4.2.1.2 Tests of the National Center for University Entrance Examinations

I will start with validation of the DNC test. The Cronbach’s alpha, .557 was obtained from the test of 11 items in total. Next, alphas were computed by passage, which showed that \( \alpha = .423 \) was given for the fiscal 2003 test with 5 items, and that \( \alpha = .386 \) was given for the fiscal 2004 test with 6 items. Both of them were unexpectedly low. Therefore, it was concluded that, in the present context, it would not be preferable to separate total scores into the two scores for the fiscal 2003 test and for the fiscal 2004 test. However, it is important to examine the effect of question sequence passage by passage without combining two passages; therefore, it is hoped that this issue will be examined in future research.

Again, let us examine the reliability, \( \alpha = .557 \), for a total of 11 items, from the viewpoints of the number of items and item discrimination. The Spearman-Brown formula (Henning, 1987) predicts that a test with 1.9 times as many items as the present version would attain .700 and greater of reliability.

Table 5.11

*Point Biserial Correlation Coefficients of the DNC Test Used in Study 3*

<table>
<thead>
<tr>
<th>Item Number</th>
<th>Passage 1</th>
<th></th>
<th>Passage 2</th>
<th></th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>( r_{pb} )</td>
<td>.296</td>
<td>.394</td>
<td>.275</td>
<td>.022</td>
<td>.263</td>
</tr>
</tbody>
</table>

*Note. N = 794.*

\( r_{pb} \) = point biserial correlation coefficients.

With regard to item discrimination, the point biserial correlation coefficients ranged
between .022 (item 4 in Passage 1) and .406 (item 3 in Passage 2). With reference to the Henning’s (p. 53) criteria, six items among 11 could effectively differentiate between good readers and poor readers. The coefficients of the DNC tests were generally higher than those of the STEP test (see Tables 5.6 and 5.11), even though the test lengths should be considered.

The point biserial correlation coefficients of the present study, $r_{pb} = .238$ on average, were relatively lower than those of Sugino et al. (2003), who reported $r_{pb} = .312$ on average in the 6th section in the 2001 DNC test. Although Sugino et al.’s contained 5 more items which were in Part A but not the present study, it would be possible to compare these statistics as a rough indication because the same part of the DNC test were targeted. In addition, Sugino et al. report one more interesting finding with regard to wild guessing. To investigate the effects of wild guessing on item reliability, they replaced their participants’ blank responses with random marks. The result showed that the point biserial correlation coefficients through that operation were lowered from .312 to .291 (p. 227). Taking Sugino et al.’s study into consideration, the participants in the present study may have conducted random marking more often than usual due to limited time for each answer, which may have lead to lower reliability.

Another aspect of construct validity of Messick’s (1989, 1996) framework will be considered, as was discussed in section 5.4.2.1.1, viz., considering the academic levels of the four schools. Average scores of the DNC test are shown in Table 5.12 by school. A one-way ANOVA was calculated across the four means. The $F$ value was significant at the .001 level, suggesting that the students from different schools performed differently on the DNC test ($F (3, 793) = 53.032, p < .001$); in addition, a large effect size was reported, $\eta^2 = 0.168$. A post hoc comparison of group means, using the Tukey’s HSD test, revealed differences at a significance level of .05 between School A and School B, School A and School C, School A
and School D, School B and School C, and School B and School D. In other words, the following descending order was found: School A > School B > School C = School D.

Table 5.12

 Means and SDs of DNC Test by School

<table>
<thead>
<tr>
<th>School</th>
<th>n</th>
<th>Mean</th>
<th>SD</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>195</td>
<td>6.764</td>
<td>2.052</td>
<td>6.474</td>
<td>7.054</td>
<td>1.000</td>
<td>11.000</td>
</tr>
<tr>
<td>B</td>
<td>272</td>
<td>6.040</td>
<td>2.158</td>
<td>5.783</td>
<td>6.298</td>
<td>1.000</td>
<td>11.000</td>
</tr>
<tr>
<td>C</td>
<td>216</td>
<td>4.606</td>
<td>1.882</td>
<td>4.354</td>
<td>4.859</td>
<td>0.000</td>
<td>9.000</td>
</tr>
<tr>
<td>D</td>
<td>111</td>
<td>4.441</td>
<td>2.198</td>
<td>4.028</td>
<td>4.855</td>
<td>0.000</td>
<td>10.000</td>
</tr>
<tr>
<td>Total</td>
<td>794</td>
<td>5.605</td>
<td>2.260</td>
<td>5.447</td>
<td>5.762</td>
<td>0.000</td>
<td>11.000</td>
</tr>
</tbody>
</table>

Note. Full score = 11.000.

This was also supported by the 95% confidence interval for means (see Table 5.12). According to the information regarding the courses of the alumnae of each school, it was expected that School A and School B would exceed School C and School D in means. Except for the significant difference between School A and School B, the data matched my expectation. To summarize the validation of the DNC test, the satisfactory level of validity seemed to be achieved from the external aspect, but not from the structural aspect.

A total of 794 participants’ descriptive statistics will be shown first, and a description of means and standard deviations by L2 reading proficiency group will follow. The mean scores were 2.525 (SD = 1.302) on the fiscal 2003 test, and 3.082 (SD = 1.438) on the fiscal 2004 test, respectively. When we did not differentiate the years of these tests, the participants obtained 5.605 out of 11.000 on average. The means of Group A and Group B were 5.792 (SD = 2.178) and 5.412 (SD = 2.330), respectively. Thus, considering that the average percentage score was around 50%, the test facility was judged to be of moderate
difficulty for the present participants.

Table 5.13

*Means and SDs of DNC Tests*

<table>
<thead>
<tr>
<th>DNC Test (Full Score)</th>
<th>Group A</th>
<th></th>
<th>Group B</th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n M SD</td>
<td>n M SD</td>
<td>N M SD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fiscal 2003 Test (5.000)</td>
<td>410 2.661 1.272</td>
<td>401 2.387 1.318</td>
<td>811 2.525 1.302</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fiscal 2004 Test (6.000)</td>
<td>409 3.132 1.405</td>
<td>393 3.031 1.472</td>
<td>802 3.082 1.438</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total (11.000)</td>
<td>403 5.792 1.782</td>
<td>391 5.412 2.330</td>
<td>794 5.605 2.260</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* 1. The number of participants in Group A who took both tests.
2. The number of participants in Group B who took both tests.

To examine these scores more precisely, the average scores and standard deviations were computed by item (see Table 5.14).

Table 5.14

*Means and SDs by Sequence Group and Item*

<table>
<thead>
<tr>
<th></th>
<th>Passage 1</th>
<th></th>
<th>Passage 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5</td>
<td></td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>Group A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>410 410 410 410 410</td>
<td></td>
<td>409 409 409 409 409</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>.661 .412 .646 .268 .673</td>
<td></td>
<td>.560 .518 .680 .296 .411 .667</td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>0.474 0.493 0.479 0.444 0.470</td>
<td></td>
<td>0.497 0.500 0.467 0.457 0.493 0.472</td>
<td></td>
</tr>
<tr>
<td>Group B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>401 401 401 401 401</td>
<td></td>
<td>393 393 393 393 393</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>.738 .379 .569 .229 .471</td>
<td></td>
<td>.560 .440 .650 .310 .440 .630</td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>0.440 0.486 0.496 0.421 0.500</td>
<td></td>
<td>0.500 0.500 0.480 0.460 0.500 0.480</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>811 811 811 811 811</td>
<td></td>
<td>802 802 802 802 802</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>.699 .396 .608 .249 .573</td>
<td></td>
<td>.560 .480 .666 .303 .423 .651</td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>0.459 0.489 0.489 0.433 0.495</td>
<td></td>
<td>0.497 0.500 0.472 0.460 0.494 0.477</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Full score = 1.000.
Since the test consisted of typical multiple-choice discrete-point items, either zero or a perfect answer was given to an item. The facility values varied considerably in accordance with items, ranging from .249 (item 4 of Passage 1) to .699 (item 1 of Passage 1). The results of Group A and Group B are also shown in the table. Again, you should note that Group B took a set of questions in the opposite order of Group A; that is, Group A answered from item 1 to item 5 (or to item 6 in the case of Passage 2), whereas Group B answered from item 5 (or from item 6 in the case of Passage 2) to item 1. In order to make a single table containing the results of both groups, Group B’s order was converted tentatively. The same column indicates the mean score and standard deviation for that question.

Table 5.15 shows descriptive statistics from another viewpoint.

Table 5.15

<table>
<thead>
<tr>
<th></th>
<th>Passage 1</th>
<th>Passage 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td><strong>Group</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>A</strong> Correct</td>
<td>266 166 263 108 272</td>
<td>226 206 274 119 166</td>
</tr>
<tr>
<td>Incorrect</td>
<td>137 237 140 295 131</td>
<td>177 197 129 284 237</td>
</tr>
<tr>
<td><strong>B</strong> Correct</td>
<td>288 146 223 90 185</td>
<td>220 172 255 121 169</td>
</tr>
<tr>
<td>Incorrect</td>
<td>103 245 168 301 206</td>
<td>171 219 136 270 222</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>554 312 486 198 457</td>
<td>446 378 529 240 335</td>
</tr>
<tr>
<td>Correct</td>
<td>240 482 308 596 337</td>
<td>348 416 265 554 459</td>
</tr>
<tr>
<td>Incorrect</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The number of participants who chose a correct answer and those who chose an incorrect answer were counted by group and item. To be more specific, the following numbers were counted by item: (a) the number of persons who chose a *correct* answer in Group A, (b) the number of persons who chose an *incorrect* answer in Group A, (c) the number of persons who chose a *correct* answer in Group B, and (d) the number of persons who chose an
incorrect answer in Group B. The last two lines of the table were for all counts. As was mentioned above, there were large differences among items, such as 198 correct answers and 596 incorrect answers to item 4 of Passage 1, and 554 correct answers and 240 incorrect answers to item 1 of Passage 1.

The descriptive statistics by L2 reading proficiency group will be covered in the rest of this section. The results can be summarized as follows (see Table 5.16): (a) “Lower” in Group A obtained 5.536 ($SD = 1.876$), while “Lower” in Group B obtained 4.826 ($SD = 2.159$), (b) “Middle” in Group A obtained 5.824 ($SD = 2.263$), while “Middle” in Group B obtained 5.378 ($SD = 2.229$), and (c) “Upper” in Group A obtained 6.289 ($SD = 2.529$), while “Upper” in Group B obtained 6.803 ($SD = 2.332$). Looking at these statistics without considering the statistical significance, we notice that “Upper” received the highest score, “Middle” received the second highest score, and “Lower” received the lowest score. The order was the exactly same as that of the students’ L2 reading proficiency levels measured by the STEP test.

Table 5.16

<table>
<thead>
<tr>
<th>Group</th>
<th>Group A</th>
<th></th>
<th></th>
<th>Group B</th>
<th></th>
<th></th>
<th>Total</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>Mean</td>
<td>SD</td>
<td>n</td>
<td>Mean</td>
<td>SD</td>
<td>n</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Lower</td>
<td>168</td>
<td>5.536</td>
<td>1.876</td>
<td>172</td>
<td>4.826</td>
<td>2.159</td>
<td>340</td>
<td>5.176</td>
<td>2.052</td>
</tr>
<tr>
<td>Middle</td>
<td>159</td>
<td>5.824</td>
<td>2.263</td>
<td>143</td>
<td>5.378</td>
<td>2.229</td>
<td>302</td>
<td>5.613</td>
<td>2.254</td>
</tr>
<tr>
<td>Upper</td>
<td>76</td>
<td>6.289</td>
<td>2.529</td>
<td>76</td>
<td>6.803</td>
<td>2.332</td>
<td>152</td>
<td>6.546</td>
<td>2.438</td>
</tr>
</tbody>
</table>

Note. Full score = 11.000.

Table 5.17 shows means and standard deviations, and Table 5.18 shows counts by L2 reading proficiency group, sequence group and item. The former table corresponds with
Table 5.14, and the latter table corresponds with Table 5.15. The data presented in Table 5.18 will be used to examine the degree of association between question sequence and question type by utilizing the chi-square test (the results will be reported in the next section).

Table 5.17

Means and SDs by L2 Reading Proficiency Group, Sequence Group, and Item

<table>
<thead>
<tr>
<th>Group</th>
<th>Passage 1</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Passage 2</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Lower</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>172</td>
<td>172</td>
<td>172</td>
<td>172</td>
<td>172</td>
<td>171</td>
<td>171</td>
<td>171</td>
<td>171</td>
<td>171</td>
</tr>
<tr>
<td>Mean</td>
<td>.674</td>
<td>.366</td>
<td>.628</td>
<td>.256</td>
<td>.634</td>
<td>.556</td>
<td>.462</td>
<td>.673</td>
<td>.222</td>
<td>.368</td>
</tr>
<tr>
<td>SD</td>
<td>0.470</td>
<td>0.483</td>
<td>0.485</td>
<td>0.438</td>
<td>0.483</td>
<td>0.498</td>
<td>0.500</td>
<td>0.471</td>
<td>0.417</td>
<td>0.484</td>
</tr>
<tr>
<td>Group B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>172</td>
<td>172</td>
<td>172</td>
<td>172</td>
<td>172</td>
<td>172</td>
<td>172</td>
<td>172</td>
<td>172</td>
<td>172</td>
</tr>
<tr>
<td>Mean</td>
<td>.674</td>
<td>.262</td>
<td>.483</td>
<td>.256</td>
<td>.413</td>
<td>.483</td>
<td>.413</td>
<td>.558</td>
<td>.267</td>
<td>.407</td>
</tr>
<tr>
<td>SD</td>
<td>0.470</td>
<td>0.441</td>
<td>0.501</td>
<td>0.438</td>
<td>0.494</td>
<td>0.501</td>
<td>0.494</td>
<td>0.498</td>
<td>0.444</td>
<td>0.493</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>344</td>
<td>344</td>
<td>344</td>
<td>344</td>
<td>344</td>
<td>343</td>
<td>343</td>
<td>343</td>
<td>343</td>
<td>343</td>
</tr>
<tr>
<td>Mean</td>
<td>.674</td>
<td>.314</td>
<td>.555</td>
<td>.256</td>
<td>.523</td>
<td>.519</td>
<td>.437</td>
<td>.615</td>
<td>.245</td>
<td>.388</td>
</tr>
<tr>
<td>SD</td>
<td>0.469</td>
<td>0.465</td>
<td>0.498</td>
<td>0.437</td>
<td>0.500</td>
<td>0.500</td>
<td>0.497</td>
<td>0.487</td>
<td>0.431</td>
<td>0.488</td>
</tr>
<tr>
<td>Middle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>161</td>
<td>161</td>
<td>161</td>
<td>161</td>
<td>161</td>
<td>162</td>
<td>162</td>
<td>162</td>
<td>162</td>
<td>162</td>
</tr>
<tr>
<td>Mean</td>
<td>.646</td>
<td>.410</td>
<td>.646</td>
<td>.242</td>
<td>.708</td>
<td>.512</td>
<td>.593</td>
<td>.673</td>
<td>.321</td>
<td>.438</td>
</tr>
<tr>
<td>SD</td>
<td>0.480</td>
<td>0.493</td>
<td>0.480</td>
<td>0.430</td>
<td>0.456</td>
<td>0.501</td>
<td>0.493</td>
<td>0.471</td>
<td>0.468</td>
<td>0.498</td>
</tr>
<tr>
<td>Group B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>148</td>
<td>148</td>
<td>148</td>
<td>148</td>
<td>148</td>
<td>144</td>
<td>144</td>
<td>144</td>
<td>144</td>
<td>144</td>
</tr>
<tr>
<td>SD</td>
<td>0.426</td>
<td>0.483</td>
<td>0.499</td>
<td>0.393</td>
<td>0.499</td>
<td>0.498</td>
<td>0.499</td>
<td>0.471</td>
<td>0.453</td>
<td>0.495</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>309</td>
<td>309</td>
<td>309</td>
<td>309</td>
<td>309</td>
<td>306</td>
<td>306</td>
<td>306</td>
<td>306</td>
<td>306</td>
</tr>
<tr>
<td>Mean</td>
<td>.702</td>
<td>.388</td>
<td>.599</td>
<td>.217</td>
<td>.586</td>
<td>.536</td>
<td>.526</td>
<td>.673</td>
<td>.304</td>
<td>.428</td>
</tr>
<tr>
<td>SD</td>
<td>0.458</td>
<td>0.488</td>
<td>0.491</td>
<td>0.413</td>
<td>0.493</td>
<td>0.500</td>
<td>0.500</td>
<td>0.470</td>
<td>0.461</td>
<td>0.496</td>
</tr>
<tr>
<td>Upper</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>77</td>
<td>77</td>
<td>77</td>
<td>77</td>
<td>77</td>
<td>76</td>
<td>76</td>
<td>76</td>
<td>76</td>
<td>76</td>
</tr>
<tr>
<td>Mean</td>
<td>.662</td>
<td>.519</td>
<td>.688</td>
<td>.351</td>
<td>.688</td>
<td>.671</td>
<td>.487</td>
<td>.711</td>
<td>.408</td>
<td>.447</td>
</tr>
<tr>
<td>SD</td>
<td>0.476</td>
<td>0.503</td>
<td>0.466</td>
<td>0.480</td>
<td>0.466</td>
<td>0.473</td>
<td>0.503</td>
<td>0.457</td>
<td>0.495</td>
<td>0.501</td>
</tr>
</tbody>
</table>
Table 5.18

Numbers of Students who Chose a Correct Answer and an Incorrect Answer, by L2 Reading Proficiency Group, Sequence Group, and Item

<table>
<thead>
<tr>
<th>Group</th>
<th>Passage 1</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Passage 2</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Group A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lower</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group B</td>
<td>n</td>
<td>81</td>
<td>81</td>
<td>81</td>
<td>81</td>
<td>81</td>
<td>77</td>
<td>77</td>
<td>77</td>
<td>77</td>
<td>77</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>.827</td>
<td>.654</td>
<td>.790</td>
<td>.247</td>
<td>.630</td>
<td>.727</td>
<td>.481</td>
<td>.818</td>
<td>.455</td>
<td>.532</td>
<td>.675</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.380</td>
<td>0.479</td>
<td>0.410</td>
<td>0.434</td>
<td>0.486</td>
<td>0.448</td>
<td>0.503</td>
<td>0.388</td>
<td>0.501</td>
<td>0.502</td>
<td>0.471</td>
</tr>
<tr>
<td>Middle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group A</td>
<td>n</td>
<td>158</td>
<td>158</td>
<td>158</td>
<td>158</td>
<td>158</td>
<td>153</td>
<td>153</td>
<td>153</td>
<td>153</td>
<td>153</td>
<td>153</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.436</td>
<td>0.494</td>
<td>0.440</td>
<td>0.459</td>
<td>0.476</td>
<td>0.460</td>
<td>0.501</td>
<td>0.426</td>
<td>0.497</td>
<td>0.502</td>
<td>0.468</td>
</tr>
<tr>
<td>Upper</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group A</td>
<td>n</td>
<td>76</td>
<td>76</td>
<td>76</td>
<td>76</td>
<td>76</td>
<td>76</td>
<td>76</td>
<td>76</td>
<td>76</td>
<td>76</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>50</td>
<td>39</td>
<td>52</td>
<td>26</td>
<td>52</td>
<td>51</td>
<td>37</td>
<td>54</td>
<td>31</td>
<td>34</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>26</td>
<td>37</td>
<td>24</td>
<td>50</td>
<td>24</td>
<td>25</td>
<td>39</td>
<td>22</td>
<td>45</td>
<td>42</td>
<td>24</td>
</tr>
<tr>
<td>Group B</td>
<td>n</td>
<td>76</td>
<td>76</td>
<td>76</td>
<td>76</td>
<td>76</td>
<td>76</td>
<td>76</td>
<td>76</td>
<td>76</td>
<td>76</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>63</td>
<td>48</td>
<td>60</td>
<td>19</td>
<td>47</td>
<td>56</td>
<td>36</td>
<td>62</td>
<td>35</td>
<td>40</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>13</td>
<td>28</td>
<td>16</td>
<td>57</td>
<td>29</td>
<td>20</td>
<td>40</td>
<td>14</td>
<td>41</td>
<td>36</td>
<td>25</td>
</tr>
</tbody>
</table>

Note: Full score = 1.000.
5.4.2.2 Comparisons of Different Sequences of Answering Questions: Interactions With Learners’ L2 Reading Proficiency

A two-way (by question sequence and L2 reading proficiency) ANOVA\(^3\) was performed to assess the significance of the observed difference in means between the treatment and L2 reading proficiency, and the interaction between them. As is shown in Table 5.19, the analysis revealed a significant main effect for the participants’ L2 reading proficiency \((F(2, 794) = 20.374, p < .001)\), and a significant interaction between question sequence and L2 reading proficiency \((F(2, 794) = 4.152, p < .05)\), which means that the main effects did not have a straightforward relationship. A line graph for mean scores by L2 proficiency group (“Lower,” “Middle,” and “Upper”) and sequence group (Group A and Group B) was made, as is shown in Figure 5.1. In contrast, there was no main effect for question sequence \((F(1, 794) = 1.676, \text{n.s.})\).

Table 5.19

Results of a Two-Way ANOVA

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>(\eta^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question Sequence</td>
<td>1</td>
<td>8.056</td>
<td>8.056</td>
<td>1.676</td>
<td>0.033</td>
</tr>
<tr>
<td>L2 Proficiency</td>
<td>2</td>
<td>195.823</td>
<td>97.911</td>
<td>20.374***</td>
<td>0.803</td>
</tr>
<tr>
<td>Question Sequence × L2 Proficiency</td>
<td>2</td>
<td>39.903</td>
<td>19.952</td>
<td>4.152*</td>
<td>0.164</td>
</tr>
<tr>
<td>Error</td>
<td>788</td>
<td>3786.902</td>
<td>4.806</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note. \(*p < .05. ***p < .001.

\(\eta^2\) = effect size.

0.010 \(\leq \eta^2\) (small) < 0.059, 0.059 \(\leq \eta^2\) (medium) < 0.138, 0.138 \(\leq \eta^2\) (large; Cohen, 1988, pp. 284-287). Effect size of 0.009 and below is regarded as having no effect.

L2 Proficiency = L2 reading proficiency.

In terms of the effect size measured by \(\eta^2\), large effect sizes were revealed for the main effect for L2 reading proficiency and the interaction between question sequence and L2
reading proficiency; that is, $\eta^2 = 0.803$ and $\eta^2 = 0.164$, respectively. In turn, the effect for question sequence was small ($\eta^2 = 0.033$). These results were consistent with the results of significant testing.

![Figure 5.1](image)  
*Figure 5.1. Mean scores of tests of DNC by L2 reading proficiency and question sequence.*

To examine what patterns were found with regard to the interaction, one-way ANOVAs and $t$-tests of the Bonferroni approach, in which a significance level of $.05/k$ (group numbers) is used for each individual test, were carried out. One-way ANOVAs were performed to test the effect of the participants’ L2 reading proficiency level on reading comprehension scores by question sequence, whereas $t$-tests were performed to test the effect of question sequence on reading comprehension scores by L2 reading proficiency level.

With regard to Group A, on the one hand, the Brown Forsythe $F^*$ test showed no significant difference in mean scores among three different L2 proficiency groups ($F^* (2, 240.303) = 2.901, p = .057$). Through one of the strength of association measures, that is,
eta-squared, a small effect size was given ($\eta^2 = 0.016$) as well. Though it appeared that the higher L2 reading proficiency a learner had the higher score one would receive, that was not statistically confirmed. On the other hand, with regard to Group B, a significant effect for L2 reading proficiency level was revealed, $F (2, 390) = 20.949$, $p < .001$. Effect size was medium, $\eta^2 = 0.097$. The post-hoc comparisons (the Tukey’s HSD tests) revealed the significant differences between “Upper” and “Middle,” and “Upper” and “Lower” at the .05 level. That is, the mean score obtained by “Upper” was significantly higher than that obtained by “Middle,” and the mean score obtained by “Upper” was significantly higher than that obtained by “Lower.” There was no difference between the mean scores of “Middle” and “Lower.”

The $t$-tests were conducted repeatedly in order to examine the difference in the mean scores between Group A and Group B. In other words, I aimed to determine whether the effect of question sequences (i.e., one being the order in which readers answer from an inferential question scarcely related to the theme of passage to a question about the theme of passage, and the other one being in the opposite order) was different in each group.

