VALIDATING THE ASSUMED RELATIONSHIP BETWEEN TASK DESIGN, COGNITIVE COMPLEXITY, AND SECOND LANGUAGE TASK PERFORMANCE

A Dissertation
submitted to the Faculty of the
Graduate School of Arts and Sciences
of Georgetown University
in partial fulfillment of the requirements for the
degree of
Doctor of Philosophy
in Linguistics

By

Shoko Sasayama, M.A.

Washington, DC
December 15, 2015
VALIDATING THE ASSUMED RELATIONSHIP BETWEEN TASK DESIGN, COGNITIVE COMPLEXITY, AND SECOND LANGUAGE TASK PERFORMANCE

Shoko Sasayama, M.A.

Thesis Advisor: John M. Norris, Ph.D.

ABSTRACT

In research on Task-Based Language Teaching, it has traditionally been assumed that differing degrees of cognitive task complexity can be inferred through task design and/or observations of differing qualities in linguistic production elicited by L2 communication tasks. Without validating this assumption, however, it is unclear whether the designed or inferred difference in complexity, the key independent variable, is realized as intended. Furthermore, what exactly makes a task more or less complex has been under-investigated, leaving it unclear why task complexity manipulation worked (or did not) and what specifically it is about a task that positively or negatively affected L2 task performance.

Accordingly, this study adopted diverse methods from cognitive psychology for independently measuring cognitive complexity, including: dual-task methodology, time estimation, and self-ratings. In addition to 61 native speakers as a baseline, 120 English-L2 speakers in Japan, representing distinct proficiency levels, narrated four picture sequences, each containing different numbers of characters (hence, varying degrees of complexity by design). While performing the primary story-telling task, participants simultaneously completed a secondary task of reacting to a color change. After each task, they estimated their time-on-task, rated their perceptions of task difficulty and mental effort exerted, and provided written explanatory comments for their ratings.
Findings revealed a complicated relationship among task design, cognitive demands, and learner task performance. In general, the number of elements affected the level of cognitive task complexity as predicted. Critically, however, only large differences were detectable in terms of independent measures of cognitive load, underscoring the importance of validating the assumed relationship. Other factors identified as contributing to complexity included: conceptual input, code complexity, and performance factors. These factors were found to induce facilitative as well as extraneous cognitive demands; hence, they had both positive and negative impacts on learners’ task performances. Findings also indicated clear proficiency effects and interaction effects among learner proficiency, task design, and measures of cognitive load, as well as with performance variables.
ACKNOWLEDGMENTS

Reflecting on my long journey of choosing a topic, planning and conducting research, and actually writing up a dissertation, what stands out the most is how fortunate I have been to have such supportive mentors, colleagues, and friends throughout my career as a doctoral student at Georgetown University as well as at the University of Hawai‘i at Mānoa. First and foremost, my deepest sincere gratitude goes to my advisor, Dr. John Norris, who has guided me from the beginning of this dissertation project through the very end. His excellent and generous support has given me desire for continuous self-improvement, confidence in myself as a researcher, and above all the key asset of my academic life – knowledge and expertise in a wide variety of domains. Words cannot express how grateful I am, but sincerely “Thank you!” would not hurt!

Secondly, I would like to extend my deep appreciation to my committee members, Dr. Darlene Howard, Dr. Yoshiko Mori, and Dr. Peter Skehan, for their wonderful guidance and continuous encouragement. Without their generous support, this dissertation would not have been possible.

Thirdly, I want to thank those who willingly participated in the dissertation research. You have given me (and the field as a whole) numerous valuable insights into the issue of cognitive task complexity. My gratitude also goes to my colleagues who have kindly helped me recruit these participants as well as those who have kindly given me their time and effort to transcribe and code data. Your help was truly invaluable and indispensable to this dissertation – thank you! Also, I’d like to express my appreciation to my mentors, colleagues, and friends who have given me lots of valuable feedback on the earlier versions of the dissertation (in the form of in-class and conference presentations) – it helped improve and polish my dissertation greatly.
Next, my sincere gratitude also goes to Dr. Heidi Byrnes, editor of *The Modern Language Journal*, for her sustained interest in my work. I am also deeply grateful for the dissertation award sponsored jointly by *The Modern Language Journal, The American Council on the Teaching of Foreign Languages*, and *The National Federation of Modern Language Teachers’ Association*.

Last, but not least, I would like to thank my dear family for their love, support, and trust. Their love has been the fuel for making progress and moving forward. *Otosan* (or Dad), thank you for showing me how fun and rewarding it is to be in academia. I have seen you design and create your sculptures with great joy, and it always made me feel that I would want to have a job that I would love for the rest of my life. *Okasan* (or Mum), thank you for letting me explore my potentials since early childhood and for always having trust in me. Towards the end of my dissertation writing, *obento* (a lunch box) that you made for me every day (including weekends!) motivated me to do the best I could and to keep moving forward. Finally, *Bun-chan* (my brother), it has been great to be able to share with you how fun and tough at the same time it is to be in academia. It has helped me a lot to know that I could always talk to you and that you would understand without much explanation what it takes to be a researcher.

Thank you all for your love and support. Without you, I literally would not have been able to write what is written here. I would like to dedicate this dissertation to everyone who has given me such generous support along this long journey – thank you.
# Table of Contents

CHAPTER I: INTRODUCTION ........................................................................................................... 1  
1.1 Tasks and second language acquisition .................................................................................. 1  
1.2 Cognitive task complexity as a task design and sequencing principle ................................. 2  
  1.2.1 Cognitive task complexity in educational research ......................................................... 3  
  1.2.2 Cognitive task complexity in instructed SLA research .................................................... 8  
    1.2.2.1 Early attempts to use task complexity as a design and sequencing principle ............ 8  
    1.2.2.2 Robinson’s Cognition Hypothesis ................................................................................ 11  
    1.2.2.3 Skehan’s Limited Attentional Capacity approach ......................................................... 18  
    1.2.2.4 Robinson’s Cognition Hypothesis and Skehan’s Trade-Off Hypothesis: How are they different? ........................................................................................................................................ 21  
  1.3 Problem statement .................................................................................................................. 24  
  1.4 Definitions of constructs ........................................................................................................ 27  
  1.5 Organization of the dissertation ............................................................................................ 29  

CHAPTER II: COGNITIVE TASK COMPLEXITY AS A CAUSAL MECHANISM IN SECOND LANGUAGE PERFORMANCE .................................................................................................................. 33  
2.1 Introduction ............................................................................................................................ 33  
2.2 Predicted effects of cognitive task complexity on L2 performance and learning ............... 33  
2.3 Understanding the status quo .................................................................................................. 35  
2.4 Chapter summary .................................................................................................................... 46  

CHAPTER III: MEASURING COGNITIVE TASK COMPLEXITY ............................................... 48  
3.1 Introduction ............................................................................................................................ 48  
3.2 Measures of cognitive load used in cognitive psychology and multimedia learning .......... 50
3.3 Measures of task complexity used in the field of TBLT ............................................. 57
3.4 What does data elicited by measures of cognitive load really tell us? ......................... 62
3.5 Chapter summary ........................................................................................................... 64

CHAPTER IV: THE ROLE OF PROFICIENCY IN UNDERSTANDING TASK COMPLEXITY ........................................................................................................ 66
4.1 Introduction ...................................................................................................................... 66
4.2 Proficiency as a moderating variable in SLA and TBLT research ................................. 67
4.3 Measurement of L2 proficiency ....................................................................................... 76
4.4 Developing short-cut estimates of L2 proficiency ......................................................... 77
   4.4.1 Cloze test ..................................................................................................................... 78
   4.4.2 Elicited imitation test ............................................................................................... 79
4.5 Chapter summary ............................................................................................................. 80