Table 5.20

Results of T-Tests

<table>
<thead>
<tr>
<th>L2 Reading Proficiency Group</th>
<th>Group A</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n$</td>
<td>Mean</td>
<td>$SD$</td>
<td>$n$</td>
<td>Mean</td>
<td>$SD$</td>
<td>df</td>
<td>$t$</td>
<td>$g$</td>
</tr>
<tr>
<td>Lower</td>
<td>168</td>
<td>5.536</td>
<td>1.876</td>
<td>172</td>
<td>4.826</td>
<td>2.159</td>
<td>338</td>
<td>3.235***</td>
<td>0.351</td>
</tr>
<tr>
<td>Middle</td>
<td>159</td>
<td>5.824</td>
<td>2.263</td>
<td>143</td>
<td>5.378</td>
<td>2.229</td>
<td>300</td>
<td>1.723</td>
<td>0.198</td>
</tr>
<tr>
<td>Upper</td>
<td>76</td>
<td>6.289</td>
<td>2.529</td>
<td>76</td>
<td>6.803</td>
<td>2.332</td>
<td>150</td>
<td>-1.300</td>
<td>0.211</td>
</tr>
</tbody>
</table>

Note. Full score = 11.000.

$g$ = effect size. $0.200 \leq g$ (small) $< 0.500$; $0.500 \leq g$ (medium) $< 0.800$; $0.800 \leq g$ (large; Cohen, 1988). Effect size of 0.199 and below is regarded as having no effect.

A part of Table 5.16 was modified for this table.
The following results were observed: A significant difference was only found between two groups of “Lower,” showing that the mean score of Group A was higher than that of Group B ($t(338) = 3.235, p = .001$). In contrast, both in “Upper” and “Middle” the difference between Group A and Group B did not reach significance. The statistics, $t(150) = -1.300 (p = .195)$ were obtained for “Upper,” whereas $t(300) = 1.723 (p = .086)$ were obtained for “Middle.”

*Figure 5.2.* Cumulative scores of “Lower” on the fiscal 2003 DNC test (Passage 1).

*Figure 5.3.* Cumulative scores of “Lower” on the fiscal 2004 DNC test (Passage 2).
Figure 5.4. Cumulative scores of “Upper” on the fiscal 2003 DNC test (Passage 1).

Figure 5.5. Cumulative scores of “Upper” on the fiscal 2004 DNC test (Passage 2).

Figure 5.2 and Figure 5.3 reveal the mean cumulative scores from item 1 to item 5 on the fiscal 2003 test, and the cumulative scores from item 1 to item 6 on the fiscal 2004 test, respectively. As is shown, the mean cumulative scores of Group A were consistently higher than those of Group B.

In turn, Figure 5.4 and Figure 5.5 are the line graphs for the mean cumulative scores of “Upper” on the fiscal 2003 test, and on the fiscal 2004 test respectively. We can recognize
that totally different results were obtained for “Upper” than those obtained for “Lower.” Group B’s ascendancy over Group A was maintained through all procedures on both of the tests.

Furthermore, it is necessary to study the effect sizes. Due to a $t$-test, Hedges’s $g$ (see endnote 6 of Chapter 3) was computed. Small effects were seen in both “Lower” and “Upper,” and no effect was obtained in “Middle,” although the $g$ value of “Middle” was almost within the small effect size (see Table 5.20).

Following these results, in order to investigate what type of question was more likely to be affected by the question sequence, a Fisher’s exact test was run by item. Note again that the last items, item 5 in Passage 1 and item 6 in Passage 2, were about the main theme of the whole texts; the other items were all inferential questions. In addition, item 4 in Passage 1 and item 5 in Passage 2 were designed to be more strongly related to the main theme than the other questions. The data of “Lower” (see Table 5.21) and the data of “Upper” (Table 5.22) were targeted, which revealed a significant difference and a small effect size (i.e., “Lower”), or which revealed a small effect size only (i.e., “Upper”).

Table 5.21

Counts of “Lower” by Group and Results of Fisher’s Exact Tests

<table>
<thead>
<tr>
<th></th>
<th>Passage 1</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N = 340)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A (n = 168)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct</td>
<td>113</td>
<td>62</td>
<td>107</td>
<td>43</td>
<td>107</td>
<td>94</td>
<td>76</td>
<td>113</td>
<td>37</td>
<td>62</td>
<td>116</td>
</tr>
<tr>
<td>Incorrect</td>
<td>55</td>
<td>106</td>
<td>61</td>
<td>125</td>
<td>61</td>
<td>74</td>
<td>92</td>
<td>55</td>
<td>131</td>
<td>106</td>
<td>52</td>
</tr>
<tr>
<td>Phi</td>
<td>n.s.</td>
<td>0.115**</td>
<td>0.155**</td>
<td>n.s.</td>
<td>0.224**</td>
<td>n.s.</td>
<td>n.s.</td>
<td>0.117*</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>B (n = 172)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct</td>
<td>116</td>
<td>45</td>
<td>83</td>
<td>44</td>
<td>71</td>
<td>83</td>
<td>71</td>
<td>96</td>
<td>46</td>
<td>70</td>
<td>105</td>
</tr>
<tr>
<td>Incorrect</td>
<td>56</td>
<td>127</td>
<td>89</td>
<td>128</td>
<td>101</td>
<td>89</td>
<td>101</td>
<td>76</td>
<td>126</td>
<td>102</td>
<td>67</td>
</tr>
<tr>
<td>Phi</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. *p < .05. **p < .01.

$N = 340$. 

153
Firstly, with regard to “Lower,” the differences in numbers between Group A and Group B were found on the four items; that is, three inference questions less related to a main theme of a passage (i.e., items 2 and 3 in Passage 1, and item 3 in Passage 2) and one thematic question (i.e., item 5 in Passage 1).

Secondly, as to the “Upper” data, only one phi coefficient reached a statistically significant level < .05. When we compared the data of Group A with that of Group B on item 1 of Passage 1, more students in Group B selected the correct answer than in Group A. The question was one of the inferential questions which were not closely related to a main theme of the passage. Group A answered it first of all, whereas Group B answered it last. For the other items, there were no differences between Group A and Group B.

Table 5.22

Counts of “Upper” by Group and Results of Fisher’s Exact Tests

<table>
<thead>
<tr>
<th></th>
<th>Passage 1</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>A (n = 76)</td>
<td>Correct</td>
<td>50</td>
<td>39</td>
<td>52</td>
<td>26</td>
<td>52</td>
<td>51</td>
<td>37</td>
<td>54</td>
<td>31</td>
<td>34</td>
<td>52</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Incorrect</td>
<td>26</td>
<td>37</td>
<td>24</td>
<td>50</td>
<td>24</td>
<td>25</td>
<td>39</td>
<td>22</td>
<td>45</td>
<td>42</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>B (n = 76)</td>
<td>Correct</td>
<td>63</td>
<td>48</td>
<td>60</td>
<td>19</td>
<td>47</td>
<td>56</td>
<td>36</td>
<td>62</td>
<td>35</td>
<td>40</td>
<td>51</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Incorrect</td>
<td>13</td>
<td>28</td>
<td>16</td>
<td>57</td>
<td>29</td>
<td>20</td>
<td>40</td>
<td>14</td>
<td>41</td>
<td>36</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>Phi</td>
<td>0.195*</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td></td>
</tr>
</tbody>
</table>

*p < .05.

The below dendrograms show the results of cluster analyses. The dendrograms in Figure 5.6 and Figure 5.7 were made from the scores of “Lower,” while the dendrograms in Figure 5.8 and Figure 5.9 were made from the scores of “Upper.” To make it easier to compare the figure of Group A with the figure of Group B, I adopted these arrangements.
Figure 5.6. The dendrograms of the “Lower” Group A’s (see the above dendrogram) and “Lower” Group B’s (see the below dendrogram) fiscal 2003 DNC test. Item 5 is a thematic question.

Figure 5.7. The dendrograms of the “Lower” Group A’s (see the above dendrogram) and “Lower” Group B’s (see the below dendrogram) fiscal 2004 DNC test. Item 6 is a thematic question.
Figure 5.8. The dendrograms of the “Upper” Group A’s (see the above dendrogram) and “Upper” Group B’s (see the below dendrogram) fiscal 2003 DNC test. Item 5 is a thematic question.

Figure 5.9. The dendrograms of the “Upper” (see the above dendrogram) Group A’s and “Upper” Group B’s (see the below dendrogram) fiscal 2004 DNC test. Item 6 is a thematic question.
Among these four pairs, only the one pair in Figure 5.8 proved coincident to each other. They were made through the following progression: (a) At first, item 1 and item 3 were combined, (b) item 2 was integrated into the cluster, (c) item 5 (a thematic question) comprised one cluster with item 1, item 3, and item 2, and (d) finally, the big cluster was combined with item 4. However, as to the other pairs the dendrograms of Group A and Group B did not match.

5.4.3 Discussion

Since test validation has already been discussed in the Results section, this section will focus on a discussion of hypothesis testing. As is shown in Table 5.19, the significant interaction between the variable of question sequence and the variable of L2 reading proficiency was found by using a two-way ANOVA. The further analyses revealed that there was no statistical difference in means among three L2 reading proficiency groups in Group A, but that there were between “Upper” and “Middle,” and between “Upper” and “Lower” in Group B. The line of Group B in Figure 5.1 also shows a much steeper rise than that of Group A, and two lines are crossed between “Middle” and “Upper.”

Furthermore, these findings were supported by the comparisons between Group A and Group B in each of L2 reading proficiency groups (see Table 5.20); to be more specific, that “Lower” Group A obtained the higher mean score than “Lower” Group B, and “Upper” Group B showed a tendency to obtain a higher mean score than “Upper” Group A. Note that these results were not only based on the results of significant testing, but also on the results of strength of association (i.e., the effect size measure by \( \eta^2 \)); in addition, even the results of small effect sizes were included in the present discussion. This was because the present study employed a rather weak operation as to the question sequence; for example,
(a) firstly, a series of questions were presented on paper, (b) secondly, one and a half minutes were given to each question, which appeared to be relatively short if a student wanted to deliberate upon them thoroughly, and (c) thirdly, no correct answer was given during their answering of 11 questions. Generally speaking, questions by teachers in the classroom are often accompanied by interactions between teachers and students, and correct answers are going to be confirmed immediately after questions are given. Skinner’s (1968) programmed instruction was developed to intensify this point. Considering these differences between the present experimental context and the usual teaching context, it would be fruitful as a suggestion for future research to discuss them even with small effect size.

The results of the hypothesis testing as to Hypothesis 6A and Hypothesis 6B were as follows: Hypothesis 6A (i.e., The question sequence in which learners answer inferential questions during the early stages of the sequence, far from thematic questions which are presented at the end of sequence, is more beneficial for learners of lower L2 reading proficiency than the opposite sequence) was supported by both the significance testing (i.e., the t-tests) and the strength of association; however, with regard to Hypothesis 6B (i.e., The question sequence in which learners answer a thematic question during the early stages and answer inferential questions far from a thematic question during the final stages is more beneficial for learners of higher L2 reading proficiency than the opposite sequence), a tendency was shown by the strength measure of association, but it was not supported by significance testing. Therefore, it would be safe to say that Hypothesis 6B was partly supported.

With regard to the results of Hypotheses 6A and 6B, two interpretations can be offered; that is, one is from the viewpoint of learners’ linguistic deficiency, and the other one is from the viewpoint of cognitive loads on learners. I do not believe that they are mutually exclusive,
but rather, complementary ones. These interpretations will be mentioned one by one in the following paragraphs.

The first interpretation focuses on learners’ L2 knowledge, in particular, their vocabulary knowledge and grammatical knowledge. Basically, it can be considered that learners of lower L2 reading proficiency do not have enough vocabulary knowledge and grammatical knowledge, and that they cannot understand propositional meanings of a text correctly; furthermore, they are not able to use inference generation processes even if they may have the ability to make inferences. For such learners (learners in “Lower” in this study), providing questions which ask for why a character in a text did a certain behavior, how a character felt about a certain event, and so forth, described in each situation would have been helpful to guide their reading to proper understanding, and finally to comprehend a theme of a passage. This sequence of the local-to-global transition, or the step-by-step way of questioning may have functioned as helpful guides for “Lower.” However, you may still ask how such inference questions can complement their deficiency in vocabulary knowledge and grammatical knowledge. One of the answers seems to be consistent with the results of Study 2. This study showed that inference questions increased learners’ inference making and understanding of literal meanings when there were some relationships between inference making and understanding of literal meanings.

Learners of higher L2 reading proficiency, on the other hand, are basically considered to have adequate knowledge both of vocabulary and grammar, and they can grasp propositional meanings of a text correctly by themselves; furthermore, they are able to use their ability to make inferences as freely as they want. The local-to-global transition or the step-by-step way of questioning did not have larger effects on higher L2 reading proficiency learners (i.e., “Upper”) than lower L2 reading proficiency learners (i.e., “Lower”). In other
words, the sequence may not have functioned as helpful guides of their reading comprehension because the learners have already completed those processes by themselves with their richer vocabulary knowledge and grammatical knowledge. In contrast, the global-to-local transition had small but positive effects on learners of higher L2 reading proficiency (see Table 5.20). A thematic question placed before the other minor questions may have revealed successfully a framework of the passages. This assumption is in line with previous studies which investigated the relationships between readers’ schema (or background knowledge) and their understanding. I would like to refer to Bransford and Johnson’s (1973) research, which used a well-known reading material entitled “washing clothes.” The following is the original passage.

The procedure is actually quite simple. First you arrange things into different groups. Of course, one pile may be sufficient depending on how much there is to do. If you have to go somewhere else due to lack of facilities, that is the next step, otherwise you are pretty well set. It is important not to overdo things. That is, it is better to do too few things at once than too many. In the short run this may not seem important but complications can easily arise. A mistake can be expensive as well. At first the whole procedure will seem complicated. Soon, however, it will become just another facet of life. It is difficult to foresee any end to the necessity for this task in the immediate future, but then one never can tell. After the procedure is completed, one arranges the materials into different groups again. Then they can be put into their appropriate places. Eventually they will be used once more and the whole cycle will then have to be repeated. However, that is part of life. (Bransford & Johnson, 1973, p. 400)

You may recognize that if you are given the title, “washing clothes” before you start reading, you would be able to comprehend these sentences easily; however, if you are not given it before reading, you would have many troubles understanding each sentence.
Whether a title would be given beforehand or not is followed by the discussion of whether readers are able to activate their schema effectively or not. It is because almost all persons would have had experience in washing clothes; and then, when they are informed that the passage is about such housework, they can remember how they did it in the past and use that knowledge effectively. As can be expected, Bransford and Johnson (1973) showed that those participants who were given the title before reading were able to recall the passage better than those who read the passage without the title. This experiment was replicated by Muramoto (1998) with Japanese university students in their L1, in which both a recall technique and a recognition technique were employed. The results revealed that, in spite of the data collection methods, the students who were given the title in advance achieved significantly higher scores than the students who were not given the title. In sum, questions as to a theme of a text may have activated some knowledge of the higher L2 reading proficiency learners; and although the positive effect was small it helped them to answer the following questions.

The second interpretation on the results of Hypotheses 6A and 6B focuses on readers’ cognitive load or working memory (Gavens & Barrouillet, 2004). Thus, the following discussion will proceed based on the assumption that these terms (i.e., cognitive resources and working memory capacity) have, in this section, some elements in common. This discussion takes the same direction as that in Nassaji (2002). The central issue of Nassaji’s paper was that if a reader can construct mental representations at the textbase level effectively, their working memory capacity is not so much consumed with this lower-level processing, and is distributed among higher-level processing or inference generation processing.

To answer a thematic question immediately after reading a passage, without looking at
any other questions, must place some overwhelming cognitive load on the present participants. The narrative stories were taken from the previously administered DNC tests, which were designed to be taken by 12th graders or by students who had already left high school for graduation. Moreover, around a half of the students in the present study were 10th-grade students, and more than a quarter of the students were 11th-grade students (see Table 5.2). In brief, the participants were one year or two years behind the assumed examinees of the original tests. Considering that most of these students started taking English classes at school when they were in the 7th grade, these gaps in ability appear to be large.

In addition to difficulties inherent in the materials, finding a theme of a passage calls on complex reading processes; readers are required to be engaged simultaneously in both bottom-up processing and top-down processing, and to recognize what the story implied or what message the writer hoped to convey to the readers through the entire passage. Although four choices were given to the present participants, the decisions must have been challenging for them. Let us look at the thematic questions and the choices used in the present study. Both thematic questions were the exactly same wording, viz., “What is the main theme of this story?” Firstly, the choices of Passage 1 (the fiscal 2003 test) were as follows: “Brazilian lifestyle,” “intercultural understanding,” “Japanese education” and “childhood memories.” The passage dealt with all of these topics; for example, the story was about the main character who had lived in Brazil in her childhood, but after she moved to Japan with her parents she tried to adapt to Japanese school life. Through several conflicts with other Japanese students, the main character and some of the students recognized the importance of intercultural understanding. The correct answer was “intercultural understanding.” This was because the choice described the author’s message. Secondly, the
choices to the thematic question of Passage 2 (the fiscal 2004 test) were “swimming,” “friendship,” “a rival,” and “competition.” “Friendship” was the right answer, and the other three were distractors. These choices were made through the pilot study and a discussion between the author and a native speaker of English (see section 5.2), and used to infer whether a student was able to understand the author’s message.

Taking all of these factors into consideration, to answer thematic questions first and immediately after reading passages almost certainly would have been cognitively demanding for the EFL learners, especially for lower L2 reading proficiency learners. It should be noted, therefore, that this seemed to be one of the reasons why the mean score obtained by the “Lower” Group B was significantly lower than the mean score obtained by the “Lower” Group A.

Here, let us review Horiba (1996) and Ikeno (2002), which focused on the limited amount of readers’ cognitive resources or working memory capacity. Horiba’s study was a comparative study of American university students of Japanese and native speakers of Japanese, and investigated how their limited cognitive resources were distributed to several levels of reading processes by examining their think-aloud protocols. The conclusion was that there existed a large difference between L2 learners and native speakers in terms of the distribution of cognitive resources; to be more specific, the L2 learners devoted much of their cognitive resources to lower-level processes whereas the native speakers of the target language devoted much of their cognitive resources to higher-level processes (see section 2.1). Horiba’s study indicates the following crucial points: (a) Readers’ language proficiency levels determine what processes their cognitive resources are spent on, and (b) readers with less proficient language ability are likely to consume much of their cognitive resources in processing smaller parts of texts such as characters, words and sentences; as a result, not
enough of their resources remain to construct textual coherence.

A correlation study was conducted by Ikeno (2002) by administering several tests to Japanese’ EFL learners. The following results were reported in his study. Moderate or almost moderate correlations were found between the scores of the L2 reading span test (a modified version of the reading span test developed by Osaka & Osaka, 1994, as cited in Ikeno) and the scores of the following tests: (a) “a text structure prediction test,” which was related to one of the higher-level processes (see section 2.1), (b) “a sentence verification test,” (c) “a lexical judgment semantic test,” and (d) “a grammaticality judgment test.” In sum, the main finding was that individual differences in the capacity of working memory were reliably correlated with lower-level processing efficiency as well as a higher-level processing. Parallels can be drawn with the Horiba’s (1996) results; readers’ limited amount of cognitive resources or working memory capacity is apportioned to multiple processes. Therefore, if readers are deficient in lower-level processes, as is often found in L2 learners’ reading, too few cognitive resources or working memory capacity remain to be distributed into higher-level processes.

We can summarize the discussion from the perspective of readers’ cognitive load or working memory as follows: When the lower L2 reading proficiency learners looked at a question concerning a whole passage immediately after reading the passages, they had already depleted their cognitive resources or working memory capacity in processing more basic units concerning lower-level processes, and they did not have any more resources or working memory to process the content of what the question inquired. Therefore, giving such a cognitively demanding question to the lower L2 reading proficiency learners first of all had no effect, or rather, had an ill effect on their comprehension. In contrast, this was not the case with higher L2 reading proficiency learners; they could afford to deliberate about a
theme of an entire passage even when they were obliged to do so immediately after reading the passage. That is to say, this deliberation must have been easier, rather than being cognitively demanding for the learners.

More interestingly, the related previous studies in fields other than English language teaching can also be explained reasonably from the point of view of learners’ cognitive resources or working memory. Firstly, van den Broek et al. (2001) implemented L1 reading research for 4th-, 7th-, 10th-graders and college students to investigate the effect of inferential questions (see section 2.3). Their results of a recall test revealed that the positive effect was found only in the college students, not in the other students. Given as many as 14 questions per text, this task must have been cognitively demanding; therefore, we can speculate that the positive effect was only seen in the students who still had many resources available to process those questions. Another study, which should be referred to here, is Fushimi’s (1992) research. The sequence in which students were given “a well-known case” first, and “a misunderstood case” second had a more positive effect than the opposite sequence, that is, in which students were given “a misunderstood case” first, and “a well-known case” second. The latter sequence must have been more difficult for those students who had at that time not yet acquired much scientific knowledge. Thus, as to Fushimi’s results, completing the less demanding task first also had a positive effect on the students’ acquisition.

I would finally like to discuss the results of the Fisher’s exact tests and the cluster analyses in the remainder of this section. These statistical analyses were employed to examine if there was a meaningful effect of question sequence on EFL learners’ reading comprehension, and on what item the effect would particularly be found. It should be noted, however, that the Fisher’s exact test and the cluster analysis were almost certainly based on
different viewpoints concerning the data analysis. The former, on the one hand, concerned the number of students who chose a correct answer and an incorrect answer by group (i.e., Group A and Group B), and determined whether these were an even distribution or an uneven distribution. On the other hand, the latter analysis was for the grouping of items in accordance with similarities in patterns.

The following findings common to “Lower” and “Upper” were found by the Fisher’s exact tests. First, there seemed to be no particular question type which had been directly affected by the question sequence. Results of “Lower” showed that the differences were found in three inferential questions, which were only slightly related to the central theme of the passages, and in one thematic question (see Table 5.21). The data of “Upper” showed that in one inferential question, which was scarcely related to the central theme of the passages, the difference between Group A and Group B was found (see Table 5.22). Second, considering the sequences of these five items, four for “Lower” and one for “Upper,” the effect of question order was likely to appear on the latter questions. It may take a certain time before the effect of question sequence starts to work positively. This does not contradict our common knowledge.

With regard to cluster analyses, the two dendrograms in Figure 5.8 represented almost the same figures. That is to say, both of the data obtained from the fiscal 2003 test of “Upper” Group A and Group B showed almost the same pattern. In contrast, one dendrogram was different from the other in Figures 5.6, 5.7, and 5.9, that which showed the different patterns of the data obtained from Group A and Group B. In sum, we could not find in the results of the cluster analysis any consistent pattern for the same question sequence groups.

Thus, the Fisher’s exact tests and the cluster analyses implied that the effect of question
sequence was not limited to a particular question. It would be more accurate to say that the effect was seen across a series of questions.

In sum, Study 3 offered an important implication as to how we should present questions to learners in order to develop their reading comprehension. This is related to the context of reading instruction. How inference questions are used in the context of assessment constitutes another possible avenue for future research.

5.5 Summary

Study 3 was implemented to find out, in order to help Japanese EFL learners comprehend passages better, in what order a series of inference questions should be arranged. In addition, interaction between question sequence and learners’ L2 reading proficiency levels was examined. Inference questions were ranked from the viewpoint of the degree of proximity to a text theme, and questions asking for themes of entire passages were labelled as thematic questions in particular. The following sequences, which were exactly opposite to each other, were prepared: (a) the sequence in which learners answer inference questions during the early stages of the sequence, far from thematic questions which are presented at the end of sequence, and (b) the sequence in which learners answer a thematic question during the early stages and answer inference questions far from a thematic question during the final stages.

The main finding was that for the lower L2 reading proficiency learners, the sequence (a) mentioned in the previous paragraph had a positive effect on their reading comprehension. In contrast, the learners of higher L2 reading proficiency indicated the tendency that they benefited from the sequence (b), which was mentioned in the previous paragraph. However, it should again be noted that the results regarding the higher L2
reading proficiency did not reach a significant level. Finally, it was shown that these sequence effects were not limited to a particular type of question, but were spread over a series of questions which had appeared during later stages.