CHAPTER V: THE CURRENT STUDY ................................................................................. 82
5.1 Rationale for the current study ....................................................................................... 82
5.2 Research questions .......................................................................................................... 85
5.3 Methods .......................................................................................................................... 86
   5.3.1 Participants ............................................................................................................... 86
      5.3.1.1 Native speakers .................................................................................................. 87
      5.3.1.2 English-L2 speakers ......................................................................................... 88
   5.3.2 Materials and instruments ....................................................................................... 90
      5.3.2.1 Primary tasks ..................................................................................................... 90
      5.3.2.2 Secondary task .................................................................................................. 96
      5.3.2.3 Time estimation ................................................................................................. 97
5.3.2.4 Self-assessment of perceived task difficulty and mental effort .......................... 98
5.3.2.5 Proficiency tests ........................................................................................................ 99
5.3.2.6 Background questionnaire ....................................................................................... 100
5.3.3 Procedure ...................................................................................................................... 100
  5.3.3.1 Pilot study 1 ............................................................................................................... 101
  5.3.3.2 Pilot study 2 ............................................................................................................... 102
  5.3.3.3 Procedures in the main study .................................................................................. 104
5.3.4 Measurement and data analysis ................................................................................... 105
  5.3.4.1 Measures and data analysis for Study Part 1 ............................................................ 106
    5.3.4.1.1 Dual-task methodology ...................................................................................... 106
    5.3.4.1.2 Time estimation ................................................................................................. 107
    5.3.4.1.3 Self-assessment of task difficulty and mental effort ........................................... 108
    5.3.4.1.4 Inter-coder reliability ......................................................................................... 113
  5.3.4.2 Measures and data analysis for Study Part 2 ............................................................ 119
    5.3.4.2.1 Unit of analysis ................................................................................................. 119
    5.3.4.2.2 Accuracy ............................................................................................................ 120
    5.3.4.2.3 Fluency .............................................................................................................. 121
    5.3.4.2.4 Complexity ....................................................................................................... 122
    5.3.4.2.5 Inter-coder reliability ......................................................................................... 125
  5.3.4.3 Proficiency test ........................................................................................................ 126
    5.3.4.3.1 Cloze test .......................................................................................................... 126
    5.3.4.3.2 Elicited imitation test ....................................................................................... 126
5.3.4.3.3 Inter-rater reliability ................................................................. 129
5.3.4.4 Statistical analysis ................................................................. 130
CHAPTER VI: RESULTS ............................................................................. 133
6.1 Introduction ............................................................................... 133
6.2 Measured cognitive load: Quantitative results ........................................ 133
  6.2.1 Native speakers of English .............................................................. 133
    6.2.1.1 Dual-task methodology ......................................................... 136
    6.2.1.2 Time estimation ................................................................ 137
    6.2.1.3 Self-ratings of perceived difficulty and mental effort ................. 140
    6.2.1.4 Summary of complexity measures between tasks for native speakers .... 143
  6.2.2 English L2 speakers .................................................................... 144
    6.2.2.1 Overall trends .................................................................... 144
      6.2.2.1.1 Dual-task methodology ............................................... 146
      6.2.2.1.2 Time estimation ......................................................... 147
      6.2.2.1.3 Self-ratings of perceived difficulty and mental effort .......... 150
      6.2.2.1.4 Summary of complexity measures between proficiency tasks for English L2 speakers ......................................................... 153
    6.2.2.2 Role of proficiency .............................................................. 153
      6.2.2.2.1 Proficiency tests .......................................................... 154
      6.2.2.2.2 Cognitive load measures by proficiency ............................ 156
      6.2.2.2.3 Summary of proficiency differences ................................. 169
  6.3 Measured cognitive load for English L2 speakers: Open-ended responses ...... 170
<table>
<thead>
<tr>
<th>Section</th>
<th>Subsection</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.3.1</td>
<td>Overall patterns</td>
<td>171</td>
</tr>
<tr>
<td></td>
<td>Conceptual input</td>
<td>171</td>
</tr>
<tr>
<td></td>
<td>Task 1</td>
<td>171</td>
</tr>
<tr>
<td></td>
<td>Task 2</td>
<td>179</td>
</tr>
<tr>
<td></td>
<td>Task 3</td>
<td>187</td>
</tr>
<tr>
<td></td>
<td>Task 4</td>
<td>195</td>
</tr>
<tr>
<td></td>
<td>Code complexity</td>
<td>201</td>
</tr>
<tr>
<td></td>
<td>Task 1</td>
<td>201</td>
</tr>
<tr>
<td></td>
<td>Task 2</td>
<td>207</td>
</tr>
<tr>
<td></td>
<td>Task 3</td>
<td>213</td>
</tr>
<tr>
<td></td>
<td>Task 4</td>
<td>218</td>
</tr>
<tr>
<td></td>
<td>Performance factors</td>
<td>222</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>226</td>
</tr>
<tr>
<td></td>
<td>Summary of the overall patterns</td>
<td>227</td>
</tr>
<tr>
<td>6.3.2</td>
<td>Explanatory comments by proficiency</td>
<td>230</td>
</tr>
<tr>
<td></td>
<td>Task 1</td>
<td>230</td>
</tr>
<tr>
<td></td>
<td>Task 2</td>
<td>233</td>
</tr>
<tr>
<td></td>
<td>Task 3</td>
<td>237</td>
</tr>
<tr>
<td></td>
<td>Task 4</td>
<td>239</td>
</tr>
<tr>
<td></td>
<td>Summary of open-ended response patterns by proficiency</td>
<td>248</td>
</tr>
<tr>
<td>6.4</td>
<td>Performance measures</td>
<td>250</td>
</tr>
<tr>
<td></td>
<td>Native speakers of English</td>
<td>251</td>
</tr>
</tbody>
</table>
6.4.1.1 Syntactic complexity .............................................................. 253
  6.4.1.1.1 Mean length of T-unit .................................................... 253
  6.4.1.1.2 Mean length of clauses .................................................. 255
  6.4.1.1.3 Clauses per T-unit ....................................................... 257
6.4.1.2 Lexical variety .................................................................. 260
6.4.1.3 Fluency ............................................................................ 264
6.4.1.4 Summary of performance measures for native speakers .......... 268
6.4.2 English L2 speakers ............................................................... 269
  6.4.2.1 Overall trends ................................................................. 269
    6.4.2.1.1 Syntactic complexity .................................................... 272
    6.4.2.1.2 Accuracy .................................................................. 278
    6.4.2.1.3 Lexical variety .......................................................... 280
    6.4.2.1.4 Fluency .................................................................... 284
    6.4.2.1.5 Summary of performance measures for English L2 speakers 288
  6.4.2.2 Performance measures by proficiency .................................. 289
    6.4.2.2.1 Syntactic complexity .................................................... 290
    6.4.2.2.2 Accuracy .................................................................. 300
    6.4.2.2.3 Lexical variety .......................................................... 303
    6.4.2.2.4 Fluency .................................................................... 308
    6.4.2.2.5 Summary of proficiency differences .............................. 313
6.4.3 Influence of the dual-task methodology on task performance .......... 314
6.5 Chapter summary .................................................................... 315
CHAPTER VII: DISCUSSION

7.1 Introduction ........................................................................................................................................... 320

7.2 Validating the assumed relationship between task design and the level of cognitive complexity ....................................................................................................................................... 320

7.2.1 Summary of overall patterns ................................................................................................................. 320

7.2.2 Further discussion ................................................................................................................................. 324

7.2.2.1 Relationship between learner factors and the level of cognitive load .............................................. 324

7.2.2.2 Relationship between the number of elements and the level of cognitive load ............................ 327

7.2.2.3 Types of cognitive load: Facilitative versus extraneous cognitive load ......................................... 329

7.2.2.4 Differences among the cognitive load measures ................................................................................ 334

7.2.2.4.1 Effect sizes ....................................................................................................................................... 334

7.2.2.4.2 Interaction with proficiency ........................................................................................................... 335

7.2.2.4.3 Possibilities of the cognitive load measures: Do they function as intended in the context of TBLT research? ..................................................................................................................................... 340

7.3 Effects of measured cognitive complexity on task performance .............................................................. 346

7.3.1 Summary of overall patterns ................................................................................................................. 346

7.3.2 Further discussion ................................................................................................................................. 348

7.3.2.1 Why did Task 3 elicit the best performance of the four tasks? ......................................................... 349

7.3.2.2 Proficiency effects ........................................................................................................................... 352

7.3.2.3 Differences among performance measures ....................................................................................... 352

CHAPTER VIII: CONCLUSIONS .................................................................................................................. 355

8.1 Introduction ..................................................................................................................................................... 355