**Endnotes**

1. Z transformation was applied for the computation of means.

2. Although the homogeneity of variance was not assumed by the Levene procedure, the $F$ statistic was used in the present case. This was because its robustness is empirically reported (Glass & Hopkins, 1996, pp. 402-405), and neither the Brown-Forsythe $F^*$ test nor the Welch $F'$ were provided by even the latest version of the statistical package SPSS for a two-way ANOVA.

3. This is the same as is mentioned in endnote 2.

4. This analysis was used because of the lack of homogeneity of variance.
6. Study 4: Questions on Reading Tests Which Measure Higher-Level Processes and Lower-Level Processes

6.1 Purposes and Research Hypotheses

Study 3 indicated that L2 learners of a lower reading proficiency level have much more difficulty answering, in particular, thematic questions than L2 learners of a higher reading proficiency level. From the perspective of reading processes, a thematic question employs higher-level processing skills—Grabe (2000) defines higher-level processing as “working with larger units of information and information contributed by the reader” (p. 233; for more information, see section 2.1). This provides us with some implications concerning language assessment. That is to say, reading comprehension tests should include enough questions to assess examinees’ higher-level processing skills as they do to assess lower-level processes. Therefore, the first aim of the present study was determined in order to investigate how some popular English reading tests for EFL or ESL learners are organized. Considering the number of the examinees and L2 proficiency levels of the current participants in Study 4, the following tests were selected: (a) the Society for Testing English Proficiency (STEP) test, (b) the test administered by the National Center for University Entrance Examinations (Daigaku Nyushi Center: DNC), and (c) TOEFL. The following hypothesis was then constructed.

Hypothesis 7. The STEP test, the DNC test, and TOEFL contain as many questions that measure higher-level processes as the questions that measure lower-level processes.

The second aim of Study 4 was to survey the correlation between the scores for higher-level processing questions and the scores for lower-level processing questions.
Under the administration of the present research, the following assumptions were made: (a) If we obtain a higher correlation coefficient between these scores, that is, the ranking of the participants’ scores are not changed by the types of questions (higher-level processing questions and lower-level processing questions), the claim that test developers should make higher-level processing questions as well, would be weakened, however, (b) if we, on the contrary, obtain a lower correlation coefficient between these scores, that is, the ranking of the participants’ scores are changed by the types of questions (higher-level processing questions and lower-level processing questions), the claim would be strengthened.

Before predicting the correlation coefficient between the scores of higher-level processing questions and the scores of lower-level processing questions, we must consider Ikeno’s (2002) results (see section 5.4.3). His study revealed that the scores of “a text structure prediction test,” which is higher-level processing, had moderate or almost moderate correlations with the scores of “a L2 lexical semantic judgment test” and the scores of “a L2 sentence verification test,” all of which are lower-level processing. However, the scores of “a text structure prediction test” had no correlation with the scores of “number matching,” “L2 word matching,” and “L2 grammaticality judgment,” all of which are also lower-level processing. Considering this, the second hypothesis of the current research, Hypothesis 8 was formulated.

Hypothesis 8. Scores of the questions which measure lower-level processes correlate low or moderately with the scores of the questions which measure higher-level processes.

The criteria for judging the correlation coefficients such as “low” and “moderate” will be adopted from Kiyokawa (1990; see section 4.2.5). That is to say, the coefficients $| .700 |$
and greater are considered to be “high”; those between \( |.400| \) and \( |.699| \) are considered to be “moderate”; and those between \( |.200| \) and \( |.399| \) are considered to be “low.” Finally, coefficients \( |.199| \) and below are considered to have almost no correlation.

6.2 Pilot Study

6.2.1 Method

A pilot study was conducted to test Hypothesis 7 that the STEP test, the DNC test, and TOEFL contain as many questions that measure higher-level processes as the questions that measure lower-level processes. This pilot study is also intended to provide foundations for the main study.

These reading tests (see Table 6.1) were selected because they are popular as an EFL or ESL test in Japan and other parts of the world, and because the number of examinees is quite large. Various types of methods have been developed as assessment instruments in recent years, for example, a cloze test, short-answer questions, multiple-choice questions, a matching technique, a gap-filling test, and true/false questions (Alderson, 2000). Without exception, these three tests contained several types of test methods; however, the current research targeted question-answer, multiple-choice discrete-point items only. This was because the following procedures for categorization were adopted: Raters examined both the content of questions and the part of texts related to the answers, and categorized a question into either type (What types and how many types were made for the categorization will be explained later.) For example, with regard to true/false questions, different processes are included within one item so as to identify antecedents of pronouns, to understand information explicitly written in texts and so on; thus, because it was assumed beforehand that allotting these questions to one category would be impossible, the test methods other
than the question-answer, multiple-choice technique were not targeted in the present study.

Table 6.1

*Reading Tests Used for Categorization in Pilot Study*

<table>
<thead>
<tr>
<th>Test (Source)</th>
<th>Part (Number of Items and Test Methods)</th>
<th>Excluded Item From the Categorization</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. STEP 2nd Grade Test (administered in June 2004; Society for Testing English Proficiency, n.d.)</td>
<td>Parts 4A to 4C (12 four-option, multiple-choice questions)</td>
<td>8 items in 3A and 3B</td>
</tr>
</tbody>
</table>
| 2. DNC Test (administered in fiscal 2004) | Part B of the 4th section and Part A of the 6th section (9 four-option, multiple-choice questions) | (a) 7 items in the 3rd section  
(b) 1 item in Part A of the 4th section  
(c) 5 items in the 5th section  
(d) 3 items in Part B of the 6th section |
| 3. TOEFL (ETS, 1999, pp. 30-41) | Section 3 of the Practice Test A (50 four-option, multiple-choice questions) | — |

*Note.* ETS = Educational Testing Service.

A part of the STEP test and TOEFL are in Appendix 6-A and Appendix 6-B.

The following six types were prepared for reading questions, “Paraphrase Questions,” “Inference Questions,” “Thematic Questions,” “Referential Questions,” “Vocabulary Questions,” and “Text Organization Questions.” These types, which were made following previous studies (Davey, 1988; Freedle & Kostin, 1993), were expected to function properly as a framework for categorization. A definition of each question type and an example is shown in Table 6.2.
Table 6.2

Six Categories Used for Reading Questions in Pilot Study

<table>
<thead>
<tr>
<th>Question Type</th>
<th>Definition</th>
<th>Example of Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Paraphrase</td>
<td>Correct answers can be obtained when a stem and part of a text are paraphrased. And frequently, vocabularies in stems overlap considerably with part of a text. These are labelled as explicitly stated questions in some previous studies.</td>
<td>What happened soon after HOV lanes were introduced?</td>
</tr>
<tr>
<td>2. Inference</td>
<td>Correct answers can be obtained when inferences are generated.</td>
<td>What can be inferred about the illuminating gas described in the second paragraph?</td>
</tr>
<tr>
<td>3. Thematic</td>
<td>These ask a main topic of a paragraph or a passage, or an appropriate title to a passage.</td>
<td>What is the author’s main point in the first paragraph?</td>
</tr>
<tr>
<td>4. Referential</td>
<td>These ask antecedents of pronouns.</td>
<td>The word ‘this’ in line 8 refers to…</td>
</tr>
<tr>
<td>5. Vocabulary</td>
<td>These ask the meanings of words, without reference to contextual information.</td>
<td>The phrase ‘served as’ in line 6 is closest in meaning to…</td>
</tr>
<tr>
<td>6. Text Organization</td>
<td>These ask the structures of texts (e.g., comparison or contrast, time order), or the parts where a certain topic is stated.</td>
<td>Which of the following best describes the organization of the passage?</td>
</tr>
</tbody>
</table>

Moreover, “Inference Questions” were divided into subcategories, as can be seen in Table 6.3, with reference to Graesser et al.’s (1994) inference types (see Table 2.1). However, note that some modifications were made in the framework particularly in terms of the four points: (a) “Case structure role assignment,” which appeared in Graesser et al., was excluded since these questions seemed to appear more frequently in a grammar test, (b) although “thematic” and “referential” were considered to be members of inferences in Graesser et al.’s framework, categories of “Thematic Questions” and “Referential Questions” were made independently of “Inference Questions,” (c) “author’s intent,” which appeared in Graesser et al., was integrated into “Thematic Questions,” because these appeared to share concepts, and (d) “instantiation of noun category,” which appeared in Graesser et al., was integrated into “Vocabulary Questions.”
A total of 71 items (see Table 6.1) were allotted to either type, respectively. Judgments were made by two nonnative speakers of English independently, who were both graduate students majoring in English language education as a foreign language, one of whom was the author. The following points were confirmed before we started categorizing them: (a) If an item seemed to fit into more than one category, it should be judged focusing on the most salient feature (Lumley, 1993), and (b) if causal relations were explicitly marked such as “because,” it would fall into a category of “Paraphrase Questions”; in contrast, if it was not marked, it would be in a subcategory of “Inference Questions,” “Causal Antecedent” or “Causal Consequence.”

6.2.2 Results and Discussion

The inter-rater reliability was high, 85% agreement--because the data was categorical data, a correlation coefficient was not computed--and all of the disagreements were resolved through discussion between the raters.
Table 6.4

Types of Questions in STEP Test, DNC Test and TOEFL

<table>
<thead>
<tr>
<th></th>
<th>STEP Test [Section]</th>
<th>DNC Test [Section]</th>
<th>TOEFL [Passage]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Paraphrase Questions</td>
<td>35 [4A], 36 [4A], 37 [4B], 38 [4B], 39 [4B], 40 [4B], 41 [4C], 42 [4C], 43 [4C], 44 [4C], 45 [4C]</td>
<td>34 [4], 35 [4], 36 [4], 37 [4], 40 [4B], 41 [4C], 42 [4C], 43 [4C], 44 [4C], 45 [4C]</td>
<td>1 [1], 4 [1], 7 [1], 12 [2], 15 [2], 23 [3], 27 [3], 28 [3], 29 [3], 31 [3], 32 [3], 33 [3], 36 [4], 38 [4], 40 [4], 50 [5]</td>
</tr>
<tr>
<td>2b. Superordinate Goal</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>2c. Character Emotional Reaction</td>
<td>--</td>
<td>43 [6], 47 [6]</td>
<td>--</td>
</tr>
<tr>
<td>2d. Causal Consequence</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>2e. Instrument</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>2f. Subordinate Goal-Action</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>2g. State</td>
<td>--</td>
<td>--</td>
<td>5 [1], 13 [2], 18 [2], 21 [2], 42 [5], 46 [5]</td>
</tr>
<tr>
<td>2h. Emotion of Reader</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>3 Thematic Questions</td>
<td>--</td>
<td>--</td>
<td>11 [2], 22 [3], 35 [4], 48 [5]</td>
</tr>
<tr>
<td>4. Referential Questions</td>
<td>--</td>
<td>--</td>
<td>3 [1], 8 [1], 14 [2], 26 [3], 37 [4], 45 [5]</td>
</tr>
<tr>
<td>5. Vocabulary Questions</td>
<td>--</td>
<td>--</td>
<td>2 [1], 6 [1], 16 [2], 17 [2], 19 [2], 24 [3], 25 [3], 30 [3], 39 [4], 41 [5], 44 [5], 47 [5], 49 [5]</td>
</tr>
<tr>
<td>6. Text Organization Questions</td>
<td>--</td>
<td>--</td>
<td>9 [1], 10 [1], 20 [2], 34 [3]</td>
</tr>
</tbody>
</table>

*Note.* Underlined items did not receive an agreement before discussion.

“2g. State” was the most difficult category to be judged. This may have been because it
required careful judgment of the raters. For example, when there appeared to be a causal relationship, it would be in “2a. Causal Antecedent” or “2d. Causal Consequence”; however, when there did not appear to be a causal relationship, it would be in “2g. State.” In contrast, all of the judgments on what were finally categorized into “3. Thematic Questions,” “4. Referential Questions,” “5. Vocabulary Questions” and “6. Text Organization Questions” were perfectly agreed (see Table 6.4).

The number of items and proportion of each category are shown in Table 6.5. Results of TOEFL, the STEP test and the DNC test will be reported one by one. Furthermore, information open to the public such as in test manuals and materials on the web will be mentioned as well, to compare the empirical results with what these tests intend to measure.

With regard to TOEFL, the questions were across all of the types; the largest group was “Paraphrase Questions” to which 32% of questions on the test belonged; the second largest group was “Vocabulary Questions,” to which 26% of questions on the test belonged; and the smallest groups were “Thematic Questions” and “Text Organization Questions,” to which 8% of questions on the test belonged, respectively. What should be noted about TOEFL is that at least 4 out of 50 items were classified into each category.

Let us review the overview of TOEFL and the test manual here. TOEFL measures “the ability of nonnative speakers of English to use and understand North American English as it is spoken, written and heard in college and university settings” (Educational Testing Service, n.d.), and most of the examinees are planning to enter and study at universities in English speaking countries such as the U.S., Canada and Australia. That is, this proficiency test targets English for an academic purpose. The following statements were found regarding the reading comprehension section in TOEFL Test & Manual (Educational Testing Service, 1997): “The questions test information that is stated in or implied by the passage, as well as
knowledge of some of the specific words as they are used in the passage” (p. 12). In accordance with this statement, we can expect that reading questions as to explicit statements in a passage, those as to what can be inferred during or after reading a text, and vocabulary questions. The above-mentioned results, which we obtained, were almost coincident with these expectations.

Table 6.5

<table>
<thead>
<tr>
<th>Question Type</th>
<th>STEP Test</th>
<th>DNC Test</th>
<th>TOEFL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Paraphrase Questions</td>
<td>11 (92%)</td>
<td>4 (44%)</td>
<td>16 (32%)</td>
</tr>
<tr>
<td>2. Inference Questions</td>
<td>1 (8%)</td>
<td>5 (56%)</td>
<td>7 (14%)</td>
</tr>
<tr>
<td>3. Thematic Questions</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>4 (8%)</td>
</tr>
<tr>
<td>4. Referential Questions</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>6 (12%)</td>
</tr>
<tr>
<td>5. Vocabulary Questions</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>13 (26%)</td>
</tr>
<tr>
<td>6. Text Organization Questions</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>4 (8%)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>12 (100%)</td>
<td>9 (100%)</td>
<td>50 (100%)</td>
</tr>
</tbody>
</table>

With regard to the STEP test, all of the questions except one were grouped into “Paraphrase Questions,” which are described as “Correct answers can be obtained when a stem and part of a text are paraphrased. And frequently, vocabularies in stems overlap considerably with part of a text” (see Table 6.2). It is critical to examine these results in a comparison with the aim of this test development. The object of the STEP test is described on the web to measure ability of the examinees to use English language necessary for daily social life, and the following statements are found as well.

1. All test items are constructed based on practicality.3
2. In addition to daily materials, topics discussed outside of Japan in the newspaper, in magazines, on television, on the web, or on business are adopted...

(Society for Testing English Proficiency, n.d.; translation mine)

Considering that this test focuses on “practicality” or daily social life, and that reading materials are often selected from areas of mass media and business (see the above citation), the emphasis seems to be on the following ability: (a) to scan the critical points in texts and understand their literal meaning as fast as possible and as accurately as possible, and (b) to understand the main idea of a passage as fast as possible and as accurately as possible. Thus, we can reasonably expect that most of the questions on the STEP test would ask for information explicitly written in texts and the main ideas of a passage.

Thus, some of the results as to question types could be assumed after considering the aims of assessing learners’ practical English language ability. Nevertheless, the result of there being no question which asks for main idea did not meet my anticipation. If I interpret these results favorably, we should also consider time allotment.

A total of 75 minutes are devoted to the STEP 2nd grade test, in which four sections are comprised (i.e., vocabulary section, grammar section, composition section, and reading comprehension section). Among these sections, the reading comprehension section accounts for approximately half of this test. Therefore, by simple calculation, about 35-40 minutes are going to be taken for reading section. In contrast, TOEFL (paper and pencil version) specifies 55 minutes for “Reading Comprehension” (Educational Testing Service 1999). If we think of these differences in time allotment between the STEP test and TOEFL, we may be able to interpret the results of the STEP test, which is in favor of “Paraphrase Questions,” according to the lack of time for assessment. This is because the STEP test may not have enough time to assess multiple levels of reading processing within this time frame.
Questions on the DNC test were categorized into two categories of either “Paraphrase Questions” or “Inference Questions.” This was the same as the STEP test; however, the test differed from the other in the proportion of “Paraphrase Questions” and “Inference Questions.” To be more specific, the questions on the DNC test were fairly evenly distributed into these types. Thus, the DNC test can be placed in the middle of the continuum whose ends are TOEFL and the STEP test in terms of question types. Moreover, note the questions grouped into “Inference Questions”; TOEFL was biased in favor of “State” inferences, whereas the DNC test contained “Superordinate Goal” and “Character Emotional Reaction” as well as “Causal Antecedent.” In sum, with a focus on “Inference Questions” on TOEFL and the DNC test, more variety of questions were found on the DNC test than TOEFL (see Table 6.4).

The results of the DNC test will be discussed with reference to the Course of Study for Upper Secondary School (Ministry of Education, Science, Sports and Culture, 1999). The Japanese education system is organized on the basis of this bulletin. Among six subjects as to foreign language education (i.e., “Aural/Oral Communication I,” “Aural/Oral Communication II,” “English I,” “English II,” “Reading,” and “Writing”), we will cite the subject of “Reading” in the followings:

2 Contents

(1) Language Activities

The following communicative activities should be conducted in concrete language-use situations so that students play the role of receivers and senders of information, ideas, etc.

A To obtain necessary information, to make an outline, and to summarize the main points by reading texts.

B To understand the writer’s intentions etc. and to organize and transmit one’s own ideas
etc. about theme by reading texts.

C To read stories etc. and talk or write about one’s own impressions.

D To read passages aloud so that the content and one’s interpretation can be transmitted to the listener. (p. 125)

The statement C in the Course of Study is concerned with “2h. Emotion of Reader,” a subcategory of “Inference Questions.” Thus, the following questions were expected to be seen in the DNC test before I start categorizing questions: (a) questions asking for explicit information, that is, “Paraphrase Questions” (from the statement A), (b) questions asking for main ideas, that is, “Thematic Questions” (from the statements of A and B), and (c) questions asking for inferences, that is, “Inference Questions” (from the statement of C). The results indicated that four questions were “Paraphrase Questions,” that five were “Inference Questions,” but that no question was a “Thematic Question” (see Table 6.5). With regard to there being no “Thematic Questions,” it should be noticed again that a lot of items in the DNC reading test were excluded from analysis because of their test methods such as gap-filling and true/false. It may be possible that if we examine the DNC test as a whole we may be able to find more various types of questions. This should be examined in future research by employing different means of categorization from the present study.

While it may deviate from the intent of this pilot study, it would be valuable to refer to the next generation iBT (Internet-based Testing) TOEFL. The next generation iBT TOEFL has been in use since September 2005 in the U.S., and it has been decided that it will also be introduced into Japan in May 2006. Following some revision, a number of considerable changes were made to this test, such as the introduction of speaking section, the measurement of integrated skills (e.g., read a passage, listen to a related passage, and then speak about how the ideas in the two passages are related) and so forth. Furthermore, the
new test in reading would include new kinds of questions on “rhetorical purpose, paraphrasing, overall inferences, and overall organization of ideas, in addition to the more familiar questions on main ideas, details, vocabulary, and reference” (Phillips, 2005, p. 29). Interestingly, these categories of question types closely match the categories that the present study used. The present study revealed that the old version of the paper and pencil TOEFL had already included questions which asked for paraphrasing or explicit information (Paraphrase Questions), inferences (Inference Questions), main ideas (Thematic Questions), reference (Referential Questions), vocabulary (Vocabulary Questions), and organization of ideas (Text Organization Questions), but this tendency may be further enhanced in the next generation iBT TOEFL by stating it clearly. Also, it must be interesting for us to know what questions can be seen as having a rhetorical purpose.

6.3 Main Study

6.3.1 Method

6.3.1.1 Participants

The purpose of this study was to examine whether the data consistent with Hypothesis 8 would be empirically provided or the different data from Hypothesis 8 would be provided. Hypothesis 8 was that scores of the questions which measure lower-level processes correlate low or moderately with the scores of the questions which measure higher-level processes.

A total of 204 students (i.e., 201 undergraduate students and three graduate students) from two universities in Japan volunteered to participate in this research. However, because of flaws in materials, the data obtained from four students were finally excluded from analysis, bringing the final total number of students whose data were analyzed to 200. To
avoid being biased in favor of students in a particular major, the data was collected from 14 different academic fields. Table 6.6 indicates the number of participants by field.

<table>
<thead>
<tr>
<th>Academic Field</th>
<th>Number of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Biology</td>
<td>68</td>
</tr>
<tr>
<td>2. Human Studies</td>
<td>33</td>
</tr>
<tr>
<td>3. Psychology</td>
<td>24</td>
</tr>
<tr>
<td>4. International Studies</td>
<td>21</td>
</tr>
<tr>
<td>5. Information Sciences</td>
<td>10</td>
</tr>
<tr>
<td>6. Comparative Culture</td>
<td>9</td>
</tr>
<tr>
<td>7. Medical Science</td>
<td>7</td>
</tr>
<tr>
<td>8. Sociology</td>
<td>6</td>
</tr>
<tr>
<td>9. Art</td>
<td>6</td>
</tr>
<tr>
<td>10. Engineering</td>
<td>6</td>
</tr>
<tr>
<td>11. English Language Education</td>
<td>3</td>
</tr>
<tr>
<td>12. Health and Sport Sciences</td>
<td>2</td>
</tr>
<tr>
<td>13. Japanese Language Education</td>
<td>2</td>
</tr>
<tr>
<td>14. Natural Science</td>
<td>1</td>
</tr>
<tr>
<td>15. Others</td>
<td>2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>200</strong></td>
</tr>
</tbody>
</table>

*Note. These majors are arranged from the largest group to the smallest group.*

Biology was the largest group, with dominating more than a quarter of the total number. Human studies majors, psychology majors, international studies majors, and information sciences majors followed biology majors. No student was a double major.

In terms of their L2 proficiency, participants had taken English classes at junior school and at senior high school for at least six years. Some students had studied at university in English speaking countries for one year.
6.3.1.2 Materials

Among the six types of questions described in the pilot study, “Paraphrase Questions” are labelled lower-level processing questions, whereas “Inference Questions,” “Thematic Questions,” and “Text Organization Questions” are labelled higher-level processing questions in the current research. They were grouped as the following procedures, based on Grabe’s (2000) classification (for more information, see section 2.1): (a) Answering “Paraphrase Questions” is closely related to syntactic parsing and propositional integration; therefore, this type can be considered the lower-level processing, (b) answering “Inference Questions” is closely related to the construction of situation model, and inference making has been classified into higher-level processing by many researchers (e.g., Grabe; Horiba, 1993, 1996), and (c) finally, considering that in order to find correct answers to “Thematic Questions” and “Text Organization Questions,” we need to “work[ing] with larger units of information” (Grabe, p. 233), these processing should be classified as higher-level processing as well.

Basically, a part of TOEFL and the STEP test, which were examined in Pilot Study, were administered in this main study. TOEFL was chosen, on the one hand, because it contained many types of questions which measure examinees’ higher-level processing as well as questions which measure examinees’ lower-level processing. These higher-level processing questions on TOEFL were to become a criterion when I conducted a correlation analysis between the scores of higher-level processing questions and those of lower-level processing questions. The STEP test was chosen, on the other hand, because most of the questions were classified in the lower-level processing questions (“Paraphrase Questions”). It then became possible to run two correlation analyses, that is, the correlation between the higher-level processing questions on TOEFL and the lower-level processing questions on
TOEFL, and the correlation between the higher-level processing questions on TOEFL and the lower-level processing questions on the STEP test. Table 6.7 shows sources from which test materials were selected, the number of items, and test methods.

Table 6.7

*Reading Tests Used in Study 4*

<table>
<thead>
<tr>
<th>Test (Source)</th>
<th>Part (Number of Items and Test Methods)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. STEP Test (2nd grade test and pre-1st grade test administered in June 2004; Society for Testing English Proficiency, n.d.; see Appendix 6-A)</td>
<td>4A and 4C on the 2nd grade test, and Section 3 on the pre-1st grade test (11 four-option, multiple-choice questions)</td>
</tr>
<tr>
<td>2. TOEFL (ETS, 1999, pp. 30-36; see Appendix 6-B)</td>
<td>The first three passages in Section 3 on the Practice Test A (34 four-option, multiple-choice questions)</td>
</tr>
</tbody>
</table>

*Note.* ETS = Educational Testing Service.

With regard to the STEP test, three questions on the pre-1st grade test were newly categorized by the two raters, who made judgments in the Pilot Study. Their judgments matched perfectly, that is, all three questions were considered to be “Paraphrase Questions.” As a result, out of 11 items, 10 items were classified as “Paraphrase Questions,” and one item was classified as an “Inference Question.” With regard to TOEFL, question types had already been allotted to all of the items in Pilot Study. In sum, the higher-level processing questions accounted for 10 items (30%), and the lower-level processing questions accounted for 12 items (34%) among a total of 34 items from TOEFL (see Table 6.8).
Table 6.8

Administered Questions on the STEP Test and TOEFL in Study 4

<table>
<thead>
<tr>
<th>Question Type</th>
<th>Number of Items</th>
<th>Proportion (%)</th>
<th>Item Number [Section]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>STEP Test</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Paraphrase Questions</td>
<td>10</td>
<td>91</td>
<td>2nd Grade: 35 [4A], 36 [4A], 41 [4C], 42 [4C], 43 [4C], 44 [4C], 45 [4C]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Pre-1st Grade: 35 [3], 36 [3], 37 [3]</td>
</tr>
<tr>
<td>2. Inference Questions*</td>
<td>1</td>
<td>9</td>
<td>2nd Grade: 34 [4A]</td>
</tr>
<tr>
<td>3. Thematic Questions*</td>
<td>0</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>4. Referential Questions</td>
<td>0</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>5. Vocabulary Questions</td>
<td>0</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>6. Text Organization Questions*</td>
<td>0</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td><strong>Higher-Level Processes</strong></td>
<td>1</td>
<td>9</td>
<td>2nd Grade: 34 [4A]</td>
</tr>
<tr>
<td>(2 + 3 + 6)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>11</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td><strong>TOEFL</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Paraphrase Questions</td>
<td>12</td>
<td>34</td>
<td>1 [1], 4 [1], 7 [1], 12 [2], 15 [2], 23 [3], 27 [3], 28 [3], 29 [3], 31 [3], 32 [3], 33 [3]</td>
</tr>
<tr>
<td>2. Inference Questions*</td>
<td>4</td>
<td>12</td>
<td>5 [1], 13 [2], 18 [2], 21 [2]</td>
</tr>
<tr>
<td>3. Thematic Questions*</td>
<td>2</td>
<td>6</td>
<td>11 [2], 22 [3]</td>
</tr>
<tr>
<td>4. Referential Questions</td>
<td>4</td>
<td>12</td>
<td>3 [1], 8 [1], 14 [2], 26 [3]</td>
</tr>
<tr>
<td>5. Vocabulary Questions</td>
<td>8</td>
<td>24</td>
<td>2 [1], 6 [1], 16 [2], 17 [2], 19 [2], 24 [3], 25 [3], 30 [3]</td>
</tr>
<tr>
<td>6. Text Organization Questions*</td>
<td>4</td>
<td>12</td>
<td>9 [1], 10 [1], 20 [2], 34 [3]</td>
</tr>
<tr>
<td><strong>Higher-Level Processes</strong></td>
<td>10</td>
<td>30</td>
<td>5 [1], 13 [2], 18 [2], 21 [2], 11 [2], 22 [3], 9 [1], 10 [1], 20 [2], 34 [3]</td>
</tr>
<tr>
<td>(2 + 3 + 6)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>34</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Categories with an asterisk indicate the higher-level processing.

6.3.1.3 Data Collection Procedures and Data Analyses

These tests were administered during or after school hours. Every participant took TOEFL for 33 minutes, then the STEP test for 25 minutes. The order of these tests was not
counterbalanced since all of these procedures were completed within an hour. In contrast, regarding TOEFL, the order of reading passages was counterbalanced. This was because it was predicted that a number of students might leave many questions unmarked due to a lack of time. Therefore, the following three versions of booklets were made: (a) A total of 63 students received a booklet 1, in which passage 1, passage 2, and passage 3 were presented in this order, (b) a total of 68 students received a booklet 2, in which passage 2, passage 3, and passage 1 were presented in this order, and (c) a total of 69 students received booklet 3, in which passage 3, passage 1, and passage 2 were presented in this order. However, it should be noted that the examinees were not obliged to follow these sequences when they answered questions.

The participants were instructed that they could not take notes while answering TOEFL; but that they could not during answering the STEP test, which followed the original TOEFL and the original STEP test. This research continued for three months from January 19 to April 19, 2005. With regard to data analysis, the Pearson product-moment correlation analysis was employed. The statistical software package SPSS for Windows (Version 10.0E) and spreadsheet software developed by Aoki (2003) were used for data entry and analysis.

6.3.2 Results and Discussion

Validation of the instruments used for measurement was examined from the external aspect and the structural aspect on the basis of Messick’s (1989, 1996) framework. These aspects were selected in this case for the same reasons as those mentioned for Study 1 (see section 3.3.1). With regard to the external aspect, the Pearson product-moment correlation coefficient was computed between the scores on the STEP test and the scores on TOEFL. A
moderate correlation, \( r = .592 \) (\( p < .01 \)), was provided. This coefficient is nearly the same as that reported in Ushiro et al. (in progress), \( r = .52. \) Although there are some differences in the numbers of items (34 items on TOEFL and 11 items on the STEP test in the current study, and 50 items on TOEFL and 12 items on the STEP test in Ushiro et al.’s study), and the number of participants (200 learners in the current study and 114 learners in Ushiro et al.’s study) between these studies, the correlation coefficients of these studies between the scores on TOEFL and the scores on the STEP test were roughly matched. Therefore, this moderate correlation is to be reasonable. In sum, in terms of the validation of the instruments from the external aspect, relatively positive evidence was provided.

With regard to the STEP test containing 11 items, on the one hand, the Cronbach’s alpha coefficient was .635. It was neither too high nor too low. “Correlation magnitude is partly a function of sample size and ability range” (Henning, 1987, p. 53), and I speculated about this test from this perspective, firstly by using the Spearman-Brown formula. As a result, in order to attain the Cronbach’s alpha of .700, the number of items should have been 1.3 times as many as that of the current research. In other words, at least 15 similar items were required to be beyond \( \alpha = .700 \). Furthermore, point biserial correlation coefficients ranged from .184 to .426 (see Table 6.9). On the Henning’s (1987) criteria, eight items were acceptable with indicating .250 or more than .250, and three items were unacceptable, indicating less than .250. If I followed these criteria strictly, I should have dropped the three items with lower point biserial correlation coefficients; however, considering the present aim to examine the relationship between higher-processing questions and lower-processing questions, the content validity should be given priority. In addition, the mean of the point biserial correlation coefficients was over .250 (\( r_{pb} = .299 \)). After considering these factors, it was decided that all of the items on the STEP test would be entered for further analyses.
Table 6.9

Point Biserial Correlation Coefficients of STEP Test Used in Study 4

<table>
<thead>
<tr>
<th>Item Number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>Mean</th>
</tr>
</thead>
</table>

Note: $r_{pb}$ = point biserial correlation coefficients.

Table 6.10

Point Biserial Correlation Coefficients of TOEFL Used in Study 4

| Item Number | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 |
|-------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| Item Number | 31 | 32 | 33 | 34 | Mean |
| $r_{pb}$    | .304 | .266 | .205 | .311 | .271 |

Note: $r_{pb}$ = point biserial correlation coefficients.

With regard to TOEFL, on the other hand, Cronbach’s alpha was .778, which was higher than that of the STEP test; and as to point biserial correlations, 22 items were .250 and above, and 12 items were under .250. When the point biserial correlations of all of the items were averaged, $r_{pb} = .271$ was obtained as the mean coefficient. For the same reasons as were mentioned about the STEP test above, it was determined that all of the items on TOEFL should be analyzed. Thus, in terms of the STEP test and TOEFL, relatively positive evidence was provided for validity from the structural aspect as well as from the external aspect.

Descriptive statistics are shown in Tables 6.11 for the STEP test and Table 6.12 for TOEFL. With regard to the STEP test, the mean percentage score for a total was .667 ($SD = 0.205$), and the mean percentage score for “Paraphrase Questions” was .645 ($SD = 0.215$).
Table 6.11

Means and SDs of STEP Test Used in Study 4

<table>
<thead>
<tr>
<th>Question Type</th>
<th>Number of Items</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paraphrase Questions</td>
<td>10</td>
<td>.645</td>
<td>0.215</td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>.667</td>
<td>0.205</td>
</tr>
</tbody>
</table>

Note. Percentage scores were computed by dividing the scores by a total number of items in each category. Full score = 1.000.

Table 6.12

Means and SDs of TOEFL by Question Type Used in Study 4

<table>
<thead>
<tr>
<th>Question Type</th>
<th>Number of Item</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Paraphrase Questions</td>
<td>12</td>
<td>.554</td>
<td>0.207</td>
</tr>
<tr>
<td>2. Inference Questions*</td>
<td>4</td>
<td>.423</td>
<td>0.281</td>
</tr>
<tr>
<td>3. Thematic Questions*</td>
<td>2</td>
<td>.548</td>
<td>0.357</td>
</tr>
<tr>
<td>4. Referential Questions</td>
<td>4</td>
<td>.714</td>
<td>0.252</td>
</tr>
<tr>
<td>5. Vocabulary Questions</td>
<td>8</td>
<td>.509</td>
<td>0.225</td>
</tr>
<tr>
<td>6. Text Organization Questions*</td>
<td>4</td>
<td>.459</td>
<td>0.267</td>
</tr>
<tr>
<td>Higher-Level Processing Questions (2 + 3 + 6)</td>
<td>10</td>
<td>.462</td>
<td>0.208</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>.534</td>
<td>0.284</td>
</tr>
</tbody>
</table>

Note. Categories with an asterisk indicate the higher-level processing. Percentage scores were computed by dividing the scores by a total number of items in each category. Full score = 1.000.

With regard to TOEFL, the following statistics were shown: (a) The mean percentage score for a total was .534 ($SD = 0.284$), (b) that for “Paraphrase Questions” was .554 ($SD = 0.207$), and (c) that for higher-level processing questions was .462 ($SD = 0.208$).

The Pearson product-moment correlations among the following variables were computed: (a) “Paraphrase Questions” on TOEFL (containing 12 items), (b) “Paraphrase Questions” on the STEP test (containing 10 items), and (c) higher-level processing questions on TOEFL (containing 10 items).
Table 6.13

Correlation Coefficients Among “Paraphrase Questions” on TOEFL, “Paraphrase Questions” on STEP Test, and Higher-Level Processing Questions on TOEFL

<table>
<thead>
<tr>
<th>Measure</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Paraphrase Questions on TOEFL</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>2. Paraphrase Questions on STEP</td>
<td>.464 (.348-.566)</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>3. Higher-Level Processing Questions on TOEFL</td>
<td>.465 (.349-.567)</td>
<td>.479 (.364-.579)</td>
<td>---</td>
</tr>
</tbody>
</table>

Note. N = 200.

All of the coefficients were significant at \( p < .001 \).
The values in parentheses indicate lower and upper bounds of 95% confidence interval.

As Table 6.13 reports, all of the coefficients were moderate. Let us focus on the relationships between the higher-level processing questions and lower-level processing questions (i.e., “Paraphrase Questions”) here. Within TOEFL, \( r = .465 \ (p < .001) \) was obtained between the scores of these different levels of questions. And, between the scores of higher-level processing questions on TOEFL and those of “Paraphrase Questions” on the STEP test, \( r = .479 \ (p < .001) \) was computed. Considering these results, Hypothesis 8 that scores of the questions which measure lower-level processes correlate low or moderately with the scores of the questions which measure the higher-level processes was tenable. It should also be noted that positive evidence in support of Hypothesis 8 was consistently shown in spite of the tests.

I will review Ikeno (2002) again more in detail. His study showed that significant correlations were found between “a text structure prediction test” and “a L2 lexical semantic judgment test,” \( r = .412 \ (p < .05) \), and between “a text structure prediction test” and “a L2 sentence verification test,” \( r = .397 \ (p < .05) \). There was no significant correlation between the other pairs. Taking both the results of the current study and Ikeno’s study into consideration, higher-level processing might correlate, at most, moderately with lower-level
Furthermore, the correlational analyses of the current research indicate the following: (a) The scores of “Paraphrase Questions” on TOEFL account for about 22% ($r^2 = .216$) of the variance of the scores of higher-level processing questions on TOEFL, and (b) the scores on “Paraphrase Questions” of the STEP test account for about 23% ($r^2 = .229$) of the variance of the scores of higher-level processing questions on TOEFL. However still, it should be noted that 77% to 78% of the variance of the higher-level processing questions could not be explained by the lower-level processing questions. If you remember the reading questions on the STEP 2nd grade test administered on June 2004, which has few higher-level processing questions (see Table 6.5), these small proportions should not be overlooked.

These results have implications for test developers who are interested in the examinees’ reading ability, which consists of multiple levels of reading processes. It is desirable that following types of questions are contained as well as questions related to lower-levels processing such as what the present study labelled as “Paraphrase Questions”: (a) a question which measures ability to elicit a correct response based on their inferences, (b) a question which measures ability to understand text organization, and (c) a question which measures ability to understand main ideas of a paragraph or a reading passage. Furthermore, when we consider the readers’ processes of constructing mental representations, in which they integrate text information with their background knowledge and have coherent mental representations for a sequence of events, actions and states (Gernsbacher, Robertson, Palladino, & Werner, 2004; Horiba, 1996, 2001; Kintsch, 1988, 1994, 1998; Muramoto, 2000; van Dijk & Kintsch, 1983; Zwaan & Brown, 1996), to set higher level processing questions on reading tests leads to assess a part of examinees’ mental representations.
Finally, the author acknowledges that further research is necessary because an unexpected consequence has been obtained in the present study at the same time. The correlation coefficient between “Paraphrase Questions” on the STEP test and “Paraphrase Questions” on TOEFL was moderate, \( r = .464 \) \( (p < .001) \), although they were classified into the same question type. It may imply that in addition to a question type, other variables should be studied in the future.

### 6.4 Summary

The present study aimed to accomplish two goals; to find (a) whether, if we adopt the viewpoint of reading processes, some well-known English reading tests include questions which ask for what can be inferred, questions which ask for a theme of a passage, and other types of questions, and (b) what relationship exists between the questions which focus on lower-level processing and those which focus on higher-level processing (see section 2.1).

Firstly, questions were categorized into either one of the following six classes: “Paraphrase Questions,” “Inference Questions,” “Thematic Questions,” “Referential Questions,” “Vocabulary Questions,” and “Text Organization Questions.” As a result of these classifications, TOEFL was found to be comprised of various types of questions, whereas most of the questions of the STEP test were grouped into “Paraphrase Questions”; furthermore, the DNC test revealed a character that could place it in the middle of the continuum from TOEFL to the STEP test.

Secondly, when we labelled “Paraphrase Questions” as lower-level processing questions, and “Inference Questions,” “Thematic Questions” and “Text Organization Questions” as higher-level processing questions altogether, moderate correlation coefficients were obtained between these different levels of questions. Considering these results, that is,
that the one level processing questions account for only around 20% of the variance of the other level processing questions, it is desirable that a reading examination should include both higher-level processing questions and lower-level processing questions.

Endnotes

1. This research was conducted under a grant from Society for Testing English Proficiency. An earlier version of this chapter will be published in Shimizu (in press).

2. Test specifications are useful to understand the objective of the test, theoretical framework underlying the test, test tasks, and the other important issues about the test. However, in general, it is difficult for an outsider to attain test specifications because “detailed descriptions of tests are frequently considered proprietary information” (Alderson, 2000, p. 125). Without exception, I was not able to find test specifications of the STEP test, the DNC test, and the paper and pencil TOEFL; therefore, I substituted this information for test specifications in the present study.

3. The use of this term is different from the traditional one in language testing. It does not mean “the relationship between the resources that will be required in the design, development, and use of the test and the resources that will be available for these activities” (Bachman & Palmer, 1996, p. 36); rather, it seems to express that their interest is to select English items in terms of how useful the expression is.

4. Listening section is not included in these 75 minutes.

5. This is the value which is not corrected for attenuation. If it is corrected for attenuation, it increases to $r = .74$ (Ushiro et al., in progress).

6. Z transformation was applied for the computation of means.
7. Conclusion

7.1 Overview of Findings

The present study is concerned with the following two areas of L2 acquisition research, that is, the development of L2 reading comprehension theory, and an exploration of effective reading tasks. In the following paragraphs, I will begin by discussing the theory development, and then, reading questions. Answers to hypotheses will not be restated in this section, but the results will be summarized in general terms.

First, the current research provides a foundation for an understanding of EFL learners’ mental representations constructed during or after reading texts. When we focused on their inference types, lower L2 proficiency learners were found to be weak at constructing textual coherence, in particular, within a few sentences. It can be concluded that what accords with teachers’ intuition was validated to a large degree by research (see Chapter 3).

Second, we can see the possibility that some inferences are generated differently between L1 and L2 readers. This discussion is concerned with the timing which generates inferences; that is, “on-line,” which is “during the course of comprehension,” or “off-line,” which is “during a later retrieval task but not during comprehension” (Graesser et al., 1994, p. 371). “Causal antecedent” inferences (i.e., the inference is on a causal chain between the current explicit action, event, or state and the previous passage context), which is targeted in Study 2 (see Chapter 4), are well known as on-line inferences in L1 reading research. However, the data of the present study did not show that L2 learners, even those with higher language proficiency (whose L2 reading proficiency was high enough to be allowed to enter universities in English speaking countries such as the U.S. and Canada) did not always generate the inferences on-line; rather, it was the off-line condition that they generated much of the inferences in and that revealed the difference in amounts or the generation processes
between higher and lower L2 reading proficiency learners. Thus, it is confirmed by the present study that the predictions based on L1 reading research about timing of inference generation cannot always be applied directly to L2 reading comprehension. Therefore, much more research of L2 reading comprehension in this area is necessary.

Third, Study 4 (see Chapter 6) reported that the correlation between higher-level processing and lower-level processing was moderate \((r = .465 \text{ to } .479)\). We cannot disregard these coefficients, because this indicates that lower-level processing can only account for about 22% to 23% of the variance of the higher-level processing, and vice versa. Like many other researchers, the author acknowledges that a process cannot be clearly divided into two groups, higher-processes and lower-processes; instead, each process can be placed at a certain point in a continuum between the two extremes. However, when we follow the quite agreeable classification among researchers, such as in Study 4, their differences can be revealed.

Fourth, “why” questions were found to function as guided questions in enabling EFL learners to achieve deeper comprehension. That is, when given such questions, they inferred more accurately the causes of some events in the text (for example, the punch lines of humorous stories), and moreover, they understood more correctly not only the information overtly targeted by the questions, but information only slightly related to the questions as well. What kinds of questions improve learners’ comprehension, and how those questions have positive effects on learners’ comprehension is still a central interest of many researchers and ESL or EFL teachers. This study provides them with one pedagogical implication of questioning (see Chapter 4).

Fifth, one of the answers as to how reading questions should be arranged, or how teachers should give a series of questions to their students in the classroom, was suggested
by the present study. With regard to the question sequence in which learners start at a question designed to tap some details written in a certain paragraph, and gradually come to questions designed to tap a main theme of an entire passage, this sequence which is frequently used in class was found to be particularly acceptable for poor readers. In contrast, good readers showed the tendency that they benefit more from the exact opposite sequence, that is, that during the early stages they are given a question asking for a main theme of an entire passage. From these, it can be concluded that there is an interaction between learners’ L2 reading proficiency and effective question sequences (see Chapter 5).

7.2 Limitations, and Suggestions for Future Research

In this section limitations of this research and directions for future research will be mentioned. Please also note that for each study, these will be discussed one by one.

With regard to Study 1 conducted in order to examine Japanese EFL learners’ mental representations in reading, I need to refer to the texts used for immediate written protocols. These were different in length and readability, producing more factors as a result than the researcher initially expected. For further research, carefully controlling reading materials is necessary. This will almost certainly clarify the interpretation of experimental results.

With regard to Study 2, which focused on the inference generation processes of Japanese EFL learners, with particular emphasis on their causal antecedent inference generation, I can point to a direction for future research in that of methodologies for data collection. Study 2, employing a think-aloud procedure, revealed that in the case of EFL learners’ reading, their L2 reading proficiency was more closely related to the inferences measured off-line than on-line, and that the causal antecedent inferences were more likely to be generated off-line. This differentiates the EFL learners’ processes from L1 readers’
processes. This finding, however, needs to be confirmed by future research by using other on-line measures such as a measurement of response time latency.

With regard to Study 3 conducted to look for an effective question sequence while acknowledging the levels of learners’ L2 reading proficiency, two points should be addressed to be considered as limitations of the present research and as suggestions for further research. The first point regards the relatively low reliability of both the items in the L2 reading proficiency test taken from the STEP tests, and items in the test used for observing the effect of a question sequence, which were taken from the tests of the National Center for University Entrance Examinations. Because of the lower reliability of the L2 reading proficiency test, I could not help reducing the 10 items to only four to judge the students’ L2 reading proficiency levels. And as with the tests of the National Center for University Entrance Examinations, I had to abandon aims to examine the data separately by reading passage; in other words, I could not help combining all of the questions (i.e., 11 items) attached to two passages together, and analyzing the scores of 11 items in total. If the reliability was higher, we would be able to examine the data of Passage 1 and Passage 2 separately.

The second point concerning Study 3 is a suggestion for further research. If I do not limit my foci to inference questions only, but if I shift my foci to all of the types of reading questions, I would be able to more easily find a set of questions for this design. For example, some questions ask for details explicitly written in texts, other questions ask for information elicited by integrating some sentences, and furthermore, the remainders ask for a theme of a passage. These systematically arranged questions might possibly be in some reading textbooks for EFL or ESL learners. Furthermore, these questions are likely to more frequently appear in expository passages.
With regard to Study 4, which was conducted to examine what types of test items are contained in EFL or ESL reading tests and what relations we can find between higher-level processing questions and lower-level processing questions, there are three concerns to be discussed in this section. The first point is that further research should be pursued to explain the following finding: The scores of the questions taken from TOEFL practice tests and classified as paraphrase questions, correlated only moderately with those of the questions taken from the tests of the Society for Testing English Proficiency and classified as paraphrase questions. It was expected that these scores would have a high correlation because the questions were of the same type. In further research, we need to control some conditions as much as possible, for example, the readability of a passage, the time frame for one question, permission or prevention of notetaking, and other controllable factors.

The second point addressed here about Study 4, I selected paraphrase questions as one of the lower-level processing questions. However, as is listed by Grabe (2000; see section 2.1), there are many other lower-level processes such as orthographic processing and phonological coding. In the future, we should encourage research whose focus is on a variety of lower-level processes.

The third point concerning Study 4 is about the assumptions of the study. As was mentioned in section 6.1, Study 4 was conducted based on the following assumptions: (a) If we obtain a higher correlation coefficient between these scores, the claim that test developers should make higher-level processing questions as well would be weakened, in contrast, (b) if we obtain a lower correlation coefficient between these scores, the claim would be strengthened. These assumptions may not be acceptable for some researchers, especially those who consider that the content validity, to be specific, the match between construct and test items, is extremely important in constructing a test. In other words, they
are probably going to argue for making higher-level processing questions even after they have received a high correlation between higher-level processing questions and lower-level processing questions. Although I acknowledge that there are still some conflicting viewpoints in the field of test development, the results of the current study offer evidence for the first position.

Finally, from an overall perspective the author would like to deal with an issue common to Studies 1, 2, 3, and 4. Among these, three of them targeted Japanese EFL learners at the university level (Studies 1, 2, and 4), and one targeted learners at the high school level (Study 3). Considering that some results may be closely related to learners’ grades (e.g., van den Broek et al., 2001), much further research is necessary in this area including the other levels of learners such as junior high school students and so forth.

7.3 Methodological and Pedagogical Implications

This study suggests several implications for methodology for data collection, and language pedagogy, which includes reading instructions in the EFL classroom and language testing. Some may have already been mentioned in the Discussion sections in Chapters 3, 4, 5, and 6, but I will repeat them in this section.