8.2 Limitations ..................................................................................................................................................... 355
8.3 Concluding remarks ........................................................................................................ 358
  8.3.1 Importance of the current study: So what? ......................................................... 358
  8.3.2 Implications .............................................................................................................. 361
    8.3.2.1 Implications for researching the effects of cognitive task complexity .......... 361
    8.3.2.2 Implications for teaching ................................................................................ 365
  8.3.3 Future research ........................................................................................................ 367
Appendix A ......................................................................................................................... 370
Appendix B ......................................................................................................................... 371
Appendix C ......................................................................................................................... 372
Appendix D ......................................................................................................................... 374
Appendix E ......................................................................................................................... 375
Appendix F ......................................................................................................................... 376
Appendix G ......................................................................................................................... 377
Appendix H ......................................................................................................................... 378
Appendix I ......................................................................................................................... 379
Appendix J ......................................................................................................................... 381
Appendix K ......................................................................................................................... 383
Appendix L ......................................................................................................................... 385
Appendix M ......................................................................................................................... 387
Appendix N ......................................................................................................................... 388
Appendix O ......................................................................................................................... 389
Appendix P ......................................................................................................................... 390
List of Tables

Table 1. Example Dual-Tasks With (Supposedly) Distinct Difficulty for Each Attentional Pool .................................................................................................................................................................................. 12

Table 2. Summary of Task Complexity Operationalizations with Example Studies ................. 36

Table 3. Characteristics and Results of Studies That Operationalized Task Complexity by [+/- few elements] and Investigated its Effects on CALF Indices .................................................................................................................. 40

Table 4. Characteristics and Results of Studies That Used a Narrative Task, Operationalized Task Complexity by [+/- pre-task planning], and Investigated its Effects on CALF Indices .... 43

Table 5. List of L1-English Participants’ Majors in Study Part 1 ............................................. 87

Table 6. List of L1-English Participants’ Majors in Study Part 2 ............................................. 88

Table 7. Participants’ Self-Reported Scores of Standardized English Tests in Study Part 1 .... 89

Table 8. Participants’ Self-Reported Scores of Standardized English Tests in Study Part 2 .... 90

Table 9. Summary of Task Features ....................................................................................... 94

Table 10. Specific Measures Extracted ................................................................................... 106

Table 11. Superordinate/Subordinate Categories and Their Definitions and Examples ........... 110

Table 12. Superordinate/Subordinate Categories and Their Definitions and Examples (Revised) .................................................................................................................................................. 115

Table 13. Means/Medians (and Standard Deviations) and Effect Sizes (d) for Each Cognitive Load Measure on the Four Tasks in Part 1 (for Native Speakers of English, n = 30) ............... 135

Table 14. Means (and Standard Deviations) and Effect Sizes (d) for Each Cognitive Load Measure on the Four Tasks in Part 2 (for Native Speakers of English, n = 31) ....................... 136

Table 15. Means/Medians (and Standard Deviations) for Each Cognitive Load Measure on the Four Tasks and Effect Sizes (d) in Part 1 (for English-L2 Speakers, n = 53) ................. 145

Table 16. Means (and Standard Deviations) for Each Cognitive Load Measure on the Four Tasks and Effect Sizes (d) in Part 2 (for English-L2 Speakers, n = 67) .......................... 146

Table 17. The Mean Score, Standard Deviation, and Range on Elicited Imitation Test for Each Proficiency Group .................................................................................................................. 155
Table 18. Means/Medians (and Standard Deviations) and Effect Sizes \((d)\) for Dual-Task Methodology and Time Estimation on the Four Tasks by Proficiency in Part 1 ....................... 157

Table 19. Means/Medians (and Standard Deviations) and Effect Sizes \((d)\) for Task Difficulty and Mental Effort on the Four Tasks by Proficiency in Part 1 .................................................. 157

Table 20. Means (and Standard Deviations) and Effect Sizes \((d)\) for Each Cognitive Load Measure on the Four Tasks by Proficiency in Part 2 ................................................................. 158

Table 21. Summary Table With the Number of Comments for Each Category of Conceptual Input (Task 1, Part 1) ................................................................. 173

Table 22. Summary Table With the Number of Comments for Each Category of Conceptual Input (Task 1, Part 2) ................................................................. 177

Table 23. Summary Table With the Number of Comments for Each Category of Conceptual Input (Task 2, Part 1) ................................................................. 183

Table 24. Summary Table With the Number of Comments for Each Category of Conceptual Input (Task 2, Part 2) ................................................................. 185

Table 25. Summary Table With the Number of Comments for Each Category of Conceptual Input (Task 3, Part 1) ................................................................. 189

Table 26. Summary Table With the Number of Comments for Each Category of Conceptual Input (Task 3, Part 2) ................................................................. 193

Table 27. Summary Table With the Number of Comments for Each Category of Conceptual Input (Task 4, Part 1) ................................................................. 196

Table 28. Summary Table With the Number of Comments for Each Category of Conceptual Input (Task 4, Part 2) ................................................................. 199

Table 29. Summary Table With the Number of Comments for Each Category of Code Complexity (Task 1, Part 1) ................................................................. 204

Table 30. Summary Table With the Number of Comments for Each Category of Code Complexity (Task 1, Part 2) ................................................................. 204

Table 31. Summary Table With the Number of Comments for Each Category of Code Complexity (Task 2, Part 1) ................................................................. 209
Table 32. Summary Table With the Number of Comments for Each Category of Code Complexity (Task 2, Part 2) ................................................................. 210

Table 33. Summary Table With the Number of Comments for Each Category of Code Complexity (Task 3, Part 1) .......................................................................................... 214

Table 34. Summary Table With the Number of Comments for Each Category of Code Complexity (Task 3, Part 2) .......................................................................................... 215

Table 35. Summary Table With the Number of Comments for Each Category of Code Complexity (Task 4, Part 1) .......................................................................................... 219

Table 36. Summary Table With the Number of Comments for Each Category of Code Complexity (Task 4, Part 2) .......................................................................................... 219

Table 37. Summary Table With the Number of Comments for Each Category of Performance Factors (Part 1) ......................................................................................... 222

Table 38. Summary Table With the Number of Comments for Each Category of Performance Factors (Part 2) ......................................................................................... 223

Table 39. Example Comments Related to Performance Factors for Each Task Provided in Part 1 and Part 2 ........................................................................................................ 224

Table 40. Summary Table With the Number of Comments for Each Superordinate Category by Proficiency (Task 1, Part 1) ........................................................................... 232

Table 41. Summary Table With the Number of Comments for Each Superordinate Category by Proficiency (Task 1, Part 2) ........................................................................... 233

Table 42. Summary Table With the Number of Comments for Each Superordinate Category by Proficiency (Task 2, Part 1) ........................................................................... 236

Table 43. Summary Table With the Number of Comments for Each Superordinate Category by Proficiency (Task 2, Part 2) ........................................................................... 237

Table 44. Summary Table With the Number of Comments for Each Superordinate Category by Proficiency (Task 3, Part 1) ........................................................................... 238

Table 45. Summary Table With the Number of Comments for Each Superordinate Category by Proficiency (Task 3, Part 2) ........................................................................... 239

Table 46. Summary Table With the Number of Comments for Each Superordinate Category by Proficiency (Task 4, Part 1) ........................................................................... 240
Table 47. Summary Table With the Number of Comments for Each Superordinate Category by Proficiency (Task 4, Part 2) ................................................................. 241

Table 48. Number of Comments with Some Example Responses on ‘Number of Elements’ as Justifications for High Ratings on Difficulty or Mental Effort for Task 4 by Proficiency ........ 242

Table 49. Number of Comments with Some Example Responses on General versus Specific Task-Induced Code Complexity as Justifications for High Ratings on Difficulty or Mental Effort in Part 1 (Low Proficiency Group) ........................................................................................................ 244

Table 50. Number of Comments with Some Example Responses on General versus Specific Task-Induced Code Complexity as Justifications for High Ratings on Difficulty or Mental Effort in Part 2 (Low Proficiency Group) ........................................................................................................ 245

Table 51. Number of Comments with Some Example Responses on General versus Specific Task-Induced Code Complexity as Justifications for High Ratings on Difficulty or Mental Effort in Part 1 (High Proficiency Group) ........................................................................................................ 246

Table 52. Number of Comments with Some Example Responses on General versus Specific Task-Induced Code Complexity as Justifications for High Ratings on Difficulty or Mental Effort in Part 2 (High Proficiency Group) ........................................................................................................ 247

Table 53. Means/Medians (and Standard Deviations) and Effect Sizes ($d$) for Each Performance Measure on the Four Tasks in Part 1 (for Native Speakers of English, $n = 30$) ................. 252