The first methodological implication concerns van den Broek et al.’s (1993) model, which was used in Study 1. I could conclude that it functioned well as a framework for observing EFL learners’ mental representations, moreover, for a comparison of mental representations constructed by different levels of L2 proficiency learners. This offers one possible direction for future research. For example, although the present study targeted learners of relatively higher L2 proficiency, university students in Japan, who had learned English for six and a half years and more on average, this framework may be applicable to
examining mental representations of lower L2 proficiency learners as well. Furthermore, it can be useful for not only cross-sectional studies such as the present study but also longitudinal studies in the future. How EFL learners develop their reading ability is one of the central interests for L2 reading researchers.

The second methodological implication regards on-line and off-line data collection methods. It seems that some reading processes are difficult to observe if we employ off-line methods only. For example, in Study 1, which adopted an off-line procedure, immediate written protocols, there was a relatively strong relationship between readers’ L2 proficiency and making connecting inferences (i.e., bridging inferences). In contrast, such a relation disappeared when I used an on-line procedure, think-aloud protocols, in Study 2, but was observed again with the participants’ protocols retrieved by questioning examined. To sum up, division between on-line and off-line makes us understand more profoundly EFL learners’ reading processes.

The remaining implications are all related to language pedagogy. The first pedagogical implication is for instructors. It is important to encourage EFL learners to make inferences during and after reading, if we expect them to construct rich mental representations. In particular, learners of lower L2 proficiency appear to be poor at making coherence of an argument written in a text. This process includes finding antecedents of pronouns, substitutes for proverbs, and implicit connections which are not marked by any special grammatical units. Let us look at some good examples of reading tasks from Pearson and Johnson (1978, as cited in Temma, 1989, p. 148).

■ What do you think the following underlined words or phrases indicate? (translation mine)
John and his cousin went to the fair last week. *They* had a great time *there*. First, they took the roller coaster. It was really fast. John got sick. So *did* his cousin. Then they went to see the gorilla. What a crazy animal! First *it* threw peanuts at the crowd. Then *it* pounded *its* chest…

This task requires students to identify the antecedents of pronouns such as “They,” “It,” and “its,” substitutes for proverbs such as “did,” and the places which adverbs such as “there” refer to. Another example shown below concerns making a student recognize implicit connections in a text. This task comes from Temma (1989, p. 152), and the answer keys are indicated with asterisks added.

- Read the following sentences. Which do you think appropriate for them, “and” or “but”? (translation mine)

  1. His sister gave him a watch for his birthday, (and*/ but) his brother gave her a book for her birthday.
  2. His sister gave him a watch for his birthday, (and/ but*) her brother gave her nothing for her birthday.
  3. The farmers grow food, (and*/ but) the people in the cities eat it…

Because these complete sentences (or clauses) are explicitly marked with “and” or “but,” you may think these tasks are not related to readers’ inferences, especially to those which are for constructing local coherence. However, this type of task enables students to direct their attention to the relationships between sentences (or clauses). I can, therefore, expect that this type of task can be a starting point for eventually making a good reader who can infer sentence connections even when sentences are not connected by explicit markers.

Although there may be many techniques to improve these inference generation
processes in EFL learners, the effectiveness of questioning was shown by the present research as a place to start. To be more specific, the question types on which the present research focused can be summarized, by following Ikeno’s (2000) classification, as follows: (a) inference questions, (b) posed during the course of or after reading, (c) in a short-answer format, (d) using learners’ L1 (Japanese in this case) for both questioning and answering, (e) presented orally, and (f) posed by a teacher.

One more point we should note is that “why” questions—“why do you think the event in the last sentence occurred?”—improved not only their inference generation but also their performance on understanding the literal meaning of some related parts in a text. In sum, this question had indirect effects. In addition to “why” questions, Graesser et al. (1994, pp. 389-390) propose “how” questions and “what-happened-next” questions. According to Graesser et al., “how” questions can draw inferences of “subordinate goals-actions” or “causal antecedents events.” For example, if teachers give a question such as “how did the character attain the treasure?” students would answer with the procedures through which the character came to obtain the treasure (i.e., this answer refers to “subordinate goals-actions” or “causal antecedents events”). Furthermore, “what-happened-next” questions can draw inferences of “causal consequences.” For example, if teachers give a question such as “what would happen after this accident?” students would be predicting the following story plots (i.e., this answer refers to “causal consequences”). These seem to be quite obvious; however, it must be helpful for instructors to keep it in mind that there are some connections between the questions and inference types which would be elicited effectively by each question. This is the second pedagogical implication.

The third pedagogical implication is also for reading instructors. When instructors give a series of reading questions to their students in the classroom, they should carefully devise
their sequence while considering their vocabulary and grammatical knowledge, and the cognitive load on their students. A small step sequence, in which a set of questions are given from those which ask for details to those which ask for a main theme of a passage, can be less demanding than the opposite sequence. This sequence also seems to be particularly effective for poor readers. The poor readers in the present study were high school students, who had taken formal English language education for at least three and a half years in Japan, but failed to score above chance levels (25% in the present case) on a reading section of the STEP 2nd grade test (see Table 5.10). This implication is parallel to the programmed instruction proposed by Skinner (1968). Furthermore, this implication is based on the viewpoint of “aptitude-treatment interaction” (Cronbach & Snow, 1977; Kitao, 1991), which is defined as “the relationship between a learner’s personal strengths and weaknesses in learning and the learning situation, including the type of programme one is enrolled in” (Richards & Schmidt, 2002, s.v. aptitude-treatment interaction).

The last pedagogical implication concerns language testing. Even widely used language tests such as the STEP tests administered in Japan have only a small number of reading questions measuring examinees’ higher-level processing. Thus, I recommend that questions which ask for text structures, main themes of paragraphs or an entire passage, and inferences, should be included in addition to questions which ask for information explicitly written in a text. This seems to be a promising approach in order to more accurately assess EFL learners’ reading ability.

It is hoped by the author that the present study contributes to the future theoretical research with regard to reading processes, and EFL reading instruction in classroom; moreover, it is believed that the present research could yield some insights into language test development as well.
References


Gavens, N., & Barrouillet, P. (2004). Delays of retention, processing efficiency, and


Japan Association of College English Teachers (JACET) Basic Words Revision Committee (Ed.). (2003). *JACET list of 8000 basic words*. Tokyo: JACET.


Appendixes

Appendix 3-A

L2 Proficiency Test Used in Study 1

Choices with asterisks indicate answer keys.

英語基礎力テスト
Ⅰ 語彙 1点×20問 = 20点

1. The light was too (  ) for me to read the magazine.
   (l) vague  (2) steady  (3) thin  (4) dim*

2. Anne lost (  ) of herself and began to cry.
   (l) balance  (2) confidence  (3) control*  (4) temper

3. The computer room is kept at a constant (  ).
   (l) climate  (2) temperature*  (3) circumstance  (4) atmosphere

4. The (  ) for the new branch office is not determined yet.
   (l) estate  (2) alley  (3) ground  (4) site*

5. Our parents took care of us, and now it’s our (  ) to take care of them.
   (l) attempt  (2) charge  (3) turn*  (4) order

6. Under the new government proposal, heavy taxes will be (  ) on luxury goods.
   (l) disposed  (2) exposed  (3) imposed*  (4) transposed

7. The (  ) drinking age varies from state to state in the United States.
   (l) adult  (2) advanced  (3) liable  (4) legal*

8. My sister usually takes a shower first and then eats breakfast, but this morning she (  ) the order.
   (l) resumed  (2) returned  (3) revolved  (4) reversed*

9. One version of this video comes with subtitles and the other has been (  ) into Japanese.
   (l) twisted  (2) shifted  (3) dubbed*  (4) framed

10. The restaurant had a (  ) range of main dishes on its menu.
    (l) deep  (2) spacious  (3) central  (4) wide*

11. Peter’s unique performance on the street (  ) the attention of all passersby.
    (l) drew*  (2) paid  (3) caused  (4) released

12. The president’s autobiography (  ) half of its pages to his hard and penniless school days.
    (l) devotes*  (2) serves  (3) divides  (4) contains

13. Please do not (  ) to contact me if you need any further information.
    (l) worry  (2) mind  (3) hesitate*  (4) pause

14. I was so late for the party that I decided to take a taxi. To my surprise there weren’t any taxis (  ).
    (l) usable  (2) available*  (3) capable  (4) suitable

15. The police department (  ) that the number of violent crimes will increase by about 5% this year.
    (l) estimates*  (2) evaluates  (3) rates  (4) values

16. A: This soup is delicious, Jenny. You must give me the (  ).
    B: Sure. I’d be glad to.
    (l) recipe*  (2) script  (3) method  (4) manual

17. The old church needs to be (  ) because it’s one of the most important historical buildings in the city.
18. Bob’s been studying German for eight years but still isn’t as (   ) as he wants to be.
   (1) mild   (2) verbal   (3) fluent*   (4) plain

19. Susan’s income has dropped since she changed jobs, so she’s had to give up eating out to reduce her living (   ).
   (1) expenses* (2) amounts   (3) prices   (4) charges

20. Although Claire looks young for her age, she is actually one of the most (   ) people I know.
   (1) prime   (2) mature*   (3) ancient   (4) childish

Ⅱ 語法・文法 1点×20問 = 20点

1. May I (   ) this jacket to see if it fits?
   (1) try on*   (2) take on    (3) pull on  (4) wear on

2. Let’s stop at a cash machine. We’re running (   ) of money.
   (1) aware   (2) apart     (3) short*  (4) stock

3. Jane walked out of the house without (   ) a word to her mother.
   (1) as little as  (2) better than  (3) less than  (4) so much as*

4. I ran to the station as fast as I could (   ) to miss the train.
   (1) as     (2) only*     (3) until   (4) before

5. I was surprised to find that (   ) students in my class went to the rock concert.
   (1) all of    (2) almost   (3) so many* (4) most of

6. When (   ) noodles in Japan, don’t worry about making a noise.
   (1) eaten    (2) to eat     (3) eating*  (4) one eats

7. I received a letter (   ) I was told to pay additional taxes.
   (1) of that    (2) in which*   (3) by what  (4) on where

8. I will see to (   ) that everything is ready for the upcoming conference.
   (1) it*     (2) one      (3) which  (4) this

9. People in small towns often go (   ) their way to be kind to visitors.
   (1) out of*   (2) into      (3) along with (4) throughout

10. These are my (   ) shoes, which I bought in Italy.
    (1) new leather walking* (2) new walking leather (3) walking new leather
    (4) leather walking new

11. A: “Come and see me whenever (   ). I’ll be in my office all afternoon tomorrow.”
    B: “Thank you very much, Mr. Lansky.”
    (1) you are convenient  (2) you will be convenient  (3) it is convenient of you
    (4) it is convenient for you*

12. Because of her three small children, Nancy must keep things like knives and medicine safely out of
    their (   ).
    (1) position   (2) care   (3) touch   (4) reach*

13. Amy couldn’t make herself (   ) because of the noise of the approaching train.
    (1) hear (2) to be heard   (3) heard*  (4) hearing

14. (   ) Mike’s plan to go to Italy after graduation, I asked him if he would go to England with me.
    (1) Not knowing* (2) Not known   (3) Not to know (4) Neither knowing

15. Be sure to keep the receipt for your new skirt. Otherwise, you (   ) be able to exchange it.
    (1) should not   (2) might not*   (3) may hardly (4) must not

16. We have decided to put (   ) ¥20,000 a month, to save up for our trip to Europe next March.
    (1) across    (2) aside*   (3) beside (4) behind
17. Leaving clean air and water for future generations is something ( ) by all people.
   (1) to desire  (2) to desiring  (3) to be desired* (4) to have desired
18. It wasn’t ( ) Dick said that annoyed me, but the way he said it.
   (1) as (2) which (3) that (4) what*
19. Now that my father is in his fifties, he has his health ( ) every six months.
   (1) checking (2) checked* (3) be checked (4) to be checked
20. With her ankle ( ), Anne couldn’t take part in Sunday’s basketball game.
   (1) injuring (2) injured* (3) to injure (4) to be injured

Ⅲ 作文
を並べかえて、( )の中で 3 番目と 5 番目に入るものの番号のマーク欄を塗りつぶしなさい。
ただし、( )の中では文頭にくるべき語も小文字で示してあります。

1. 妹の世間知らずには、よく驚かされます。
2. はるばる手伝いに来てくれてどうもありがとう。
3. たとえどんな職業に従事するとしても、誠実でなければなりません。
4. 彼らは少なくとも 10 分前に出発したはずです。
5. 雨が突然降ってきたので、少年たちは仕方なく野球をやめました。
6. 子供たちは Jリーグの試合を見て、すっかり夢中になっていきました。
   League game.
7. この試験管の中には、このフラスコのおおよそ 5 分の 1 の量の溶液が入ります。
8. 当地での生活費は、母国に比べるとずっと安い。
9. パーティには、お一人でいられてもお友達を同伴なさっても、どちらでもかまいません。
   party.
10. 私たちの製品を改善するのにいちばん良い方法を考え直さないといけませんね。

IV 読解
次の英文を読んで、1 から 5 までの ( ) に入れるのに最も適切なものを、下記の (1)、(2)、(3)、
(4) の中から一つずつ選び、その番号のマーク欄を塗りつぶしなさい。

Teaching in Everyday Life

We all teach and learn throughout our lives. Each father and mother does an astonishing amount of
teaching in everyday life. However, we sometimes forget that the relationship between parents and
children is essentially based on teaching. Some parents may love their children so much that they try
to keep them (1) the real world. The important thing is not to treat children that way, but to explain the rules
required to live in the real world.
When they are very young, children ask thousands of questions because the world is new to them. They are curious about everything they see, and want every question to be answered. (2) difficult for parents to answer all their questions perfectly, but children do not expect their parents to know everything. They just need to be responded to properly so they can trust adults to help them understand the world. If parents don’t know the answer, it’s (3) to tell their children honestly that they don’t, and promise to try to find the answer. (4) children will likely become more frustrated, less interested in learning, and may turn away from their parents. As a result, family members may grow distant from each other.

Some parents say children are sometimes difficult to talk with because of the generation gap. That’s true in a way if parents ignore children’s viewpoints. When they talk to children about subjects that worry children, such as studying, money and love, children will open their hearts and talk without hesitation. It is also important to let children know that parents like (5) as much as children do and that the world is a place to explore.

1. (1) along (2) inside (3) out of* (4) above
2. (1) When it is (2) Never is it (3) In case it is (4) Of course it is*
3. (1) all right* (2) not true (3) a mistake (4) embarrassing
4. (1) Therefore (2) Then (3) If (4) Otherwise*
5. (1) teaching (2) learning* (3) expecting (4) viewing

News Reporting in the U.S.

Several years ago, a researcher compared the number of television reports dealing with Japan shown on U.S television with the number of reports dealing with the United States shown on Japanese television. The results were remarkable. Over a seven-month period in 1992-93, there were 1,121 reports about the U.S. on Japanese TV, but only 92 about Japan in the U.S.

It is not only reporting about Japan; there is a general lack of foreign news reports in almost all American media. Most American newspaper editors agree that unless there is direct U.S. involvement in an international story—notably an American victim—there is not much of an audience for foreign news.

When editors ranked the popularity of various sections of their newspaper—such as National, Sport, Entertainment, and Business—World ranked fifth out of the seven sections. Newsmagazines that feature foreign stories on the cover reportedly sell 20 percent fewer copies on average. A former foreign-news editor at The New York Times estimated that “At any given time, there are only about two million people in the U.S. who are really interested in foreign affairs.”

But the problem does not lie only with the readers. A report by a study group in Washington, D.C., blames the lack of international news in most U.S. newspapers on the indifference of editors themselves. The report says that editors cut international stories using the excuse that readers aren’t interested. It notes that in fact readers rarely cancel their subscriptions* because of content. The document ends by suggesting that with international events becoming increasingly important to the U.S., editors should try to do a better job of educating their readers about world events even if the public isn’t interested.

*subscription: (定期刊行物) の予約購読

6. The small amount of news reporting about Japan in the U.S. is
   (1) typical of U.S. media attitudes toward foreign news.*
   (2) typical of U.S. media attitudes toward foreign audiences.
   (3) remarkable because there was so much in 1992-93.
   (4) remarkable because there are even fewer reports about other countries.

7. American newspaper editors believe that
(1) many Americans are often victims in major accidents.
(2) many Americans are not interested in foreign news.*
(3) foreign interest in American news is becoming high.
(4) foreign stories about Americans are very rare.

8. What will most likely happen if a U.S. newsmagazine puts a foreign story on the cover?
   (1) It will sell 20% more copies than usual.
   (2) It will sell 20% as many copies as usual.
   (3) It will sell 80% as many copies as usual.*
   (4) It will sell 80% fewer copies than usual.

9. The study group recommends that editors should
   (1) teach readers to regard content as important.
   (2) let readers know the reason for cutting foreign news.
   (3) help readers realize that foreign affairs are important.*
   (4) advise readers to keep their subscriptions because of content.

10. According to the passage, which of the following is NOT true?
    (1) There seems to be little interest in international news in the U.S.
    (2) U.S involvement increases Americans’ interest in international news reports.
    (3) There seem to be more foreign news reports in Japan than America.
    (4) Americans will stop reading newspapers if the amount of news is increased.*

Reducing Stress Improves Health

Stress is a natural condition that began with the earliest humans two million years ago. When we face a threat, our nervous systems get turned on. Our blood pressure increases and adrenaline is released into our body to help us fight or run away. However, running away or fighting won’t serve us well today when we’re arguing with our boss or trying to pass a test. Stress stays in our body and can damage our hearts and veins. Research shows that people who experience high psychological stress are also more likely to get sick. To test this, some researchers put cold viruses into the noses of volunteers. The result: 74% of the volunteers with low stress in their lives caught colds, while 90% of those with high levels of stress caught not only colds, but other diseases.

It has long been known that physical factors like high blood pressure, high cholesterol and smoking can lead to heart disease. Studies have also shown a relationship between certain personality types and heart problems. In 1960, two heart specialists showed that Type “A” people, those who worked too much and slept too little, were at greater risk of heart trouble than the more relaxed Type “B” people. Recent studies indicate the most dangerous Type “A” characteristic is hostility. Hostile people are their own worst enemies. They are more likely to have fewer friends than others and so be lonely. Loneliness, research shows, is also related to health risks.

Fortunately, people can learn to be less hostile. They can learn to think less about the things that make them angry, to listen to others, to be more trusting and forgiving. They can also learn to laugh at themselves and to take themselves less seriously. By doing this, they can begin to lead a healthier, happier, stress-free life.

11. Stress used to be important for humans because
    (1) it helped us act and survive.*
    (2) it helped us sleep better at night.
    (3) it helped us ignore threats.
    (4) it helped us repair our hearts and veins.
12. Stress is not good for us today because
   (1) its effects are no longer felt by people.
   (2) it remains with us and has negative effects.*
   (3) its effects disappear before we can understand them.
   (4) it never makes us fight or run away.
13. According to the passage, researchers have found that high levels of psychological stress
   (1) decrease the flow of blood to the heart.
   (2) reduce our blood pressure to unhealthy levels.
   (3) were not usually present with people long ago.
   (4) weaken our physical ability to fight off sickness.*
14. According to the passage, recent studies indicate that
   (1) Type “A” people have few enemies.
   (2) loneliness and health are not connected.
   (3) hostility is a characteristic of Type “A” people.*
   (4) hostility can be a healthy characteristic.
15. In order to lead a stress-free life we should
   (1) not trust the advice of others.
   (2) try to be less hostile.*
   (3) take ourselves more seriously.
   (4) concentrate more on our problems.
Appendix 3-B

Reading Passages Used for Recall Tasks in Study 1

The numbers and slashes were not shown in the original passages that the participants read. A slash means a border between idea units, and the number in parentheses indicates the number of idea units.

Passage 1
Do Pigeons Take the Train?

(1) Pigeons are smart birds. /
(2) They are good at finding places. /
(3) They can find their way home from many miles away. /
(4) Scientists are not sure how pigeons know their way, (5) but they do. /
(6) They almost never get lost. /
(7) Can pigeons find their way under the ground? /
(8) Some passengers say that they see pigeons on the Underground trains. /
(9) Rachel Robson says she saw a pigeon at Paddington Station. /
(10) It got on the train / (11) and then got off again at the next stop. /
(12) Some train lines are especially popular with pigeons. /
(13) They are often seen on the Northern and Piccadilly lines. /
(14) Lorna Read also sees pigeons at the Paddington Station. /
(15) Once, a passenger tried to get a pigeon off the train. /
(16) The pigeon flew back in / (17) just before the doors closed. /
(18) It seemed all upset — like a person who doesn’t want to miss the train. /
(19) Why do pigeons get on the trains? /
(20) Some people say that the pigeons are not looking for food. /
(21) They say the pigeons want to save time. /
(22) London is a big city. /
(23) The pigeons get tired of flying / (24) and they do what people do. /
(25) They take the train. /
(26) What do the scientists think of this? /

(27) They say they must have some more information—(28)—but anything is possible with pigeons! /
(Mikulecky & Jeffries, 1998, p. 45)

Passage 2
Research on the Effects of Color and Appearance

(1) At the University of Kansas art museum, investigators tested the effects of different colored walls on two groups of visitors to an exhibit of paintings. /

(2) For the first group the room was painted white; (3) for the second, dark brown. /

(4) Movement of each group was followed by an electrical system under the carpet. /

(5) The experiment revealed that those who entered the dark brown room walked more quickly, (6) covered more area, (7) and spent less time in the room than the people in the white environment. /

(8) Dark brown stimulated more activity, (9) but the activity ended sooner. /

(10) Not only the choice of colors but also the general appearance of a room communicates and influences those inside. /

(11) Another experiment presented subjects with photographs of faces that were to be rated in terms of energy and well-being. /

(12) Three groups of subjects were used; (13) each was shown the same photos, (14) but each group was in a different kind of room. /

(15) One group was in an “ugly” room that resembled a messy storeroom. /

(16) Another group was in an average room—a nice office. /

(17) The third group was in a tastefully designed living room with carpeting and drapes. /

(18) Results showed that the subjects in the beautiful room tended to give higher ratings to the faces than did those in the ugly room. /

(19) Other studies suggest that students do better on tests taken in comfortable, attractive rooms than in ordinary-looking or ugly rooms. /

(Baudoin, Bober, Clarke, Dobson, & Silberstein, 1993, p. 16)

Passage 3
Do Animals Sleep?

(1) All animals must rest, (2) but do they really sleep as we know it? /

(3) The answer to this question seems obvious. /

(4) If an animal regularly stops its activities and stays quiet and unmoving—if it looks, as though it is sleeping—then why not simply assume that it is in fact sleeping? /

(5) But how can observers be sure that an animal is sleeping? /

(6) They can watch the animal (7) and notice whether its eyes are open or closed, (8) whether it is
active or lying quietly, (9) and whether it responds to light or sound. (10) These factors are important clues, (11) but they often are not enough.

(12) Horses and cows, for example, rarely close their eyes, (13) and fish and snakes cannot close them.

(14) Yet this does not necessarily mean that they do not sleep. (15) Have you ever seen a cat dozing with one eye partly open?

(16) Even humans have occasionally been observed to sleep with one or both eyes partially open.

(17) Animals do not necessarily lie down to sleep either.

(18) Elephants, for example, often sleep standing up, with their tusks resting in the fork of a tree.

(19) Finally, while “sleeping” animals often seem unaware of changes in the sounds and light and other stimuli around them, that does not really prove they are sleeping either.

(20) Observations of animal behavior alone cannot fully answer the question of whether or not animals sleep.

(21) The answers come from doing experiments in “sleep laboratories” using a machine called the electroencephalograph (EEG).

(22) The machine is connected to animals (23) and measures their brain signals, breathing, heartbeat, and muscle activity.

(24) The measurements are different when the animals appear to be sleeping than when they appear to be awake.

(25) Using the EEG, scientists have confirmed that all birds and mammals studied in laboratories do sleep.

(26) There is some evidence that reptiles, such as snakes and turtles, do not truly sleep, although they do have periods of rest each day, in which they are quiet and unmoving.

(27) They also have discovered that some animals, like chimpanzees, cats, and moles (who live underground), are good sleepers while others, like sheep, goats, and donkeys, are poor sleepers.