Table 54. Means/Medians (and Standard Deviations) and Effect Sizes ($d$) for Each Performance Measure on the Four Tasks in Part 2 (for Native Speakers of English, $n = 31$) ................. 252

Table 55. Means/Medians (and Standard Deviations) and Effect Sizes ($d$) for Each Performance Measure on the Four Tasks in Part 1 (for English-L2 Speakers, $n = 53$) ......................... 270

Table 56. Means/Medians (and Standard Deviations) and Effect Sizes ($d$) for Each Performance Measure on the Four Tasks in Part 2 (for English-L2 Speakers, $n = 67$) ......................... 271

Table 57. Means/Medians (and Standard Deviations) and Effect Sizes ($d$) for Syntactic Complexity Measures on the Four Tasks by Proficiency in Part 1 ......................................................... 291

Table 58. Means/Medians (and Standard Deviations) and Effect Sizes ($d$) for Syntactic Complexity Measures on the Four Tasks by Proficiency in Part 2 ......................................................... 292

Table 59. Means/Medians (and Standard Deviations) and Effect Sizes ($d$) for the Accuracy Measure on the Four Tasks by Proficiency in Part 1 ................................................................. 301
Table 60. Means (and Standard Deviations) and Effect Sizes ($d$) for the Accuracy Measure on the Four Tasks by Proficiency in Part 2 .......................................................... 302

Table 61. Means/Medians (and Standard Deviations) for Lexical Variety Measures on the Four Tasks and Effect Sizes ($d$) by Proficiency in Part 1 ........................................................................ 304

Table 62. Means/Medians (and Standard Deviations) for Lexical Variety Measures on the Four Tasks and Effect Sizes ($d$) by Proficiency in Part 2 ........................................................................ 304

Table 63. Means/Medians (and Standard Deviations) and Effect Sizes ($d$) for the Fluency Measures on the Four Tasks by Proficiency in Part 1 ................................................................. 309

Table 64. Means/Medians (and Standard Deviations) and Effect Sizes ($d$) for the Fluency Measures on the Four Tasks by Proficiency in Part 2 ................................................................. 310
LIST OF FIGURES

Figure 1. Relationship between ‘cognitive task complexity,’ ‘cognitive load,’ ‘task difficulty,’ and ‘mental effort.’ ................................................................. 29

Figure 2. Effect sizes between the simple and complex tasks for low and high proficiency groups in Ishikawa’s (2006) study. ................................................................. 71

Figure 3. Effect sizes between the simple and complex tasks for low and high proficiency groups in Yoshida’s (2012) study. ................................................................. 72

Figure 4. Effect sizes between the simple and complex tasks for low and high proficiency groups in Kim’s (2009) study. ................................................................. 74

Figure 5. Screenshot of a hypothetical primary narrative task and the secondary color-detection task. ........................................................................................................ 97

Figure 6. Reaction time of the dual-task methodology for the four tasks with 95% confidence intervals in Part 1 (for native speakers of English). ........................................ 137

Figure 7. Time estimation for the four tasks with 95% confidence intervals in Part 1 (for native speakers of English). ................................................................. 139

Figure 8. Time estimation for the four tasks with 95% confidence intervals in Part 2 (for native speakers of English). ................................................................. 139

Figure 9. Task difficulty ratings for the four tasks with 95% confidence intervals in Part 1 (for native speakers of English). ................................................................. 141

Figure 10. Mental effort ratings for the four tasks with 95% confidence intervals in Part 1 (for native speakers of English). ................................................................. 142

Figure 11. Task difficulty ratings for the four tasks with 95% confidence intervals in Part 1 (for native speakers of English). ................................................................. 142

Figure 12. Mental effort ratings for the four tasks with 95% confidence intervals in Part 2 (for native speakers of English). ................................................................. 143

Figure 13. Reaction time of the dual-task methodology for the four tasks with 95% confidence intervals in Part 1 (for English L2 speakers). ........................................ 147

Figure 14. Time estimation for the four tasks with 95% confidence intervals in Part 1 (for English L2 speakers). ................................................................. 149
Figure 15. Time estimation for the four tasks with 95% confidence intervals in Part 2 (for English L2 speakers). .......................................................... 149

Figure 16. Task difficulty ratings for the four tasks with 95% confidence intervals in Part 1 (English L2 speakers). .......................................................... 151

Figure 17. Mental effort ratings for the four tasks with 95% confidence intervals in Part 1 (English L2 speakers). .......................................................... 151

Figure 18. Task difficulty ratings for the four tasks with 95% confidence intervals in Part 1 (English L2 speakers). .......................................................... 152

Figure 19. Mental effort ratings for the four tasks with 95% confidence intervals in Part 2 (English L2 speakers). .......................................................... 152

Figure 20. Reaction time of the dual-task methodology for the four tasks with 95% confidence intervals by proficiency in Part 1. .................................................. 160

Figure 21. Time estimation for the four tasks with 95% confidence intervals by proficiency in Part 1. .......................................................... 161

Figure 22. Time estimation for the four tasks with 95% confidence intervals by proficiency in Part 2. .......................................................... 163

Figure 23. Task difficulty for the four tasks with 95% confidence intervals by proficiency in Part 2. .......................................................... 165

Figure 24. Mental effort for the four tasks with 95% confidence intervals by proficiency in Part 1. .......................................................... 166

Figure 25. Task difficulty for the four tasks with 95% confidence intervals by proficiency in Part 2. .......................................................... 168

Figure 26. Mental effort for the four tasks with 95% confidence intervals by proficiency in Part 2. .......................................................... 169

Figure 27. Input as a source of easiness and difficulty (Part 1, Task 1). ...................... 174

Figure 28. Input as a source of easiness and difficulty (Part 2, Task 1). ...................... 178

Figure 29. Input as a source of easiness and difficulty (Part 1, Task 2). ...................... 184

Figure 30. Input as a source of easiness and difficulty (Part 2, Task 2). ...................... 186
Figure 31. Input as a source of easiness and difficulty (Part 1, Task 3). ................................ 190
Figure 32. Input as a source of easiness and difficulty (Part 2, Task 3). ................................ 194
Figure 33. Input as a source of easiness and difficulty (Part 1, Task 4). ................................ 197
Figure 34. Input as a source of easiness and difficulty (Part 2, Task 4). ................................ 200
Figure 35. Code complexity as a source of easiness and difficulty (Part 1, Task 1). ........... 205
Figure 36. Code complexity as a source of easiness and difficulty (Part 2, Task 1). ........... 206
Figure 37. Code complexity as a source of easiness and difficulty (Part 1, Task 2). ........... 211
Figure 38. Code complexity as a source of easiness and difficulty (Part 2, Task 2). ........... 212
Figure 39. Code complexity as a source of easiness and difficulty (Part 1, Task 3). ........... 216
Figure 40. Code complexity as a source of easiness and difficulty (Part 2, Task 3). ........... 217
Figure 41. Code complexity as a source of easiness and difficulty (Part 1, Task 4). ........... 220
Figure 42. Code complexity as a source of easiness and difficulty (Part 2, Task 4). ........... 221
Figure 43. MLT for the four tasks with 95% confidence intervals in Part 1 (for native speakers of English). .......................................................... 254
Figure 44. MLT for the four tasks with 95% confidence intervals in Part 2 (for native speakers of English). .......................................................... 255
Figure 45. MLC for the four tasks with 95% confidence intervals in Part 1 (for native speakers of English). .......................................................... 256
Figure 46. MLC for the four tasks with 95% confidence intervals in Part 2 (for native speakers of English). .......................................................... 257
Figure 47. C/T for the four tasks with 95% confidence intervals in Part 1 (for native speakers of English). .......................................................... 259
Figure 48. C/T for the four tasks with 95% confidence intervals in Part 2 (for native speakers of English). .......................................................... 260
Figure 49. D score for the four tasks with 95% confidence intervals in Part 1 (for native speakers of English). .......................................................... 261
Figure 50. $D$ score for the four tasks with 95% confidence intervals in Part 2 (for native speakers of English). ................................................................. 262

Figure 51. TTR for the four tasks with 95% confidence intervals in Part 1 (for native speakers of English). ......................................................................................... 263

Figure 52. TTR for the four tasks with 95% confidence intervals in Part 2 (for native speakers of English). ......................................................................................... 264

Figure 53. Total number of words for the four tasks with 95% confidence intervals in Part 1 (for native speakers of English). ......................................................................................... 265

Figure 54. Total number of words for the four tasks with 95% confidence intervals in Part 2 (for native speakers of English). ......................................................................................... 266

Figure 55. W/S for the four tasks with 95% confidence intervals in Part 1 (for native speakers of English). ......................................................................................... 267

Figure 56. W/S for the four tasks with 95% confidence intervals in Part 2 (for native speakers of English). ......................................................................................... 268

Figure 57. MLT for the four tasks with 95% confidence intervals in Part 1 (for English L2 speakers). ......................................................................................... 273

Figure 58. MLT for the four tasks with 95% confidence intervals in Part 2 (for English L2 speakers). ......................................................................................... 274

Figure 59. MLC for the four tasks with 95% confidence intervals in Part 1 (for English L2 speakers). ......................................................................................... 275

Figure 60. MLC for the four tasks with 95% confidence intervals in Part 2 (for English L2 speakers). ......................................................................................... 276

Figure 61. C/T for the four tasks with 95% confidence intervals in Part 1 (for English L2 speakers). ......................................................................................... 277

Figure 62. C/T for the four tasks with 95% confidence intervals in Part 2 (for English L2 speakers). ......................................................................................... 278

Figure 63. Accuracy for the four tasks with 95% confidence intervals in Part 1 (for English L2 speakers). ......................................................................................... 279

xxiv
Figure 64. Accuracy for the four tasks with 95% confidence intervals in Part 2 (for English L2 speakers). ........................................................................................................................................ 280

Figure 65. D score for the four tasks with 95% confidence intervals in Part 1 (for English L2 speakers). ........................................................................................................................................ 281

Figure 66. D score for the four tasks with 95% confidence intervals in Part 2 (for English L2 speakers). ........................................................................................................................................ 282

Figure 67. TTR for the four tasks with 95% confidence intervals in Part 1 (for English L2 speakers). ........................................................................................................................................ 283

Figure 68. TTR for the four tasks with 95% confidence intervals in Part 1 (for English L2 speakers). ........................................................................................................................................ 284

Figure 69. Total number of words for the four tasks with 95% confidence intervals in Part 1 (for English L2 speakers). ........................................................................................................................................ 285

Figure 70. Total number of words for the four tasks with 95% confidence intervals in Part 2 (for English L2 speakers). ........................................................................................................................................ 286

Figure 71. W/S for the four tasks with 95% confidence intervals in Part 1 (for English L2 speakers). ........................................................................................................................................ 287

Figure 72. W/S for the four tasks with 95% confidence intervals in Part 2 (for English L2 speakers). ........................................................................................................................................ 288

Figure 73. MLT for the four tasks with 95% confidence intervals by proficiency in Part 1. ... 294

Figure 74. MLT for the four tasks with 95% confidence intervals by proficiency in Part 2. ... 295

Figure 75. MLC for the four tasks with 95% confidence intervals by proficiency in Part 1. ... 296

Figure 76. MLC for the four tasks with 95% confidence intervals by proficiency in Part 2. ... 297

Figure 77. C/T for the four tasks with 95% confidence intervals by proficiency in Part 1. ..... 299

Figure 78. C/T for the four tasks with 95% confidence intervals by proficiency in Part 2. ..... 300

Figure 79. Accuracy for the four tasks with 95% confidence intervals by proficiency in Part 1. ........................................................................................................................................ 302

Figure 80. Accuracy for the four tasks with 95% confidence intervals by proficiency in Part 2. ........................................................................................................................................ 303
Figure 81. $D$ score for the four tasks with 95% confidence intervals by proficiency in Part 1. ...................................................................................................................... 305

Figure 82. $D$ score for the four tasks with 95% confidence intervals by proficiency in Part 2. ...................................................................................................................... 306

Figure 83. TTR for the four tasks with 95% confidence intervals by proficiency in Part 1. ...................................................................................................................... 307

Figure 84. TTR for the four tasks with 95% confidence intervals by proficiency in Part 2. .... 308

Figure 85. Total number of words for the four tasks with 95% confidence intervals by proficiency in Part 1. ...................................................................................................................... 310

Figure 86. Total number of words for the four tasks with 95% confidence intervals by proficiency in Part 2. ...................................................................................................................... 311

Figure 87. W/S for the four tasks with 95% confidence intervals by proficiency in Part 1. .... 312

Figure 88. W/S for the four tasks with 95% confidence intervals by proficiency in Part 2. .... 313
CHAPTER I: INTRODUCTION

1.1 Tasks and second language acquisition

In the field of second language acquisition (SLA), second language (L2) tasks or communication activities have attracted researchers’ as well as teachers’ and materials developers’ attention for some time as an important contributor to fostering L2 learning. In particular, the use of tasks in language classrooms is assumed to have a facilitative role in instructed SLA (i.e., intentional or designed language acquisition that takes place in educational settings) for a variety of reasons, including (a) encouraging learners to see language as a functional, usage-based system intended primarily for accomplishing various kinds of communication (e.g., Bygate, Norris, & Van den Branden, 2015); (b) facilitating learners’ noticing of the holes or gaps in their interlanguage (e.g., Swain, 1993, 1995; Swain & Lapkin, 1995); (c) enabling learners to create new form-meaning connections in their existing L2 knowledge through negotiation of meaning with, and provision of feedback from, their interlocutors (e.g., Long, 1996; Mackey, 1999); and (d) motivating learners to produce more target-like language and helping them to develop positive attitudes towards L2 learning in general (e.g., Dörnyei, 2002, 2003). These theories of and research on instructed SLA have played a key role in developing Task-Based Language Teaching (TBLT), an approach to language education in which tasks are used as the core unit of analysis for syllabus design, materials development, teaching, and assessment (Long, 2015; Norris, 2009). In this sense, TBLT is increasingly referred to as a ‘researched pedagogy’, and precisely for this reason, it has been a – perhaps the – key player in informing much of the research on instructed SLA for the last few decades (Bygate, Norris, & Van den Branden, 2009, 2015; Norris, 2009, 2011).
Given the prominence of TBLT as a way of framing, investigating, and hopefully improving the practices and outcomes of instructed SLA, it is essential that the core instructional/acquisitional unit of analysis – task – be understood in terms of its potential to influence L2 teaching and learning. This dissertation therefore focuses on illuminating the various dimensions of a task that may (or may not) affect how learners engage with language and draw upon their cognitive resources as they attempt to accomplish the task. Although several theoretical frameworks have been proposed that discuss the relationship between task design features and learners’ cognitive activities, especially attention allocation (e.g., Robinson’s Cognition Hypothesis, Skehan’s Limited Attentional Capacity approach), it is as yet unclear how exactly tasks influence learners in cognitive terms and why learners react to the task at hand in a way that they do. Currently, much of the research that utilizes L2 tasks produces at best an incomplete picture of the relationship with learners’ mental responses to them, and by extension, why and how learners use and learn language as a result.

1.2 Cognitive task complexity as a task design and sequencing principle

Since TBLT is “much more than a language teaching ‘method’” (Norris, 2009, p. 581) and ultimately concerned with holistic language education programs, rather than a single task, activity, or lesson, it addresses multiple program elements, including learner needs analysis, curriculum development, syllabus design, materials and instruction, teaching, assessment, and program evaluation (Norris, 2009). Among these program elements of which TBLT consists, this dissertation is particularly concerned with issues related to syllabus design and pedagogic materials development. Within various approaches to these particular aspects of TBLT, studies based on cognitive demands of tasks, or cognitive task complexity, have become increasingly
popular among researchers who are interested in TBLT for the promotion of L2 acquisition. Tasks with differing degrees of cognitive complexity are claimed to differentially influence learners’ task performances, as well as their communicative interactions, and ultimately their language acquisition. Indeed, for some time now, it has been clear that understanding the nature of cognitive task complexity is presumed to be quite central to the TBLT endeavor, not only for the advancement of the science of the field per se, but concretely as a basis for informing task design, instructional task sequencing, and curriculum and syllabus decisions. Along these lines, Robinson (2001a) argued some 15 years ago that “research into the effects of task complexity on L2 learning … is an area of great consequence for the development of theories of instructed SLA and for pedagogic decisions about grading and sequencing tasks for learners” (pp. 316-7). Here, I will briefly summarize a few of the key sources and reasons for such interest in cognitive task complexity, both from the perspective of educational research in general, and from the perspective of instructed SLA theory and research in particular.