(28) Interestingly, the good sleepers are nearly all hunters with resting places that are safe from their enemies.

(29) Nearly all the poor sleepers are animals hunted by other animals; (30) they must always be watching for enemies, even when they are resting.

(Clarke, Dobson, & Silberstein, 1996, p. 79)
Appendix 3-C

Rubric of Recall Tasks

英 語

受験上の注意

1. これから3つの英語の文章を読んでもらいます。それぞれ読み終わったら、その内容に関
して覚えていることをできるだけ多く日本語で書くようにとの指示がなされます。ただし、
その際、本文を見直すことはできません。

2. 読む時間と書く時間はそれぞれ別々にとってあります。文章によって制限時間が違います
ので、指示があるまでページはめくらないようにしてください。

3. もしも制限時間内に文章を読み終わらなかった場合、どこまで読んだか分かるように、文
章中に「かぎ括弧」で印をつけておいてください。

4. 読んでいるときメモはとらないでください。

5. 例題: 次の文章を読んでください。

On the continent of Africa there is a great river called the Nile. In the north on both sides of the river,
the soil is very rich. Here Egypt was born four thousand years ago. The ancient Egyptians saw that with
boats they could make a great road of the Nile River. So they learned to make good boats. It turned out
that they were better boats than any that had been made before. We know what these boats looked like
because the Egyptians painted pictures of them on the walls of their palaces and tombs. They built
perfect tiny models of their boats, too. Some of these models have been found preserved in the tombs
they were buried in centuries ago.

（回答例）
アフリカ大陸には、大きなナイル川がある。ナイル川の両側の土は豊富である。エジプトは4000年前に誕生した。古代のエジプトの人々は、ナイル川をわたる船をたった。彼らは
船を作った。私たちが想像するようなもので、その上には絵が描かれていた。その船の完全なモ
デルも作った。

実施期日 2001年 月 日
学類 学籍番号 名前 性別（男 女）
Appendix 3-D

Point Biserial Correlation Coefficients of the STEP Test Used in Study 1

Underlined values are .250 and above. According to Henning (1987), point biserial correlation coefficients of .250 and above are acceptable.

<table>
<thead>
<tr>
<th>Item Number</th>
<th>( r_{pb} )</th>
<th>Item Number</th>
<th>( r_{pb} )</th>
<th>Item Number</th>
<th>( r_{pb} )</th>
<th>Item Number</th>
<th>( r_{pb} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.316</td>
<td>21</td>
<td>.270</td>
<td>41</td>
<td>.219</td>
<td>61</td>
<td>.240</td>
</tr>
<tr>
<td>2</td>
<td>.003</td>
<td>22</td>
<td>.224</td>
<td>42</td>
<td>.091</td>
<td>62</td>
<td>.225</td>
</tr>
<tr>
<td>3</td>
<td>-.080</td>
<td>23</td>
<td>.091</td>
<td>43</td>
<td>.379</td>
<td>63</td>
<td>.497</td>
</tr>
<tr>
<td>4</td>
<td>.119</td>
<td>24</td>
<td>.285</td>
<td>44</td>
<td>.346</td>
<td>64</td>
<td>.295</td>
</tr>
<tr>
<td>5</td>
<td>.007</td>
<td>25</td>
<td>.134</td>
<td>45</td>
<td>.497</td>
<td>65</td>
<td>.298</td>
</tr>
<tr>
<td>6</td>
<td>.281</td>
<td>26</td>
<td>.042</td>
<td>46</td>
<td>.326</td>
<td>66</td>
<td>.264</td>
</tr>
<tr>
<td>7</td>
<td>.316</td>
<td>27</td>
<td>.091</td>
<td>47</td>
<td>.170</td>
<td>67</td>
<td>.383</td>
</tr>
<tr>
<td>8</td>
<td>.176</td>
<td>28</td>
<td>.252</td>
<td>48</td>
<td>.035</td>
<td>68</td>
<td>.242</td>
</tr>
<tr>
<td>9</td>
<td>.023</td>
<td>29</td>
<td>.147</td>
<td>49</td>
<td>.075</td>
<td>69</td>
<td>.263</td>
</tr>
<tr>
<td>10</td>
<td>.095</td>
<td>30</td>
<td>.189</td>
<td>50</td>
<td>.221</td>
<td>70</td>
<td>.286</td>
</tr>
<tr>
<td>11</td>
<td>.222</td>
<td>31</td>
<td>.211</td>
<td>51</td>
<td>.346</td>
<td>71</td>
<td>.296</td>
</tr>
<tr>
<td>12</td>
<td>.040</td>
<td>32</td>
<td>.076</td>
<td>52</td>
<td>.356</td>
<td>72</td>
<td>.297</td>
</tr>
<tr>
<td>13</td>
<td>.111</td>
<td>33</td>
<td>.278</td>
<td>53</td>
<td>-.012</td>
<td>73</td>
<td>.394</td>
</tr>
<tr>
<td>14</td>
<td>.303</td>
<td>34</td>
<td>.213</td>
<td>54</td>
<td>.244</td>
<td>74</td>
<td>.347</td>
</tr>
<tr>
<td>15</td>
<td>.199</td>
<td>35</td>
<td>.120</td>
<td>55</td>
<td>.312</td>
<td>75</td>
<td>.342</td>
</tr>
<tr>
<td>16</td>
<td>.355</td>
<td>36</td>
<td>.288</td>
<td>56</td>
<td>.401</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>.192</td>
<td>37</td>
<td>.430</td>
<td>57</td>
<td>.419</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>.293</td>
<td>38</td>
<td>.278</td>
<td>58</td>
<td>.443</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>.237</td>
<td>39</td>
<td>.174</td>
<td>59</td>
<td>.322</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>.288</td>
<td>40</td>
<td>.255</td>
<td>60</td>
<td>-.130</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix 4-A

Reading Passages for the Think-Aloud Protocol

The target sentence of each passage is underlined, although it was not shown in the original passages that the participants read.

Passage 1

Some villagers were going to celebrate an important wine festival in a few days’ time. They borrowed a huge barrel from the nearest town and put it in the village square. They decided that each of them should empty a bottle of the best wine he had into it so that there should be plenty at the feast. The night of the feast arrived. Everybody gathered in the village square with their glasses for the wine. The tap on the barrel was opened. But what came out was pure water. (Tajika, 1999, p. 135)

Passage 2

Jim, twenty-one years old, got a job and started to live on his own. He asked Mrs. Roper to come and clean his apartment every morning for an hour. Two weeks later, he thought, “That mirror looks very dusty. Mrs. Roper’s forgotten to clean it.” He wrote a message: “I’m coughing whenever I breathe because everything in this room is very dusty.” At 7 o’clock that evening, he went into his bedroom. “That silly woman still hasn’t cleaned it!” he said to himself. He happened to find a bottle in front of the mirror. He read the words on the bottle. They were: “Cough medicine.” (Hill, 1985b, pp. 12-13)

Passage 3

Once there was an old fisherman who had a pet trout named Henry that he kept in the bathtub. The man was tired of changing the water of the tub quite often. So he began to take Henry out of the tub for a few minutes at a time. Soon Henry could stay out of the tub for a long time. One day, the fisherman started for town as usual with Henry coming along behind him. As the man crossed a bridge, he noticed that Henry had suddenly disappeared. The man could see through a hole in a bridge Henry was dead in the water below. He’d fallen through the hole into the stream and drowned. (Maruyama, 1993a, pp. 38-39)

Passage 4

Mr. Jones was very angry with his wife, and she was very angry with him. One evening Mr. Jones was very tired when he came back from work, so he went to bed soon after dinner. When Mrs. Jones went up to bed much later than her husband, she found a piece of paper on the small table near her bed. It said, “Mother, wake me up at 7 a.m.—Father.” When Mr. Jones woke up the next morning, it was nearly 8 a.m.—and on the small table near his bed he saw another piece of paper. He took it and read these words. “Father, wake up. It’s 7 a.m.—Mother.” (Maruyama, 1993b, p. 10)

Passage 5

Mrs. Kim loved her daughter, Joan, who left home upon her marriage. She received a lot of letters from Joan, but there was only news about her child named Kate. Mrs. Kim asked Joan about herself. But her replies were always about Kate. Mrs. Kim knew that parents always
thought that their own children were special and better than any others. So she did not mention it for a while. But then, after some time, Mrs. Kim wrote to Joan, “I am very glad that you have a very clever child, because I myself have never had one.”

(Hill, 1985a, p. 27)
Appendix 4-B

Questionnaire Used in Study 2

Numbers, which indicate the presentation sequence of passages, had been written in parentheses of the questions 1, 2, 3, 4, and 5. This is because the sequence was different among participants.

今調査に関するアンケート

どうもありがとうございました。最後に、アンケートにご協力を、よろしくお願いします。
選択肢があるものについては、該当するに○をつけてください。

名前：
学籍番号：
所属：
学年： 年
ここ３年以内に取得した英語関連の資格： 英検 級／TOEIC 点／TOEFL 点
その他（ ）点
海外での生活の有無： 西暦 年 月～ 年 月
国名

質問1 文章「ワイン祭り」(番目に読んだ文章)に書かれてあった出来事・内容について、以前、読んだり聞いたりしたことがありますか。
(ある／ない)
質問2 文章「Jimと掃除婦」(番目に読んだ文章)に書かれてあった出来事・内容について、以前、読んだり聞いたりしたことがありますか。
(ある／ない)
質問3 文章「おばけた魚」(番目に読んだ文章)に書かれてあった出来事・内容について、以前、読んだり聞いたりしたことがありましたか。
(ある／ない)
質問4 文章「静かな会話」(番目に読んだ文章)に書かれてあった出来事・内容について、以前、読んだり聞いたりしたことがありましたか。
(ある／ない)
質問5 文章「Mrs. Kimの娘への思い」(番目に読んだ文章)に書かれてあった出来事・内容について、以前、読んだり聞いたりしたことがありましたか。
(ある／ない)
質問6 読解中考えていることを口に出しながら文章を読むことは、難しかったですか。
難易度を1から5で表すとしましたら、どれくらいになりますか。
（1…とても易しい、2…やや易しい、3…どちらともいえない、4…やや難しい、5…とても難しい）
質問7 上の質問6の回答について、それはなぜだと思いますか。理由があれば記入してください。

ご協力ありがとうございます。
Appendix 4-C

Rubric of the Think-Aloud Tasks

All of the statements in this rubric were read aloud by the researcher before the think-aloud task was begun. Since the participants were given a paper with this rubric written, they were also able to read these by themselves during the experiment.

Appendix 4-C

Rubric of the Think-Aloud Tasks

All of the statements in this rubric were read aloud by the researcher before the think-aloud task was begun. Since the participants were given a paper with this rubric written, they were also able to read these by themselves during the experiment.

〈調査の目的〉
1 英語を学習している大学生が、実際どのように英語の文章を読んでいるのか、その過程を調べることが今回の調査の目的になっています。そのため、頭に浮かんだことを口に出してもらいながら、英語の文章を読んでいただきます。
2 この調査は、あなたが言ったことをもとに、あなた自身を評価したり、判断したりするものではありません。

〈発話の方法〉
1 頭に浮かぶことをすべて、内容を説明しながらそのまま日本語で口に出してください。
   文章に対しての感想や、文章の内容から連想したことなど、読解中に考えたことであれば、それらすべてを言葉に出してください。また、けっして正解を求めるわけではないので、ありましたので、分からない場合は「このところは分からないけど。」や「多分、ここは○○のことで。」などと言って進んでいただいて構いません。また、前の部分と矛盾していると思ったら、「矛盾しているみたいだ。前のところに戻ってみよう。」と言っって、戻ってもらって構いません。
2 自然に発話してください。こんなこと言ってもいいのかなど、自分で意味のあるなしを判断しないで、頭に浮かんだことをすべて報告してください。
3 口に出していたこととは、たとえ完全な文の形でなくても構いません。

〈パソコンの使用方法〉
1 文章はパソコンの画面にPowerPointで表示されます。
2 前読んだ文章に付け加えられる形で、一文ずつパソコン画面に表示されますので、一文を読み終わったら、パソコンの「↓」のキーを押してください、次の文が出てくる。
3 文章の最後には「END」の表示がありますので、終わったら、声をかけてください。
4 もし、パソコン操作が途中で分からなくなったら、いつでも声をかけてください。

〈状況説明〉
1 私（清水）もこの部屋に同席しますが、途中、あいづちをうったり、訂正したりするようなことはありません。気にせずに、自分のペースで進めてください。時間は十分かけてください。
2 発話はピンマイクを通してテープに録音させていただきますが、できるだけ気にしないように音質を良くするためにリーディングを行ってください。

〈調査の流れ〉
1 モデルを聞く
2 練習を行う
3 日本語で発話しながら文章の読みを行う
   最後に、「END」マークが出てきたら、「終わりました」と声をかけてください。
   その後、こちらからひとつだけ、日本語で質問をしますので、日本語で答えてください。そのまま思ったり、注釈を付けていても言叶をたくさんさせて構いません。
   その回答が終わったら、もう一度同じ文章で、同じように日本語で発話しながら読みを行います。（これが5つの文章について行われます）
Appendix 4-D

Passage Used for a Demonstration

Two friends were walking along a road when one of them spied a purse. “Look what I’ve found,” he cried. “You really mean, ‘Look what we’ve found,’” said his friend. Just then a rough-looking man came running toward them. “Oh-oh, this must be his purse.” said the one who found it. “We’re in trouble now.” “No,” said his friend. “You really mean ‘I’m in trouble now.’”

Appendix 4-E

Passage Used for a Practice

The hen was sick, poor thing. The fox went to cheer her up. “I’m so sorry to hear you are ill,” said the fox in his most sympathetic voice. “Thank you for your concern,” said the hen. “If you’ll simply leave me alone I’m sure I’ll be better soon.”
Appendix 4-F

Scoring Guide Used in Study 2

Target idea units have been written in bold face. A letter is attached to each idea unit, from A to W. Since idea unit C is embedded in idea unit B, idea unit C is boxed. Examples of correct answers regarding causal antecedent inferences and idea units are listed. Symbols ①, ②, ③, ④, and ⑤ indicate the number of idea units sandwiched between the parts connected by inferences.

Passage 1

Some villagers were going to celebrate an important wine festival in a few days' time. They borrowed a huge barrel from the nearest town and put it in the village square. A) They decided/ B) that each of them should empty C) a bottle of the best wine he had into it/ D) so that there should be plenty at the feast./ ① The night of the feast arrived. ② Everybody gathered in the village square with their glasses for the wine. ③ The tap on the barrel was opened. E) But what came out/ F) was pure water./

【アイディア・ユニット理解の正解例】
A) 彼ら (村人たち) は決めた (決定した)。
B) 彼らそれぞれが … をそれ (大きな樽, 樽, かりてきた樽) にあけることを。
C) 自分 (村人たち自身, それぞれ) が持っている最上の (極上の, 最高級の) 1 本のワイン。
D) 祭りのときに十分な量のワインがあるように。
E) しかしこutってきたのは (そこにあったのは)。
F) 喜に (単なる) 純粋な水 (真水) だった。

【推論の正解例】
0 点: (無回答), 災難が起こったから, ワインが盗まれたから, 魔法がかけられたから, 誰かのいったらだったから, 借りたとき, 隣の村人たちが水を入れてあったから
1 点: その村人たちが水を入れたから, その村人たちが利己的だったから, その村人たちがけちだったから, その村人たちは最上の (極上の, 最高級の) ワインを他人に渡したくなかったから, その村人たちは自分だけは水でもいいと思ったから

Passage 2

Jim, twenty-one years old, got a job and started to live on his own. He asked Mrs. Roper to come and clean his apartment every morning for an hour. Two weeks later, he thought, "That mirror looks very dusty. Mrs. Roper's forgotten to clean it." G) He wrote a message: H) "I'm coughing/ I) whenever I breathe/ J) because everything in this room is very dusty." ① At 7 o'clock that evening, he went into his bedroom. ② "That silly woman still hasn't cleaned it!" he said to himself. ③ He happened to find a bottle in front of the mirror. ④ He read the words on the bottle. K) They were: "Cough medicine."/

【アイディア・ユニット理解の正解例】
G) ジム (彼) はメッセージ (手紙) を書いた。
H) 私は咳をしている (する)。
I) 息 (呼吸) をするときはいつでも。
J) なぜなら, この部屋のすべてのものがとてもほこりっぽいから (ためだ)。
K) それら (ボトル) に書かれてあった言葉は, 咳薬 (咳止め, 咳止め薬) であった。

【推論の正解例】
0 点: (無回答や以下の答え以外で, 明らかに誤っているもの)
1 点: ジムが手紙でローパーさんに咳が出ると書いたから, ローパー夫人 (ローパーさん, 女の子, 掃除の人) がジムの書いてメッセージに怒ったから, ローパー夫人はジムのメッセージをよく理解できなかったから, ローパー夫人は仕返しをしたかったから, ローパー夫
人はいやみっぽかったから、ローパー夫人はジムのメッセージを単に咳がでているんだと勘違いしたから、ローパー夫人が、ジムのメッセージにジョークで返したから、ジムはメッセージでeverythingと書いたのでローパー夫人は鏡が汚いということに気がつかなかったため

Passage 3
Once there was an old fisherman who had a pet trout named Henry that he kept in the bathtub. The man was tired of changing the water of the tub quite often. L) So he began to take Henry out of the tub for a few minutes at a time. M) Soon Henry could stay out of the tub for a long time. ① One day, the fisherman started for town as usual with Henry coming along behind him. ② As the man crossed a bridge, ③ he noticed that Henry had suddenly disappeared. ④ The man could see through a hole in a bridge Henry was dead in the water below. N) He’d fallen through the hole into the stream and drowned.

【アイディア・ユニット理解の正解例】
L) それで、彼（漁師、フィッシャーマン）は、ヘンリーを一度に数分間、浴槽（バスタブ、タブ）から外に出し始めた。
M) 間もなく、ヘンリーは、長い間、浴槽（バスタブ、タブ）の外にいられるようになった（入ることができた）。
N) 彼（ヘンリー）は穴（橋にあたった穴）から川（小川、流れ）に落ちた（落ちてしまった、まっていた）。
O) 彼（ヘンリー）はおぼれ（おぼれてしまった、しまっていた）。

【推論の正解例】
0点：（無回答や以下の答え以外で、明らかに誤っているもの）
1点：漁師がヘンリーを水からたびたび出していたから、漁師に慣られても水中で生活できないようになっていたから

Passage 4
Mr. Jones was very angry with his wife, and she was very angry with him. One evening Mr. Jones was very tired when he came back from work, so he went to bed soon after dinner. When Mrs. Jones went up to bed much later than her husband, P) she found a piece of paper on the small table near her bed. / Q) It said, “Mother, wake me up at 7 a.m.—Father.”/ ① When Mr. Jones woke up the next morning, ② it was nearly 8 a.m.—③ and on the small table near his bed he saw another piece of paper. ④ He took it ⑤ and read these words. R) “Father, wake up./ S) It’s 7 a.m.—Mother.”/

【アイディア・ユニット理解の正解例】
P) 彼女（ミセスジョーンズ、ジョーンズ夫人）は、彼女のベッドの近くにある／のそばの小さなテーブルに、一枚の紙（紙切れ）を見つけて。
Q) それ（その紙、その手紙、その内容）には、お母さん（ママ、マザー）朝の7時に起こしてください、お父さん（パパ、ファザー）より、書いてあった。
R) お父さん（パパ、ファザー）起きなさい。
S) 朝7時にですよ。お母さん（ママ、マザー）より。

【推論の正解例】
0点：（無回答や以下の答え以外で、明らかに誤っているもの）
1点：夫が紙で書いたから、もともとヘンリーは泳げない生き物だったから、橋の穴から落ちるときに怪我をしてしまったので、奥さんは怒っていたから、夫が口でいえばいいものを紙で頼んだので仕返しをしたかったから、奥さんが意地悪をしたから、奥さんは嫌味な人だったから。

Passage 5
Mrs. Kim loved her daughter, Joan, who left home upon her marriage. She received a lot of letters from Joan, but there was only news about her child named Kate. Mrs. Kim asked Joan about herself. T) But her replies were always about Kate./ ① Mrs. Kim knew that parents always thought that their own children were special ② and better than any others. ③ So she did not
mention it for a while. U) But then, after some time, Mrs. Kim wrote to Joan, “I am very glad/ V) that you have a very clever child,/ W) because I myself have never had one.”

【アイディア・ユニット理解の正解例】
T) しかし (でも), 彼女の返事 (返答, 手紙) はいつも (常に) ケイト (自分の娘) のことについてだった。
U) しかし (でも), しばらくして (ほどなく, その後), キムさん (キム夫人, ミセスキム) はジョーン (ジョアン, ジョン, ジェニー) に, 私はとてもうれしい (幸せだ) と書いた。
V) あなた (ジョーン) がとても頭の良い (賢い) 子どもをもっていることを。
W) なぜならば, 私自身 (キムさん, キム夫人, ミセスキム) は, そのような (頭の良い, 賢い) 子どもをもったことがない (なかった) から。

【推論の正解例】
0点: (無回答や以下の答え以外で, 明らかに誤っているもの)
1点: ジョーンがケイトのことしか手紙に書いてくれないので, ジョーンが何度言っても, キム夫人の本当の気持ちを理解してくれなかったので, キム夫人はジョーンに対して怒ったから, キム夫人はいやみで書いた, キム夫人は皮肉を込めて書いた, キム夫人はこう書けばジョーンが答えてくれると思ったから
Appendix 4-G

Criteria Used for Dividing Passages Into Idea Units

These criteria were created by the author with reference to Stein and Glenn (1979), and Muramoto (1998).

<table>
<thead>
<tr>
<th>Criteria Example</th>
</tr>
</thead>
</table>
| 1. 原則として、文頭からピリオドまでの1文を1つの Idea Unit (IU; 例文 (a)) とする。ただし、基準2以下によっては、1文が複数の IU に分けられる場合もある。
(a) Judy is going to have a birthday party. (Stein & Glenn, 1979, p. 83) |
| 2. 副詞については、語句修飾 (例文 (b)) の場合、その副詞部分を独立した IU とはしないが、文修飾 (例文 (c)) の場合、その副詞部分を独立した IU とする。
(b) The children were playing happily. (Egawa, 1991, p. 133)
(c) Happily, the sea was calm. (Egawa, p. 133) |
| 3. 従属節は名詞節、形容詞節、副詞節いずれの場合も原則として1つの IU (例文 (d) と (e)) とする。ただし、名詞節の一部については、基準4および基準6が、副詞節については、基準7が、また、形容詞節のうち関係詞節については、基準8および基準9が使用される。
(d) I've got a hunch that I'll pass this test. (Egawa, p. 24)
(e) ... she will have to fix things when they break. (Stein & Glenn, p. 83) |
| 4. 主語＋believe (that) ～、主語＋know (that) ～、主語＋say (that) ～、主語＋think (that) ～などの肯定文 (例文 (f))、否定文 (例文 (g))、疑問文 (例文 (h)) については、that節の部分は単独の IU とはせず、それを含む全体を1つの IU とする。ただし、当該以外の基準によって、IU が that 節内に複数存在すると認められる場合には、それらを分割して、複数の独立した IU (例文 (i)) とする。
(f) I think that she should have paid the money back. (Summers et al., 1995, s.v. think)
(g) He did not think that girls should play with a hammer and saw. (Stein & Glenn, p. 83)
(h) Where do you think you’re going?! (Summers et al., s.v. think)
(i) Some historians think that when Mao died/ China was left rudderless. (Stern & Urbon, 2002, s.v. rudderless) |
| 5. 直接話法で、主語＋say [says/ said/ have said/ has said] “～” という場合については、“～” 内の部分は単独の IU とはせず、それを含む全体を1つの IU (例文 (j)) とする。ただし、当該以外の基準によって、IU が “～” 内に複数存在すると認められる場合には、それらを分割して、複数の独立した IU (例文 (k)) とする。
(j) John said, “Her story is true.” (Egawa, p. 469)
(k) he said, “If you can dream it/ you can do it.” (Stern & Urbon) |
sentence frames. 例えば It seems [seemed] that ～, It is [was] said that ～, It is [was] believed that ～, It is [was] thought that ～などの肯定文、否定文、疑問文については、that 節の部分は、独立した IU とはせず、それを含む全体を 1 つの IU (例文 (l) 参照) とする。ただし、当該以外の基準によって、IU が that 節内に複数存在する場合には、それらを分割して、複数の独立した IU とする。

(6) It would seem that someone left the building unlocked. / (Summers et al., s.v. seem)

(7) They decided to go together/ because neither one wanted to be left alone/ and they both liked fried chicken. / (Stein & Glenn, p. 82)

(8) The bear, who felt very lazy/ climbed upon the roof/ to watch. / (Stein & Glenn, p. 82)

(9) Then they ran very quickly to a nearby farm/ where they knew chickens lived. / (Stein & Glenn, p. 82)

(10) Judy liked the dress/ but she still wanted the hammer and the saw. / (Stein & Glenn, p. 83)

(11) The roof and the bear fell in/ killing five of the chickens. / (Stein & Glenn, p. 82)

(12) One day they decided to catch a chicken for supper. / (Stein & Glenn, p. 82)

Note: Slashes in a column of “Example” indicate the borders between idea units.
Appendix 4-H

Students’ Questionnaire Responses to Question 7 (Difficulty in Thinking Aloud)

Students’ comments were transcribed by the author directly from the questionnaires. No modifications were made to the original responses.