1.2.1 Cognitive task complexity in educational research

In the field of cognitive psychology, Sweller’s Cognitive Load Theory (CLT) (1988, 1994, 2010; Sweller, van Merriënboer, & Paas, 1998; van Merriënboer & Sweller, 2005) has been influential in thinking about how to design instructional materials that are amenable to more efficient learning and retention of knowledge, taking into account learners’ cognitive resources, especially their working memory capacity. As we typically assume from various human experiences, and based on established empirical evidence (e.g., Baddley, 2007), the amount of information that our working memory can retain is limited. In designing instructional materials,
then, it is critical to consider how we can reduce excessive burden on a learner’s working memory to facilitate better and more learning of given instructional content.

CLT assumes three sources of cognitive load: (a) intrinsic, (b) extraneous, and (c) germane (Sweller, 1994). Intrinsic cognitive load refers to a kind of cognitive demand or complexity that is posed by the task content itself. For example, learning about the cardio-vascular system may pose higher intrinsic cognitive load than learning about the law system in the US, for a particular group of learners, such as law majors in US universities; alternatively the same task of learning about the cardio-vascular system may pose lower intrinsic cognitive load for students majoring in Pre-Med. As can be seen in this example, the level of intrinsic cognitive load fluctuates not only by the task content per se, but depending on the learners’ level of content knowledge. Intrinsic cognitive load is claimed to be constant for a given task and for given knowledge levels (to the extent these can be established). It is, therefore, independent from instructional materials design, such as how to present or display learning materials related to the target task, and the only way to change intrinsic cognitive load is to change the task itself or to alter learners’ knowledge levels.

The second source, extraneous cognitive load, is posed by not-so-optimal instructional procedures. Given that the capacity of our working memory is limited, having to attend to information presented on two separate pages of a website while learning about the law system in the US, for example, prevents learners from paying full attention to the learning content. Extraneous cognitive load, for this reason, is a type of cognitive demand that is wasteful and does not contribute efficiently to learning of the given content. Hence, in designing instructional materials, extraneous cognitive load is best reduced, and CLT deals precisely with this issue of
how to go about reducing the amount of extraneous cognitive load to promote better and more learning.

Lastly, germane cognitive load refers to the working memory resources that are devoted to actual learning of the presented materials. Unlike the first two sources of cognitive load, that are determined largely by the learning task itself (e.g., the content, and how it is presented), germane cognitive load is largely dependent on learners themselves. In discussing this type of cognitive load, Sweller (2010) argued that:

Assuming constant levels of motivation, … if intrinsic cognitive load is high and extraneous low, germane cognitive load will be high because the learner must devote a large proportion of working memory resources to dealing with the essential learning materials. If extraneous cognitive load is increased, germane cognitive load is reduced and learning is reduced because the learner is using working memory resources to deal with the extraneous elements imposed by the instructional procedure rather than the essential, intrinsic material. (p. 126)

Using this theoretical distinction, the CLT has been applied to textbook design as well as computer-based and multimedia learning (Brünken, Plass, & Leutner, 2003), and it has contributed greatly to educational practice, in particular pedagogic materials design. What is interesting about this theory of cognition is that the goal in task design is not to reduce cognitive load altogether. Indeed, learning is shown to deteriorate in conditions where learners are both cognitively unchallenged as well as cognitively overwhelmed (van Gog & Paas, 2008). The ultimate goal of CLT then is to reduce extraneous cognitive load and thereby increase germane cognitive load to promote learning. These ideas have been influential especially in the context of
multimedia learning, where learners tend to face cognitive overload due to the massive and
disperse types of input typically available. CLT gives guidelines to material developers as to how
to keep cognitive load at the optimal level and promote better and more learning in this learning
context (e.g., Mayer & Moreno, 2003).

Historically in educational research, another prevalent idea that is relevant to cognitive task
complexity (especially in the areas of curriculum design and assessment) is Bloom’s Taxonomy
of educational objectives (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956). In this taxonomy,
cognitive skills are grouped into six categories: (a) knowledge, (b) comprehension, (c)
application, (d) analysis, (e) synthesis, and (f) evaluation. These categories are organized and
presented in the order of presumed cognitive complexity. Comprehension is thought to be
cognitively more demanding than knowledge, application is thought to be cognitively more
demanding than comprehension, and so on. Some half a century later, a revised version of
Bloom’s Taxonomy was published by his former student, Lorin Anderson, and his colleagues
(Anderson, Krathwohl, Airasian, Cruikshank, Mayer, Pintrich, Raths, Wittrock, 2001). In this
revised version, ‘knowledge’ was renamed as ‘remember,’ ‘comprehension’ was changed into
‘understand,’ ‘application,’ ‘analysis,’ and ‘evaluation’ were kept, but in their verb forms,
‘apply,’ ‘analyze,’ and ‘evaluate,’ and lastly ‘synthesis’ was renamed as ‘create.’ Remember
involves “retrieving relevant knowledge from long-term memory” (Anderson et al., 2001, p. 67),
whereas understand is slightly more complex in that a learner is required not only to “determine
the meaning of instructional messages” (p. 67), but also to express them in their own words.
Apply includes activities, such as “carrying out or using a procedure in a given [often new]
situation” (p. 67), and analyze entails “breaking material into its constituent parts and detecting
how the parts relate to one another and to an overall structure or purpose” (p. 68). Evaluate, then, is about “making judgments based on criteria and standards,” and create refers to activities of “putting elements together to form a novel, coherent whole to make an original product” (p. 68). Another prominent change made in this newer version is that ‘synthesis’ and ‘evaluation’ changed their places; hence ‘create’ is now considered to be the skill that requires the most complex cognitive processing in this taxonomy. The motivation behind this decision was more intuitive than empirical – the authors argued that creating requires cognitive operations, such as “finding things that could fit together, judging their appropriateness, and assembling them to best meet criteria” (p. 294) (i.e., induction), which are thought to be more complex than a simple evaluation of things according to set criteria (Anderson et al., 2001).

Bloom’s Taxonomy of Educational Objectives was proposed and has been used to categorize and better understand what we now call ‘learning outcomes’ in terms of types and degrees of cognitive processes required (Anderson et al., 2001; Krathwohl, 2002). It has also helped teachers to better understand statewide standards (particularly in the US), realize what possible ranges of learning outcomes might be, and design instructional practices and assessments that are based on an easily understood developmental sequence of learning outcomes (Anderson et al., 2001). This taxonomy has also been useful in designing and sequencing pedagogic tasks effectively. Tasks that require cognitively more demanding skills are claimed to entail skills that are lower in cognitive demands (Anderson et al., 2001). For example, before students are able to tackle a task of evaluating a proposal on some controversial topic (e.g., whale hunting), students should first be able to comprehend reading materials (understanding), compare and contrast opposing ideas (understanding), and relate what they have
read to their own opinions and select relevant information that can support their own arguments (applying and analyzing), and so on. For this reason, this Bloom’s Taxonomy has also helped teachers to properly design and sequence tasks and give optimal scaffolding to students in an effort to foster students’ abilities to deal with higher-order thinking skills. This idea of sequencing pedagogic tasks from simple to complex in increasing order of cognitive complexity is also supported by other educational theories (e.g., Elaboration Theory; see Reigeluth, 1979; Reigeluth & Stein, 1983).

As these major examples indicate, the cognitive complexity of learning and performance tasks has played a central role in both research and practice related to educational design. The cognitive responses or actions required by the kinds of complex tasks typical of education provide essential windows into understanding how and why learners react to and potentially benefit from certain tasks in certain ways. In relation to second language instruction and acquisition as well, interest has turned to the cognitive dimensions of valued tasks as a basis for making decisions.

1.2.2 Cognitive task complexity in instructed SLA research

In the field of instructed SLA, the notions of task complexity and difficulty have attracted researchers’ as well as practitioners’ attention as criteria for task design and sequencing. In what follows, I will summarize major proposals on L2 instructional design and sequencing that researchers have put forth to date.