1. Responses by the Students who Answered Absolutely Easy ($n = 4$)

<table>
<thead>
<tr>
<th>Student</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student A</td>
<td>多分ひとりごとが多いんだと思います。</td>
</tr>
<tr>
<td>Student B</td>
<td>頭の中で考えていることを口に出すだけだから。実際に口に出すことは行わないけれど、行ってみると自分が何を考えているのかよくわかる。話を読むことに集中できる。</td>
</tr>
<tr>
<td>Student C</td>
<td>自分で説明することによりただに口に出しているし、頭の中のことをただ口で出すだけなので易しい。声に出すか出さないかの違いのみだと思った。</td>
</tr>
<tr>
<td>Student D</td>
<td>いつも家でぶつぶつ言いながら英語の長文を読むせいかかもしれない。一文一文小刻み読んでいくのは、よく考えられる上に理解もしやすいため。</td>
</tr>
</tbody>
</table>

2. Responses by the Students who Answered Quite Easy ($n = 13$)

<table>
<thead>
<tr>
<th>Student</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student E</td>
<td>けっこうリラックスしてできたから。</td>
</tr>
<tr>
<td>Student F</td>
<td>2〜3やってみると、このパターン(皮肉話)が続くかなと思い、その対比や原因探しになってしまった。話自体は短いけど、読んだことのない文章だし、丁寧に読んでみた。</td>
</tr>
<tr>
<td>Student G</td>
<td>考えていることを言うだけだから。でも、まだ考えてはいるけど口に出していない言葉があるかもしれない。</td>
</tr>
<tr>
<td>Student H</td>
<td>文章を読むとき、日本語のときでもそうですが、私はよく1人でぶつぶつあれがああで、これがどうして口に出してしまうのだろうか。だからだと思う。</td>
</tr>
<tr>
<td>Student I</td>
<td>もともとそうしているので。</td>
</tr>
<tr>
<td>Student J</td>
<td>考えていることを言葉にするとよくもしたりつっかえたりした。考えているときは一瞬でパッと考えられるが、それを言葉にするとそれなりの時間がかかるから。</td>
</tr>
<tr>
<td>Student K</td>
<td>ただ思ったことを言うだけだから、そんなに難しいとは感じない。ただ、口にしていても、全然文章が解読できなかった。</td>
</tr>
<tr>
<td>Student L</td>
<td>口に出しながら読解を行うほうが英文を読むことのみに集中できたため。</td>
</tr>
<tr>
<td>Student M</td>
<td>私の親友によく思ったことをひとり言のように口に出す子がいるので、その子を思い出すと、何とすると出来ました。面白かったです。</td>
</tr>
<tr>
<td>Student N</td>
<td>普段家でリーディングをするときも声に出していることが多かったから。でも、考えていること全てを口にしていたわけではなくなかったので、最初少しとまどった。</td>
</tr>
</tbody>
</table>
3. Responses by the Students who Answered Either Way (n = 4)

<table>
<thead>
<tr>
<th>Student</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student R</td>
<td>最初はとまどったけれども、慣れてきているようになるのが面白かったので。</td>
</tr>
</tbody>
</table>
| Student S | よく頭の中で考えていることをいつもそのまま話しているので。決めつけることは口に出さないとすぐ忘れてしまうのです。口に出すほうが覚えていられます。その分、人に言いたいことを伝えるのが難しいです。 
| Student T | もし出されることを難しいが、それによって自分の考えていることとか考えながらできてきたのがよかった。 |
| Student U | やっぱり考えていること全部を口に出すことは無意識にできていないと思うから。 |

4. Responses by the Students who Answered Quite Difficult (n = 22)

<table>
<thead>
<tr>
<th>Student</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student V</td>
<td>頭の中では、いつも考えていることを口に出すという方法だが、考えていることをすべてを言葉にするのは難しい。はっきり言葉という形にして思考しているわけではないときもあるから。</td>
</tr>
<tr>
<td>Student W</td>
<td>普段のテストや授業中は、声に出して考えたりすると他の人の迷惑になることなので、今まですっごく、そうしないようにしてきたから。</td>
</tr>
<tr>
<td>Student X</td>
<td>ふだん考えていることをすべて口に出して読むというトレーニングはしていなかったこともあったかもしれない。また、出そうとおもうとそれに意識がいく、集中できません。</td>
</tr>
<tr>
<td>Student Y</td>
<td>リーディングの際中、様々なことを思っているものの、やはりその気持ちは言葉にするとなかならず、時間もかかり難しいなと思いました。</td>
</tr>
<tr>
<td>Student Z</td>
<td>普段、文章を読むとき、考えていることを口に出さないからかな、普通に生活している時でも考えて口に出さないから。</td>
</tr>
<tr>
<td>Student AA</td>
<td>考えていることをすべて口に出すということになっていたので、考えていることをすぐに口に出すのは難しいので、後で言った。</td>
</tr>
<tr>
<td>Student AB</td>
<td>普段リーディングのプロセスを口にして読んでいくことはしないから。はじめは躊躇してしまう感じでした。</td>
</tr>
<tr>
<td>Student AC</td>
<td>普段英語を読むとき、とばして読んだり、英語を英語のまま頭で処理するのを、口に出すと全部を訳そうとしたりしてしまうから。</td>
</tr>
<tr>
<td>Student AD</td>
<td>外国語の文章の読解中は、はっきりと理解するよりも、常に、どちらの解釈か、どういう意味についていて、迷ながら読んでいることが多い。</td>
</tr>
<tr>
<td>Student AE</td>
<td>どこまで口に出せばいいのか迷う。長い文の語句の区切をここ切るとか考えているのまで言うのか？と考え結局言わなかった。私は、英語を英語→日本語で訳しながら読んでるから違うけれど、英語を英語で読んでいる人は言いようがない気がする。</td>
</tr>
<tr>
<td>-------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Student AF</td>
<td>自分自身で考えを押し殺してしまうことがあるからです。</td>
</tr>
<tr>
<td>Student AG</td>
<td>頭で考えていることが全て言葉ではなく映像で現れたりするので、そういうのを口に出して表現するのは難しかった。</td>
</tr>
<tr>
<td>Student AH</td>
<td>読んでいると疲れるから。読んでいると目に入っていて考えることが、この場合、文章に直接化されない限り、冗長という表現が難しい。</td>
</tr>
<tr>
<td>Student AI</td>
<td>日本語もそうだけど、文章は複数続いてはじめて意味をなすものだと思うから、はじめの数文はいろいろな想像をしているため、口に出すのが難しかった。</td>
</tr>
<tr>
<td>Student AJ</td>
<td>「考えていること全てを口に出す」と言われても、口に出す前にやっぱり考えてしまって、それも口に出そうするとやはり別に考えて。。。。慣れないことなので少し迷いました。</td>
</tr>
<tr>
<td>Student AK</td>
<td>頭の中には思わぬ思い浮かぶことも含めて非常にたくさんことがあるのに、それを整理した上で口に出すというのは大変な作業だから。</td>
</tr>
<tr>
<td>Student AL</td>
<td>頭の中では常に言葉を使って思考しているわけではないから。</td>
</tr>
<tr>
<td>Student AM</td>
<td>たぶんいつもも心の中で文章を読むと同時に翻訳しているからだと思う。</td>
</tr>
<tr>
<td>Student AN</td>
<td>頭の中でも考えることは、文章になっているとは限らず、イメージだったりするので、言語化するのが難しいし、同時に多くのことを考えていることがあるから。</td>
</tr>
<tr>
<td>Student AO</td>
<td>普段ぶつぶつ言いながら英語の問題をとくとまりも、ややひやかような視線でみられることがあるから。</td>
</tr>
<tr>
<td>Student AP</td>
<td>頭には考えがうかんでいるけれど、実際にそれを言葉で表現するのが難しかったからだと思う。</td>
</tr>
<tr>
<td>Student AQ</td>
<td>頭で整理しきれていない、もしくは頭の中で考えた全ての語を口に出して言うということが難しい。全部出すのは無理でした。</td>
</tr>
</tbody>
</table>

5. Responses by the Students who Answered Absolutely Difficult (n = 5)

<table>
<thead>
<tr>
<th>Student</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student AR</td>
<td>普段読むときは、口に出さないし、口に出して読むとまわりの人や相手に迷惑がかかるから。</td>
</tr>
<tr>
<td>Student AS</td>
<td>普段口に出して読んでいないため、思ったことをそのまま口に出すということも初めてのことだったので、なかなか慣れてませんでした。</td>
</tr>
<tr>
<td>Student AT</td>
<td>いつもは口に出して考えないから。</td>
</tr>
<tr>
<td>Student AU</td>
<td>頭で読んでいるときは内容をつかまることに集中しているため、読解中で考えていることについてあまり気にならないから。</td>
</tr>
<tr>
<td>Student AV</td>
<td>いつも頭の中で考えたままに整理してから（整理してから）書いたりしゃべったりしているので、浮かんできた一語一語を口に出す、というのはやったことがなく慣れていないからだと思った。でも、これができれば自然に英語が話せるようになることにつながると思う。</td>
</tr>
</tbody>
</table>
Appendix 4-I

Point Biserial Correlation Coefficients of the STEP Test Used in Study 2

Underlined values are .250 and above. According to Henning (1987), point biserial correlation coefficients of .250 and above are acceptable.

<table>
<thead>
<tr>
<th>Item Number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_{pb}$</td>
<td>.448</td>
<td>.261</td>
<td>.075</td>
<td>.392</td>
<td>.176</td>
</tr>
<tr>
<td>Item</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>$r_{pb}$</td>
<td>.267</td>
<td>.508</td>
<td>.351</td>
<td>.418</td>
<td>.317</td>
</tr>
<tr>
<td>Item</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>$r_{pb}$</td>
<td>.402</td>
<td>.141</td>
<td>-.036</td>
<td>.305</td>
<td>.295</td>
</tr>
<tr>
<td>Item</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>$r_{pb}$</td>
<td>.598</td>
<td>.451</td>
<td>.482</td>
<td>.341</td>
<td>.282</td>
</tr>
<tr>
<td>Item</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
<td>25</td>
</tr>
<tr>
<td>$r_{pb}$</td>
<td>.205</td>
<td>.546</td>
<td>.527</td>
<td>.448</td>
<td>.442</td>
</tr>
<tr>
<td>Item</td>
<td>26</td>
<td>27</td>
<td>28</td>
<td>29</td>
<td>30</td>
</tr>
<tr>
<td>$r_{pb}$</td>
<td>.401</td>
<td>.504</td>
<td>.605</td>
<td>.452</td>
<td>.388</td>
</tr>
</tbody>
</table>

Note. $N = 37$
Appendix 5-A

Questionnaire Administered in Pilot Study 2 of Study 3

Participants read this questionnaire in Japanese, and were required to answer in Japanese. The English translations were made by the author.

■ Answer the following questions freely.
(a) As to English tests of National Center for University Entrance Examinations:
   (a-1) how do you feel about reading time?
   (a-2) how do you feel about answering time?
   (a-3) were the instructions before the tests clear?
   (a-4) were the instructions during the tests clear?

   (b) As to the Society for Testing English Proficiency tests, how do you feel about reading time and answering time?

   (c) If you have any comments write them here freely.
Appendix 5-B

English Tests of the National Center for University Entrance Examinations Used in Study 3

Question 6 in Passage 2 was constructed by the author as a result of conducting a pilot study. The other questions were taken from the tests previously administered by the National Center for University Entrance Examinations. Choices with asterisks indicate answer keys.

Passage 1 (The Fiscal 2003 Test)

A “When are you going back to Brazil?”
   I looked down at the ground, trying to stop the tears. I didn’t want to cry in front of those two girls. Why did they have to say things like that?

B “If only I could go back to Brazil,” I thought. Memories of my early years filled my head. Playing, swimming, dancing, music, laughter... When my parents decided to move to Japan, though, I had to leave all that behind. I was only seven years old at the time, but I still remember that day.
   “Elena, we’re going to Japan.”
   “For a holiday?” I knew I had some distant relatives in Japan, although I’d never met them. It would be fun to fly in an airplane and visit them.
   “No, to live. We’re going to work there, and you’ll be able to go to a new school and learn Japanese. Then you’ll be able to go to a university there, and if you can speak two languages fluently when you grow up, you’ll be able to get a good job in the future.”
   “I don’t want to go to a new school. I like my school here. And what about my friends?”
   “Don’t worry, Elena. You’ll make new friends.”

C I didn’t want new friends. I wanted my old friends, my grandparents, my aunts and uncles, and cousins in Brazil. But there was no arguing; it was decided and that was that. Two months later, we arrived in Japan.

D The first few months were hard because I didn’t speak Japanese. However, my teachers and classmates were kind. I soon picked up the language and made friends and spent five happy years at elementary school.

E The problems began when I moved to junior high school. My junior high school accepts students from three different elementary schools, and I found myself in a class with many people I had never met before. Although I spoke Japanese fluently and my behavior was in no way different from anyone else’s, two of my new classmates started to tease me after they heard me speaking Portuguese with my parents at the entrance ceremony. What hurt me most deeply was the question of when I was going back to Brazil. If I went back to school in Brazil, it would be so difficult to catch up with my former classmates there. Besides, all my friends were in Japan now. I’d spent half my life in Japan, and Japan was my home.

F The two girls started again, “When are you going back to Brazil?”
G I wished they would stop. Then suddenly, I heard a voice behind me, “Elena, what’s the matter?” Natsumi, Maiko, and Kaori were walking over to me. They had been my friends since my arrival in Japan. At first, we taught each other words in Japanese and Portuguese. After that, we always encouraged and helped each other.
“Oh, nothing. I’m just being asked again when I’m going back to Brazil.”

Natsumi turned angrily toward the two girls who were teasing me. “Don’t you understand? You have only one culture, but Elena has two. OK, that makes her different, but aren’t we all different?”

Maiko joined in, saying to the two girls, “You’re both different from me, but that’s not a bad thing. I’m certainly not nasty to you because of it.”

The two girls looked ashamed. After a few moments of tense silence, they walked off toward the classroom. As I saw them walking away, the anger inside me overflowed, “I hate it! I hate it!”

Natsumi put her hand on my shoulder, as if to calm me.

“Elena, they just don’t understand yet. They’ll realize eventually.”

“Natsumi’s right,” added Kaori. “I know that sometimes being Brazilian and Japanese makes things difficult for you, but it also makes you special.”

“Yes, that’s something to be proud of,” said Maiko.

My friends were right. Deep inside, I knew that they were right. I looked at them with gratitude and finally let the tears fall.

問1 (パラグラフBに関して) According to Elena’s parents, why would it be good for Elena to move to Japan?
1. She could live with her distant relatives.
2. She could leave her high school in Brazil.
3. She could have new educational opportunities.*
4. She could travel in an airplane.

問2 (パラグラフEに関して) What problems did Elena have after she entered junior high school?
1. She did not have any friends at school.
2. She could not speak Japanese very well.
3. Her teachers complained about her behavior.
4. A couple of students said unkind things to her.*

問3 (パラグラフEに関して) Why did Elena feel hurt when she was asked about going back to Brazil?
1. She wanted people to realize that Japan was her home.*
2. She did not want people to know she was Brazilian.
3. Her parents refused to go back to Brazil.
4. She was not allowed to enter school in Brazil.

問4 (パラグラフHに関して) What did Natsumi mean when she said Elena is “different”?
1. She went to three elementary schools.
2. She has a rich cultural background.*
3. She has had a difficult life.
4. She is proud of her family.

問5 (文章全体に関して) What is the main theme of this story?
1. Brazilian lifestyle
2. intercultural understanding*
3. Japanese education
4. childhood memories
Passage 2 (The Fiscal 2004 Test)

A My confidence as a swimmer started to disappear the day Angela moved to our small town. At the time, some members of the town’s swimming club, myself included, were preparing for the National Championships, which were just six months away. I had always been the best, and everyone thought that I would be chosen for the relay race. But now I had competition. There was only one place for the butterfly on the relay team, and we both wanted it.

B For two weeks it was awful. Angela was always the star. She was faster than I, and her form was better, too. I was jealous and scared. My chances of being selected were disappearing fast. My fear caused me to be unfriendly to Angela. I refused to speak to her and never said anything good about her.

C One day, however, our coach called me over and said, “Kate, I’ve got something to say to you. Your attitude is hurting your performance. I know you can change that. I’d like you to think about it.”

D When I arrived at the pool the next morning, I thought about what he had said as I was going through my warm-up. Angela and I were going to compete that morning, and only eight girls would enter the finals.

E My thoughts were interrupted when a voice said, “Nervous?” It was Angela. “I don’t like to talk before a race,” I replied coldly.

F “I get nervous, too,” she said. Her voice didn’t have the anger of mine, which surprised me quite a bit.

G Angela and I competed in our separate trial races. I jumped into the pool and swam like a flying fish. My mind was clear, and I could think about only one thing: swimming well.

H When the races were over, the judges announced that both of us were among the lucky eight who would be competing in the finals. Despite this good news, I noticed that Angela was sitting sadly alone. This puzzled me, but I thought that I knew how she felt and tried to be friendly to her.

I “I don’t talk before races, but I do talk after them. Sometimes it helps,” I said.

J Angela was silent for a while, but then she said, “I’m great in practices, but in competitions I just can’t do well. It’s like this all the time. I’m so worried about the finals.”

K Now I felt really bad. I realized how horrible I had been to Angela. I wanted to help her. I wanted to show her that I was sorry for my behavior of the past two weeks.

L “Listen, I have an idea,” I said. “Why don’t we help each other prepare for the final race? We have two weeks to work on things.”

M “Good idea,” said Angela.

N For the next two weeks Angela and I worked together. I taught her how to deal with stress and how to train her muscles. She helped me with my form, and at the end of those two weeks we were the best of friends and respected each other as swimmers.

O The day of the final race came and when the starter pistol was fired, I swam off like a dolphin. I thought about nothing but winning, but just before I reached the finish line, I thought of Angela and looked over into her lane.

P As I was climbing out of the pool, I said to myself, “Oh no, what have I done?” I thought that I had dropped behind Angela and lost the race when I looked in her direction.

Q While we were waiting for the official announcement, the coach came running over to us. “Congratulations, girls! It was close, but Angela has won, and so have you, Kate!”

R “What do you mean?” I asked.

S “Well, Angela will be in the relay, but you swam so fast, Kate, that you, too, have won.”

T “Really?” I screamed. “I don’t understand.”

U “I’ll explain. The 21st Century Swimsuit Company is giving both of you their Future Swimmers Scholarship this year.” “Wow, I suppose that I’m a double winner: I got a scholarship and made a
friend, too.”
“You’re not the only one,” Angela remarked.

問1 (パラグラフAに関して) How did Kate feel after Angela joined the swimming club?
1. Happy.
2. Excited.
3. Sad.
4. Anxious.*

問2 (パラグラフCに関して) Why did the coach talk to Kate before the trial races?
1. He thought she should be independent.
2. He wanted to encourage her to swim better.*
3. He wanted her to be afraid of Angela.
4. He thought she should support Angela.

問3 (パ拉グラフIに関して) Why did Kate want to help Angela?
1. She wanted Angela to win.
2. The coach insisted that she do so.
3. Angela was a newcomer.
4. She understood how Angela felt.*

問4 (パラグラフLに関して) Why was Kate disappointed immediately after the final race?
1. She thought she would not be going to the Nationals.*
2. She thought Angela had lost.
3. She had decided to give up swimming.
4. She found that Angela was not her friend.

問5 (パラグラフMに関して) What did Angela mean when she said, “You’re not the only one”?
1. She knew that both she and Kate had won the race.
2. She believed that only Kate would get a scholarship.
3. She thought that she had made a friend, too.*
4. She guessed that she would be given more than one scholarship.

問6 (文章全体に関して) What is the main theme of this story?
1. swimming
2. friendship*
3. a rival
4. competition
Appendix 5-C

Society for Testing English Proficiency 2nd Grade Test Used in Study 3

Choices with asterisks indicate answer keys.

次の英文 [A] を読み、(1) から (5) までの質問に対して最も適切なものを 1, 2, 3, 4 の中から一つ選び、その番号を解答用紙の所定欄にマークしなさい。

[A] A New Kind of Tax

While most people would like to see their taxes reduced, there is one kind of tax that wins broad support from people in many countries. This is the so-called “green,” or environmentally conscious tax which targets polluters. Taxed on cars, lawn mowers, and certain kinds of fuel are examples of green taxes.

With more people commuting long distances to work by car, air pollution has become a serious concern in most major cities of the world. Air and noise pollution affect everyone’s quality of life, and voters around the world are increasingly urging their politicians to develop programs to reduce pollution. Many government programs designed to combat pollution, however, are paid for by everyone’s taxes, not by the taxes of those responsible for generating pollution. For example, America has focused its efforts on creating laws that force companies to reduce the pollution made by factories and products. One problem with these laws, however, is that they can be very expensive to put into effect. Green taxes, on the other hand, are an increasingly popular means of battling pollution because they actually generate money that governments can then use to fight pollution.

Countries like Denmark and Norway have successfully used high green taxes to discourage polluters. A tax on highly polluting fuel in Sweden encouraged most people to switch to a less-polluting type of fuel, decreasing the amount of poisonous gases in the air. By making polluting fuels more expensive, green taxes make clean fuels more attractive and thus reduce pollution. Furthermore, only those who continue to use highly polluting fuels pay the tax. Green taxes have proven to be both effective and fair, and we can expect more governments to adopt them in the worldwide struggle against pollution.

(1) According to the passage, how do green taxes differ from other taxes?
   1. Many people are in favor of them.*
   2. Green taxes target everyone equally.
   3. Green taxes are used to help polluters.
   4. People hope to have their green taxes reduced.

(2) According to the passage, how have voters around the world reacted to increased levels of pollution?
   1. They have moved out of cities to improve their quality of life.
   2. They have demanded that governments act to reduce pollution.*
   3. They have decided to drive less and use public transportation.
   4. They have criticized governments for increasing taxes on polluters.

(3) What is one problem with the American approach to pollution?
   1. It targets any companies unfairly.
   2. It isn’t very effective in the long run.
   3. It is very costly to put into effect.*
   4. It isn’t as popular as green taxes.

(4) Green taxes help to lower levels of pollution because
   1. they are paid only by those who produce fuels.
   2. they are considered fair, and so more people pay taxes.
   3. they have no effects on the amount of harmful gasses in the air.
   4. they increases the attractiveness of fuels which are less-polluting.*
(5) Which of the following statements best summarizes the passage?
1. Taxes in America are higher than taxes in Europe.
2. Taxes on all kinds of fuels are called green taxes.
3. Taxes which target polluters are becoming more popular.*
4. Green taxes, though effective and fair, are not popular yet.