1.2.2.1 Early attempts to use task complexity as a design and sequencing principle

In his well-known Bangalore Project, Prabhu (1982, 1984, 1987) was perhaps the first to utilize the ideas of task difficulty and cognitive challenge in making L2 task sequencing
decisions and in designing pedagogic materials. The Bangalore Communicational Teaching Project was motivated by a dissatisfaction with traditional, structural syllabi, that were designed to teach one linguistic element, such as a grammar point, at a time, and by a movement towards teaching English as a functional tool of communication (Brumfit, 1984). To this end, the new syllabus that Prabhu created and implemented over some five years of experimentation was designed around tasks, which mainly took the form of mentally challenging, interactive, problem-solving activities (in the context of high school English language education in India). For example, according to Prabhu (1987), in selecting classroom activities “it was recognized that negotiation was most likely to take place – and to prove satisfying – when the demand on thinking made by the activity was just above the level which learners could meet without help” (p. 24). Pedagogic materials, then, were designed to present learners with somewhat challenging content, such as having to figure out train schedules, prices, and itineraries, so that they would be cognitively engaged in problem solving and thereby realize incidentally the need to use language for communicating in certain ways related to the problem at hand. On-going pedagogic decisions were likewise dependent on an understanding of the challenge imposed by tasks, as “the teacher’s assessment of the level of difficulty acted as an input to the planning of subsequent lessons” (p. 25), and as “tasks of the same type […] were set on successive days such that each day’s work was similar to but more complex than the previous day’s” (p. 26). The basis for larger-scale instructional sequencing was also relatively intuitive:

[T]asks within a given sequence (i.e. tasks of the same type forming the basis of several lessons) were ordered by a commonsense judgment of increasing complexity, the later tasks being either inclusive of the earlier ones or involving a
larger amount of information, or an extension of the kind of reasoning done earlier.

(Prabhu, 1987, p. 40)

Nevertheless, as intuitive as the approach to task design and sequencing might have been, this was perhaps the first example in the history of instructed SLA where a clear attempt was made to understand and sequence tasks based primarily on the extent to which their characteristics served as a source of mental challenge, rather than merely the type of language that they may elicit.

Candlin (1987) then was likely the first L2 researcher to discuss specifically the cognitive demands of tasks (rather than more subjective notion of task difficulty according to Prabhu) in relation to selection and sequencing decisions in the field of TBLT. Among other possibilities, he argued that tasks should be sequenced as their cognitive load, communicative stress, familiarity, and/or code complexity gradually increase, thereby pushing learners into gradually more target-like language performance demands and outcomes. Similarly, Long and Crookes (1992) argued that tasks should be sequenced from simple to complex, so that they provide a sort of scaffolding towards, and become increasingly more approximate to, the target task that learners are required to accomplish in their L2. The task sequencing decision, then, should be made not simply based on targeted linguistic features (as a structural syllabus does), but according to task characteristics. Long and Crookes (1992) provide the following examples:

The number of steps involved, the number of solutions to a problem, the number of parties involved and the saliency of their distinguishing features, the location (or not) of the task in displaced time and space, the amount and kind of language required, the number of sources competing for attention, and other aspects of the
An intellectual challenge a pedagogic task poses are just a few of the potential grading and sequencing criteria that have been proposed. (p. 45)

Building upon and expanding considerably beyond these early notions of cognitive complexity as a basis for L2 instructional decisions, the two most influential cognitive-interactionist models of TBLT have been proposed by Peter Robinson (2001a) and Peter Skehan (1998). Both models emphasize the importance of cognitive task complexity in pedagogic task design and task sequencing, but each model adopts a unique theoretical background and leads to somewhat distinct predictions for the role of cognitive complexity in task-based performance and, potentially, acquisition.

1.2.2.2 Robinson’s Cognition Hypothesis

Robinson (1995, 2001a, 2001b, 2003a, 2003b, 2005, 2007a, 2007b, 2010, 2011a, 2011b, 2015, Robinson & Gilabert, 2007), on the one hand, has argued that tasks with increased cognitive complexity will (a) encourage L2 learners to pay simultaneous attention to language complexity and accuracy and thereby elicit linguistically more complex and more accurate, although less fluent, language production by posing higher cognitive demands (e.g., conceptual and functional demands) and (b) facilitate interaction and uptake of the linguistic forms that learners’ attention is drawn to, and hence promote language acquisition. This claim, known as the Cognition Hypothesis (CH), is based on an idea that humans have multiple pools of attentional resources available for task execution and that our attentional resources are essentially unlimited. Key theories underpinning the CH include Wickens’ (1984, 2002, 2007) multiple resource theory model, Cromer’s (1973, 1991) Cognition Hypothesis for first language (L1) acquisition, and Givon’s (1985) theory of L1 acquisition.
In an attempt to account for differences in dual-task interference (i.e., why some tasks are easier to handle simultaneously than others), Wickens has proposed that there are multiple pools of attentional recourses available and that we are able to draw attention from distinct pools that are responsible for different aspects of cognitive activities in dealing with multiple tasks. There are three dimensions to his model of multiple resource theory: (a) perceptual modalities, (b) processing codes, and (c) processing stages. Perceptual modalities include (a) visual (e.g., reading) and (b) auditory (e.g., listening). Processing codes consists of (a) verbal or linguistic material (e.g., reading, listening) and (b) non-verbal, spatial material (e.g., driving). Lastly, processing stages can be divided into two major resources: (a) one that underlies perception and working memory, and (b) the other that is the basis for selecting and taking of actions. Key to this theory is the idea that competition for available mental attention is claimed to take place within each pool, not between them. Table 1 below illustrates some concrete examples of easier and more difficult dual-tasks (i.e., competitions) within each pool of attention, based on Wicken’s multiple resource theory.

<table>
<thead>
<tr>
<th>Perceptual modalities</th>
<th>Processing codes</th>
<th>Processing stages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Easier</strong></td>
<td>Reading a book (visual) while listening to a radio (auditory)</td>
<td>Listening to a radio (verbal) while driving (spatial)</td>
</tr>
<tr>
<td><strong>More difficult</strong></td>
<td>Reading while watching a TV (both visual)</td>
<td>Listening to a radio while reading a book (both verbal)</td>
</tr>
</tbody>
</table>

In theory, the more attentional resources we draw from the same pool, the more difficult the task becomes; and as long as we draw from different pools of attention, multi-tasking is easier to
handle. Based on this model, Robinson (e.g., 2001a) contends that L2 learners should be able to pay simultaneous attention to linguistic complexity and accuracy, for example, during the performance of a task (although why paying simultaneous attention to these two linguistic aspects of task performance should draw attention from different pools of attention instead of the same resource pool remains unspecified).

Robinson’s Cognition Hypothesis is also based on Cromer’s (1973, 1991) and Givon’s (1985) L1 acquisition theories. Cromer, in his identically named Cognition Hypothesis, argues that in L1 acquisition “our cognitive abilities at different stages of development make certain meanings available for expression” (italics in the original) and therefore, “it is cognition which determines language acquisition” (p. 246). Robinson (2003b, 2005, 2011b) then claims that this Cognition Hypothesis can also be applied to adults’ second language acquisition. Unlike young children, of course, adult L2 learners already possess highly developed cognitive abilities, and yet, just like young children, they are unable to express complex ideas (produced by their developed cognitive abilities) in their L2 at the early stages of acquisition. A series of tasks that gradually become cognitively more complex, Robinson argues, gives adult L2 learners a chance to express increasingly more complex, elaborated ideas, which in turn is thought to push their interlanguage development.

Additionally, Givon (1985) argues that in L1 acquisition, functional and processing demands motivate young children to move from the pragmatic mode to the syntactic mode. At the early stages of L1 acquisition, young children rely heavily on contexts when using the language, and its form and meaning typically show a one-to-one correspondence. Later, as their cognitive abilities develop, they move to the syntactic mode and start to use more subordinations,
complex verbs, and grammatical morphemes (e.g., plural -s). According to Givon, as young children move to the syntactic mode from the pragmatic mode, they become able to use subject predicate structure instead of topic comment structure and in turn, their language accuracy is claimed to improve. Given that both Cromer’s and Givon’s theories deal with L1 acquisition, how much cognitive challenge or demands a given task poses seems to be closely related to a child’s developmental sequences. In Robinson’s words (in discussing task sequencing), “pedagogic tasks should be sequenced solely on the basis of increases in their cognitive complexity, which mirror the sequences in which children are able to meet the cognitive demands of tasks during L1 acquisition” (p 14).