次の英文[B]を読み、(6)から(10)までの(  )に入れるのに最も適切なものを1、2、3、4の中から一つ選び、その番号を解答用紙の所定欄にマークしなさい。

[B] Home Schooling in America
One of the changing aspects of American society is the (6) trend of home education. A government survey conducted in 1990-1991 shows that between 250,000 and 300,000 children were taught at home. Another study made by a private foundation indicates that the current figure may be as high as one million. The number of parent-taught children is increasing at 15-20% annually, and it will continue to increase as more parents become dissatisfied with public school education and (7) to teach their children at home.
Home education is legal in all American states, and each state sets its own laws concerning home education. In some states, parents need only to inform the city or town school board of their (8) to teach their children at home.
On the other hand, there are people who are (9) home schooling. They claim that children who are taught at home have little or no contact with other children of the same age. They fear (10), that children will not be given a chance to learn the basic social skills which will be important when they become adults.
Generally speaking, however, children who are taught by their parents have frequently scored above national averages on achievement tests. Indeed, hundreds of colleges and universities around the country have accepted such students.

(6) 1. growing* 2. moving 3. holding 4. sinking
(7) 1. accept 2. choose* 3. wonder 4. refuse
(8) 1. manner 2. dislike 3. attitude 4. intention*
(9) 1. educated at 2. interested in 3. opposed to* 4. supported by
(10) 1. therefore* 2. however 3. otherwise 4. instead
Appendix 5-D

Questionnaire Administered After Tests in Study 3

The English translations were made by the author. Participants read this questionnaire in Japanese.

■ Have you ever read either of these two passes? Circle yes or no.

Passage 1 (Elena’s story, adopted from the fiscal 2003 English test of the National Center for University Entrance Examinations): (yes  no )

Passage 2 (Story of Kate and Angela, adopted from the fiscal 2004 English test of the National Center for University Entrance Examinations): (yes  no )
Appendix 5-E

Rubric Given to Administrators in Study 3

(3種類の冊子の配布。鉛筆またはシャープペン、消しゴムの確認。)

今日の流れ
「今日は2種類のリーディングテストの過去問題を解いてもらいます。ひとつは大学入試センター試験で、もうひとつは英語検定試験です。」

配布物の確認
「配布した冊子は合計3種類です。センター試験は、ホチキス止めされているもので『問題冊子』と『解答冊子』があります。右上に『問題冊子』、『解答冊子』とありますので確認してください。英検は、問題用紙に解答用紙がはさんであります。こちらは解答用紙がマークシートになっています。」

センター試験の説明
「それではまず、センター試験からはじめます。少し説明がありますので、『問題冊子』の表紙を見てください。」

「試験時間は約31分で、2つの読解問題、11問です。」

「注意事項は、
1. 試験開始の合図があるまでは、この『問題冊子』および別冊の『解答冊子』は開かないでください。
2. 2つの冊子に名前を記入してください。
3. 解答はすべて『解答冊子』に直接記入してください。
4. 文章を読む時間と質問に解答する時間は別々にとられています。指示によってペースをめくり、解答してください。なお、後に戻って解答することはできません。そのつど指示しますので覚える必要はありませんが、以下のような手順になります。

① 開始の合図とともにこの『問題冊子』を開き、1つ目の文章を7分間で読みます。
② 『文章1の読みをやめてください』の合図とともにもうひとつの『解答冊子』の1ページ目を開け、間1の質問に解答してください。解答は直接、記入してください。文章をみながら解答してください。なお解答時間は1分30秒とされております。
③ そして『問2へ進んでください』の合図とともに『解答冊子』の次のページをめくり問2に解答してください。解答時間は同じく1分30秒です。
④ 以下同じです。『間3へ進んでください』の合図とともに『解答冊子』のページをさらにめくり間3に解答してください。解答時間は同じく1分30秒です。
⑤ これで問題の数だけ繰り返します。文章1には間が5つ、ついています。
⑥ 2つ目の文章についても上記を繰り返します。文章2は間が6つ、ついています。
5. 読む文章には、アルファベットが付されています。これらは解答するときに必要ですが、読むときは気にしないでください。

6. 『解答冊子』は試験終了後に回収します。」

「何か質問はありませんか。」

センター試験の開始：文章1
「それでは、センター試験からはじめます。英検はその間、はじめておいてください。はじめに、『問題冊子』にある文章1を7分間で読みます。『やめ』の合図があるまで読み続けてください。それでは、『問題冊子』の1ページ目を開いてください。読み始めてください。」
(時間測定を開始。)

(7分経過した時点。)
「文章1の読みをやめてください。それでは『解答冊子』をひらいて問1に解答してください。文章をみてから
解いてください。」

(あらためて時間測定を開始。1分30秒経過した時点。)
「ページをめくり、間2へ進んでください。」

(3分経過した時点。)
「間3へ進んでください。」

(4分30秒経過した時点。)
「間4へ進んでください。」

(6分経過した時点。)
「間5へ進んでください。」

(7分30秒経過した時点。)
「解答をやめてください。」

センター試験：文章2
「それでは、2つ目の文章に移ります。『問題冊子』の2ページ目を開いてください。『解答冊子』はそのまま5
ページ目を開いたままにしてください。文章を読みはじめてください」

(あらためて時間測定を開始。7分経過した時点。)
「読みをやめてください。それでは『解答冊子』の6ページをひらいて間1に解答してください。」

(あらためて時間測定を開始。1分30秒経過した時点。)
「ページをめくり、間2へ進んでください。」

(3分経過した時点。)
「間3へ進んでください。」

(4分30秒経過した時点。)
「間4へ進んでください。」

(6分経過した時点。)
「間5へ進んでください。」

(7分30秒経過した時点。)
「間6へ進んでください。」

(9分経過した時点。)
「解答をやめてください。」
「これで、センター試験の過去問題は終わりです。」

「なお、『解答冊子』の裏にアンケートがありますので、それに答えてください。終わったら『解答冊子』を集
めます。」

(『解答冊子』を回収。)

英検の説明
「英検」の問題用紙の中に解答用紙が含まれていますので、取り出してください。今回のマークシートです。
英検の開始
(準備ができた時点。)
「それでは、解答をはじめてください。」

(あらためて時間測定を開始。8分経過した時点。)
「残り5分です。」

(続けて12分経過した時点。)
「残り1分です。」

(続けて13分経過した時点。)
「解答をやめてください。解答用紙のみ集めます。」

(『解答用紙』を回収。)
「これですべて終わりです。」

正解の配布
(正解を配布。)
Appendix 6-A

Test of Society for Testing English Proficiency in Study 4

Item numbers in this appendix are original ones. When the participants answered these questions, numbers had been changed to serial numbers. Choices with asterisks indicate answer keys.

Section 4A on the 2nd Grade Test

To: Basil Black <basil-32fds@cordline.net>
From: Mary Finnegan <finnegan-m@sunnyhols.co.uk>
Date: June 13, 2004
Subject: Re: Vacation at the Park Hotel

Dear Mr. Black,
I was sorry to hear that you were not satisfied with your recent vacation to Portugal with Sunny Tours. You mentioned in your e-mail that your flight left London five hours late because of poor weather conditions in the south of England. While I understand that this must have been frustrating, I'm afraid we cannot offer a cash refund for the time you lost, as you requested. You'll find that our brochure clearly states that Sunny Tours does not compensate customers for airport delays.

I appreciate your comments about the Park Hotel. Feedback from customers about the hotels we deal with is always useful. I telephoned Mr. Figo, the manager of the Park Hotel, to pass on your concerns about their catering. He told me that their regular chef was on vacation during the period of your stay and a less experienced chef was on duty. That may explain the poor quality of the meals you had there. We will be sending one of our staff to the Park Hotel next month. If we find that the service provided there does not meet our normal high standards, we will remove the hotel from our brochure.

I hope the experiences you had on this occasion will not stop you from traveling with Sunny Tours again in the future.

Yours sincerely,

Mary Finnegan
Customer Relations Manager
Sunny Tours

(34) What is one reason that Mr. Black was not satisfied?
1. He could not stay at the Park Hotel.
2. His flight from England was delayed.*
3. The refund he received was too small.
4. The weather was bad in Portugal.

(35) Mary Finnegan
1. got an e–mail from the hotel’s regular chef.
2. apologized for the regular chef’s lack of experience.
3. told the hotel manager about Mr. Black’s complaints.*
4. asked Mr. Black for some feedback about his vacation.

(36) Next month,
1. Sunny Tours will check the service at the hotel.*
2. Mr. Black will book another vacation at the same hotel.
3. Mr. Figo will return from his vacation in Portugal.
4. Mary Finnegan will send a brochure to Mr. Black.
Section 4C on the 2nd Grade Test

Slugging to Work

Overcrowding on highways in the United States has long made driving to and from work a slow and frustrating experience for many Americans. About 30 years ago, government agencies in the United States introduced high-occupancy vehicle (HOV) lanes on some highways to try and reduce traffic levels. These lanes were reserved for cars with three or more people. Because fewer cars used HOV lanes, traffic in those lanes moved faster than in the other lanes. The government’s intention was to reduce the amount of gasoline being used by having friends and co-workers travel together to work.

Soon after HOV lanes were set up on some highways between Washington, D.C., and its suburbs, however, something unexpected began to happen. Single drivers started picking up strangers at bus stops in order to make use of the faster lanes. This practice, which became known as slugging, gradually grew in popularity. There are now a number of different routes with specific pick-up and drop-off points where “slugs” can to seen waiting patiently in line for a ride. Slugging is not actively promoted by the government, but some individuals have set up websites that give details of slug routes and invoke people to get involved.

Over time, a number of rules have developed that are designed to make slugging safer and more pleasant. On arrival at the pick-up point, the driver first calls out his or her destination rather than asking where the slug wants to go. Slugs can refuse a ride if they are suspicious of the driver or if taking the ride would mean leaving a woman waiting alone. In the car, there should be no talking unless the driver begins the conversation, and no money should be exchanged. At the end of the ride, the driver and the slugs all say “thank you.”

Everybody benefits from slugging. Drivers who pick up slugs are able to get to work faster. And by not taking their cars into town, slugs save money on fuel and parking. Slugging also reduces the number of cars on the road. So, as many commuters in Washington, D.C., now know, taking a slug to work is a great way to help improve the environment.

(41) In the United States about 30 years ago,
1. driving to work was not a slow and frustrating experience.
2. lanes were created for cars carrying at least three people.*
3. government agencies wanted more cars to use the highways.
4. government agencies built new highways to reduce traffic levels.

(42) What happened soon after HOV lanes were introduced?
1. Drivers began giving rides to people waiting at bus stops.*
2. Websites were developed to give information about bus routes.
3. The government set up a number of pick-up points for slugging.
4. More people began commuting to work by bus instead of by car.

(43) According to the rules of slugging,
1. drivers may charge for the ride if the slug does not say “thank you.”
2. slugs should first tell the driver where they want to go.
3. slugs should not refuse a ride if the driver is a woman.
4. conversations should only be started by the driver.*

(44) The practice of slugging
1. means that drivers now spend less on parking.
2. saves people money and helps the environment. *
3. allows drivers to travel longer distances to work.
4. has enabled the government to build more bus stops.

(45) Which of the following statements is true?
1. HOV lanes were not actively promoted by the government.
2. HOV lanes were introduced to make highways in the U.S. safer.
3. Drivers who pick up slugs are able to reach their destinations more quickly.*
4 Slugging was started by individuals who no longer wanted to work in Washington, D.C.

Section 3 on the Pre-1st Grade Test

Plants as Powerful Water Pumps

One of the riddles of science is how large trees create enough pressure to lift hundreds of gallons of water from their roots to the tips of their leaves. Scientists know that water moves through a circulatory system of vessels, called xylem, that run the length of the tree. But how that water is moved still has scientists puzzled.

A few facts are known, however. First, plants require more water to survive than animals because a staggering 90 percent of plant water evaporates into the air from pores in the leaves. There are only two ways to move that much water: it can either be pulled from above or pushed from below. Scientists have long favored the pull theory, which rests on water’s chemical structure. Each water molecule has a positively charged and negatively charged pole, so water molecules attract each other, each pole bonding to the oppositely charged pole of another molecule. According to the pull theory, these bonds are strong enough for water molecules evaporating from the leaves to actually tug on the water molecules that are in the leaves—a tug that extends all the way down to the roots. The heat of the sun, as the cause of evaporation, sets the process in motion.

Not all scientists accept the pull theory. Botanist Martin Canny of Australian National University believes the water is pushed upwards. The roots of plants and trees build up pressure when the underground section of xylem gets water from the soil. This causes enough pressure, Canny believes, to overcome the downward force of gravity and any slowing caused by friction. “That is the so-called ‘pressure step’ that the root pump must overcome to put water in the bottom of the pipes,” Canny explains. “Once it is there, no further lifting is needed.” Critics of the push theory argue that root pressures powerful enough to do this have never been observed in trees. Indeed, many types of trees seem to have no root pressure at all.

Whatever the source of pressure, recent research indicates that tree size is limited by the power required to carry water to the treetop. Resistance caused by friction increases as the length of the xylem increases, causing a reduction in water pressure in higher tree elevations. Cavitation, or the formation of air bubbles in vessels, happens more frequently as pressure decreases, and since cavitation breaks water-molecule bonds, it has a disastrous effect on water transportation through the xylem.

(35) On what scientific fact does the pull theory rest?
1 Water does not evaporate easily, so it is not subjected to friction in the xylem.
2 Water molecules adhere to each other, so they naturally pull each other through the tree.*
3 Water molecules do not strongly bond together, so pulling them up the tree requires an external source of energy.
4 Water molecules flow more freely when neither negative nor positive charges are present.

(36) According to the pull theory, where does the power that moves the water through a plant originate?
1 In the many pores of the leaves.
2 In the roots of the plant.
3 In the xylem throughout the plant.
4 In the warming rays of the sun.*

(37) Scientists who dispute the push theory argue that
1 root pressures sufficient to overcome the “pressure step” have not been found.*
2 it has been impossible to measure the pressure created in the roots of any tree.
3 the theory only works for smaller plants, where little pressure is needed.
4 the vessels are simply too narrow to allow water to be pushed through them.
Appendix 6-B

Sample of Section 3, TOEFL Practice Test A (Passages 1-3)

Choices with asterisks indicate answer keys. The passages read by participants included line numbers.

Passage 1

The growth of cities, the construction of hundreds of new factories, and the spread of railroads in the United States before 1850 had increased the need for better illumination. But the lighting in American homes had improved very little over that of ancient times. Through the colonial period, homes were lit with tallow candles or with a lamp of the kind used in ancient Rome—a dish of fish oil or other animal or vegetable oil in which a twisted rag served as a wick. Some people used lard, but they had to heat charcoal underneath to keep it soft and burnable. The sperm whale provided a superior burning oil, but this was expensive. In 1830 a new substance called “camphene” was patented, and it proved to be an excellent illuminant. But while camphene gave a bright light it too remained expensive, had an unpleasant odor, and also was dangerously explosive.

Between 1830 and 1850 it seemed that the only hope for cheaper illumination in the United States was in the wider use of gas. In the 1840’s American gas manufacturers adopted improved British techniques for producing illuminating gas from coal. But the expense of piping gas to the consumer remained so high that until midcentury gaslighting was feasible only in urban areas, and only for public buildings or for the wealthy.

In 1854 a Canadian doctor, Abraham Gesner, patented a process for distilling a pitchlike mineral found in New Brunswick and Nova Scotia that produced illuminating gas and an oil that he called “kerosene” (from “keros,” the Greek word for wax, and “ene” because it resembled camphene). Kerosene, though cheaper than camphene, had an unpleasant odor, and Gesner never made his fortune from it. But Gesner had aroused a new hope for making an illuminating oil from a product coming out of North American mines.

1. Which of the following is NOT mentioned as a reason why better lighting had become necessary by the mid-nineteenth century?
   (A) Development of railroads
   (B) Demand for better medical facilities*
   (C) Increases in the number of new factories
   (D) Growth of cities

2. The phrase “served as” in line 6 is closest in meaning to
   (A) differed from
   (B) functioned as*
   (C) rested upon
   (D) reacted to

3. The word “this” in line 8 refers to
   (A) lard
   (B) charcoal
   (C) wick
   (D) oil*

4. Which of the following is NOT mentioned as a disadvantage of camphene?
   (A) High cost
   (B) Bad smell
(C) Potential to explode
(D) Greasy texture*

5. What can be inferred about the illuminating gas described in the second paragraph?
(A) It was first developed in the United States.
(B) It was not allowed to be used in public buildings.
(C) It was not widely available until midcentury.*
(D) It had an unpleasant smell.

6. The word “resembled” in line 19 is closest in meaning to
(A) was similar to*
(B) cost the same as
(C) was made from
(D) sounded like

7. According to the passage, what advantage did the kerosene patented by Gesner have over camphene?
(A) Kerosene had a more pleasant smell.
(B) Kerosene was less expensive.*
(C) Kerosene burned more brightly.
(D) Kerosene was safer to use.

8. The word “it” in line 20 refers to
(A) fortune
(B) odor
(C) camphene
(D) kerosene*

9. Which of the following best describes the organization of the passage?
(A) A description of events in chronological order*
(B) A comparison of two events
(C) The statement of a theory and possible explanations
(D) An analysis of scientific findings

10. Where in the passage does the author mention the origin of a word?
(A) Lines 4-6.
(B) Lines 7-8.
(C) Lines 12-13.
(D) Lines 16-19.*

Passage 2

The penny press, which emerged in the United States during the 1830’s, was a powerful agent of mass communication. These newspapers were little dailies, generally four pages in length, written for the mass taste. They differed from the staid, formal presentation of the conservative press, with its emphasis on political and literary topics. The new papers were brief and cheap, emphasizing sensational reports of police courts and juicy scandals as well as human interest stories. Twentieth-century journalism was already foreshadowed in the penny press of the 1830’s.

The New York Sun, founded in 1833, was the first successful penny paper, and it was followed two years later by the New York Herald, published by James Gordon Bennett. Not long after, Horace Greeley issued the New York Tribune, which was destined to become the most influential paper in America. Greeley gave space to the issues that deeply touched the American people before the Civil War—abolitionism, temperance, free homesteads, Utopian cooperative settlements, and the problems of labor. The weekly edition of the Tribune, with 100,000 subscribers, had a remarkable influence in rural areas, especially in Western communities.
Americans were reputed to be the most avid readers of periodicals in the world. An English observer enviously calculated that, in 1829, the number of newspapers circulated in Great Britain was enough to reach only one out of every thirty-six inhabitants weekly; Pennsylvania in that same year had a newspaper circulation which reached one out of every four inhabitants weekly. Statistics seemed to justify the common belief that Americans were devoted to periodicals. Newspapers in the United States increased from 1,200 in 1833 to 3,000 by the early 1860’s, on the eve of the Civil War. This far exceeded the number and circulation of newspapers in England and France.

11. What is the author’s main point in the first paragraph?
   (A) The penny press was modeled on earlier papers.
   (B) The press in the nineteenth century reached only a small proportion of the population.
   (C) The penny press became an important way of disseminating information in the first half of the nineteenth century.
   (D) The penny press focused mainly on analysis of politics.

12. What does the author mean by the statement in lines 6-7 that twentieth-century journalism was foreshadowed by the penny press?
   (A) The penny press darkened the reputation of newswriting.
   (B) Twentieth-century journalism is more important than nineteenth-century journalism.
   (C) Penny-press news reporting was more accurate than that in twentieth-century newspapers.
   (D) Modern news coverage is similar to that done by the penny press.

13. Which of the following would LEAST likely be in a penny-press paper?
   (A) A report of theft of union funds by company officials
   (B) An article about a little girl returning a large amount of money she found in the street
   (C) A scholarly analysis of an economic issue of national importance
   (D) A story about land being given away in the West

14. The word “it” in line 8 refers to
   (A) the New York Sun
   (B) the New York Herald
   (C) America
   (D) the Civil War

15. Who was Horace Greeley (line 10)?
   (A) The publisher of the first penny-press paper to make a profit
   (B) The founder of the penny-press paper that did the most to influence the thinking of the public
   (C) The most successful writer for the penny press
   (D) The man who took over James Gordon Bennett’s penny-press paper and made it successful

16. The word “remarkable” in line 14 is closest in meaning to
   (A) significant
   (B) discussable
   (C) remote
   (D) uneven

17. The word “avid” in line 16 is closest in meaning to
   (A) intelligent
   (B) eager
   (C) critical
   (D) thrifty

18. The figures concerning newspaper circulation in Pennsylvania in 1829 are relevant because they
   (A) explain why so many different periodicals were published
(B) prove that weekly periodicals were more successful than daily papers
(C) show the difference between reading habits before and after the Civil War
(D) support the belief that Americans were enthusiastic readers of periodicals*

19. The word “justify” in line 20 is closest in meaning to
(A) generate
(B) calculate
(C) modify
(D) prove*

20. The third paragraph is developed primarily by means of
(A) descriptions
(B) contrasts*
(C) ordering events in time sequence
(D) analysis of a process

21. It can be inferred that penny-press newspapers were all of the following EXCEPT
(A) inexpensive
(B) informal
(C) profitable
(D) thorough*

Passage 3

Broad-tailed hummingbirds often nest in quaking aspens, slender deciduous trees with smooth, gray-green bark found in the Colorado Rockies of the western United States. After flying some 2,000 kilometers north from where they have wintered in Mexico, the hummingbirds need six weeks to build a nest, incubate their eggs, and raise the chicks. A second nest is feasible only if the first fails early in the season. Quality, not quantity, is what counts in hummingbird reproduction. A nest on the lowest intact branch of an aspen will give a hummingbird a good view, a clear flight path, and protection for her young. Male hummingbirds claim feeding territories in open meadows where, from late May through June, they mate with females coming to feed but take no part in nesting. Thus when the hen is away to feed, the nest is unguarded. While the smooth bark of the aspen trunk generally offers a poor grip for the claws of a hungry squirrel or weasel, aerial attacks, from a hawk, owl, or gray jay, are more likely.

The choice of where to build the nest is based not only on the branch itself but also on what hangs over it. A crooked deformity in the nest branch, a second, unusually close branch overhead, or proximity to part of a trunk bowed by a past ice storm are features that provide shelter and make for an attractive nest site. Scarcely larger than a halved golf ball, the nest is painstakingly constructed of spiderwebs and plant down, decorated and camouflaged outside with paper-like bits of aspen bark held together with more strands of spider silk. By early June it will hold two pea-sized eggs, which each weigh one-seventh of the mother's weight, and in sixteen to nineteen days, two chicks.

22. What aspect of broad-tailed hummingbird behavior does the passage mainly discuss?
(A) Migration routes
(B) Mating habits
(C) Caring for the young
(D) Selection of nest sites*

23. According to the passage, in what circumstances do hummingbirds build a second nest?
(A) If the winter is unusually warm
(B) If the chicks in the first nest hatch early

260
24. The word “counts” in line 6 is closest in meaning to
   (A) weighs
   (B) estimates
   (C) matters*
   (D) numbers

25. The word “clear” in line 8 is closest in meaning to
   (A) bright
   (B) exact
   (C) unobstructed*
   (D) transparent

26. The word “they” in line 9 refers to
   (A) male hummingbirds*
   (B) territories
   (C) meadows
   (D) females

27. According to the passage, which of the following is true of the male broad-tailed hummingbird?
   (A) It finds food for the female and the chicks.
   (B) It protects the nest while the female searches for food.
   (C) It is not involved in caring for the chicks.*
   (D) It shares nesting duties equally with the female.

28. It can be inferred from the passage that the broad-tailed hummingbirds’ eggs and chicks are most vulnerable to attacks by
   (A) insects
   (B) humans
   (C) birds*
   (D) squirrels

29. Which of the following would be a good location for a broad-tailed hummingbird to build its nest?
   (A) A branch near the top of a tree
   (B) The longest branch of a tree
   (C) A thick branch
   (D) A protected branch*

30. The word “Scarcely” in line 17 is closest in meaning to
   (A) obviously
   (B) barely*
   (C) consistently
   (D) needlessly

31. Which of the following was NOT mentioned in the passage as a nest-building material of the broad-tailed hummingbird?
   (A) Paper*
   (B) Plant down
   (C) Spiderwebs
   (D) Tree bark

32. The author compares the size of the broad-tailed hummingbird’s nest to
   (A) a pea
(B) a golf ball*
(C) a spiderweb
(D) an egg

33. According to the passage, how long does it take for broad-tailed hummingbird eggs to hatch?
   (A) Less than a week
   (B) Two to three weeks*
   (C) One month
   (D) More than six weeks

34. Where in the passage does the author mention the number of eggs generally found in the nests of
    broad-tailed hummingbirds?
   (A) Line 5
   (B) Lines 10-11*
   (C) Lines 15-17
   (D) Lines 20-22*