Putting together these theories from cognitive psychology as well as child language acquisition, Robinson (2001a, 2005) developed his Cognition Hypothesis of language learning and performance. In his model, Robinson argues that task complexity should firstly be distinguished from task difficulty and task conditions (the three dimensions taken together are referred to as the Triadic Componential Framework). Task complexity, on the one hand, is solely dependent on inherent task characteristics (e.g., [+/- few elements], [+/- no reasoning]) and refers to the level of cognitive demands that the task itself requires for successful completion. Task difficulty and task condition, on the other hand, are learner- or situation-dependent factors. Task difficulty refers to learners’ perceptions of how difficult the task is, and task condition refers to interactional factors such as participation (e.g., one-way/two-way, convergent/divergent) and participant variables (e.g., gender, familiarity). Robinson argues that a proactive (i.e., a priori) syllabus design for TBLT should be based solely on task complexity, as

---

1 The plus (+) indicates lower task complexity (i.e., simpler) and the minus (−) indicates higher task complexity (i.e., more complex).
it is the primary dimension that can be controlled by teachers or syllabus designers regardless of learner or situation factors. Task difficulty and task conditions should have little to contribute in terms of task sequencing, but might have considerable impact for on-line decisions about how tasks should be implemented in the classroom (e.g., intentional grouping of students).

In his Triadic Componential Framework, Robinson (2001a, 2005) has described how cognitive task complexity can be operationalized in a predictable, systematic way. According to him, cognitive task complexity can be manipulated along two distinct dimensions: resource-directing and resource-dispersing. As the name suggests, tasks that are complex along the resource-directing dimension are hypothesized to direct learners’ attentional resources to specific aspects of task performance, such as linguistic complexity and accuracy (Robinson, 2003b). Cognitive task complexity along these lines presumably can be manipulated in a variety of ways, for example by decreasing or increasing the number of elements involved in the task (i.e., [+/- few elements]), by requiring reference to events happening here and now or elsewhere in time and space (i.e., [+/- here-and-now ]), by decreasing or increasing reasoning demands of the task (i.e., [+/- no reasoning demands]), or by requiring simply first-person perspective or second- or third-person perspectives (i.e., [+/- perspective-taking]) (Robinson, 2007b). In a picture-based narrative task, for instance, a picture with only one character to be referred to (i.e., [+ few elements]) is considered to be less complex than a picture with nine characters (i.e., [- few elements]). Referring to nine characters, instead of one, is conceptually and linguistically more demanding in general and ostensibly taxes learners’ attentional recourses more heavily in a particular way, by directing their attention to task-relevant forms, such as noun modifiers (Robinson & Gilabert, 2007). When there is only one character involved in the story, learners
can simply refer to him/her by using a simple noun (e.g., a man, a woman). When there are more than two characters involved, however, learners will have to distinguish among them by using more complex language forms, such as relative clauses and pre- and post-modifications of noun phrases. Hence, the complex task may draw learners’ attention to focus on task-relevant, communicatively non-redundant language (e.g., noun modifiers in this case) and, as a result, may promote a focus on linguistic accuracy in their task performance (Robinson, 2001b). Cognitively complex tasks of the resource-directing variety, according to the CH, thus promote an attentional focus on linguistic complexity and accuracy of learners’ task performance. They are also claimed to have some impact on interaction in dialogic or group tasks; that is, cognitively complex tasks may elicit more breakdowns in communication, negotiations of meaning, confirmation checks, clarification requests, and uptake of linguistic forms that learners’ attention is directed to (e.g., the previous example of use of noun modifiers) in a particular task.

By contrast, if we make tasks cognitively complex along the resource-dispersing dimension, for example, by depriving learners of pre-task planning time or relevant prior knowledge or task familiarity, tasks will divert or disperse learner’s attention and memory resources away from linguistic aspects to other information processing activities, such as thinking about what to say and organizing these thoughts, or focusing on maintaining spontaneous online communication (Robinson, 2005). Making tasks complex along these dimensions, therefore, may improve learners’ ability to access and deploy their existing knowledge during task performance (e.g., for fluent, rapid speech); however, it is presumed not to direct learners’ attention to focus on any particular aspects of their task performance or stretch their interlanguage towards further development (Robinson, 2005).
Robinson (2003b, 2010) has recommended the use of these theoretical ideas of cognitive task complexity for classroom application, arguing that in order to foster L2 development, instructional tasks should be sequenced from simple to complex, requiring increasingly more cognitive demands for their successful completion. This idea of task sequencing based on the CH is known as the SSARC model (Robinson, 2010). SS stands for “simple/stabilizing interlanguage,” A stands for “automatizing access to interlanguage,” RC stands for “restructuring and complexifying,” (p. 252), and it is argued that a task should be made cognitively complex by manipulating it both along resource-directing and resource-dispersing dimensions. In his model, the first task should be made cognitively simple along both resource-directing and resource-dispersing dimensions. As an example, Robinson provides a direction-giving map task. If the target task or the final outcome is for L2 learners to be able to give correct and appropriate directions to a driver, we would want to first have them practice giving directions with a map of a small ([+ few elements]), familiar area ([+ prior knowledge]), and with a route already marked ([+ single task]); in addition, we would want to give them time to plan before they engage in the direction-giving task ([+ planning time]). The provision of planning time, the marked route, and the use of a map from a familiar area make the task cognitively simple along the resource-dispersing dimension, and the use of a map of a small area makes the task simple along the resource-directing dimension as well. This is what Robinson would call an SS stage. Moving on to a more complex stage (Stage A), L2 learners would have to complete a similar task of direction-giving, but this time without pre-task planning time. This, as compared to the first task, is cognitively more complex along the resource-dispersing dimension. We can make the task even more cognitively demanding (along the resource-dispersing dimension) by getting rid of the
marked route and by using the map of which learners have no prior knowledge. The last version (Stage RC) then is a direction-giving task with a map that covers a larger area, which poses increased cognitive demands along the resource-directing dimension given a larger number of elements entailed in the map ([− few elements]), without time to plan ([− planning time]), a marked route ([− single task]), or prior knowledge of the area ([− prior knowledge]). As learners engage in progressively more complex tasks with higher functional and communicative demands, they gradually stretch their interlanguage to meet the increased task demands. This, in turn, is argued to have the consequence of facilitating their L2 development (Robinson, 2003b, 2010).

1.2.2.3 Skehan’s Limited Attentional Capacity approach

A somewhat different notion of cognitive task complexity in SLA has been proposed by Skehan (1996, 1998, 2009, 2014, 2015) in the form of his Limited Attentional Capacity approach (LAC). Skehan’s model assumes that human attention as well as working memory are limited in capacity. He argues that “this is a fairly standard account from contemporary cognitive psychology, with a long tradition of relating working memory operations to attentional availability (Baddeley, 2007)” (Skehan, 2009, p. 510). Within the field of SLA, Skehan’s model is also underpinned by VanPatten’s (1990, 1996) Input Processing Theory (Foster & Skehan, 1996). To illustrate, in one experiment, VanPatten (1990) had his Spanish-L2 participants listen to a passage in Spanish in four experimental conditions. The first group was instructed simply to attend to its content, whereas the other groups were asked to pay attention not only to the content of the passage but also to a lexical or grammar item, including a definite article, a verb morpheme, and a content word that is contextually important in comprehending the passage. The
results indicated that the participants, especially with lower proficiency, had difficulty in paying simultaneous attention to meaning (i.e., content of the passage) and form that is not essential for passage comprehension (i.e., a definite article and a verb morpheme). This suggests that L2 learners, if not humans in general, have limited attentional capacity that has to be shared when attending to multiple elements.

Based on his LAC, Skehan (1996, 1998) has suggested analyzing cognitive complexity of tasks based on three principle areas, which are different from Robinson’s Triadic Componential Framework. These include: code complexity, cognitive complexity, and communicative stress, all of which are inherent in the demands made by tasks. Code complexity refers to the level of complexity or difficulty (e.g., relative frequency, developmental stage) of linguistic forms that a task requires for its successful completion. Cognitive complexity in Skehan’s model has to do with the content or meaning of task performance, rather than linguistic aspects, and can be divided into processing and familiarity. Processing refers to online processing that a task requires for its successful completion and is closely related to the intrinsic cognitive demands that a task poses to participants (specifically of task content rather than language). Familiarity, on the other hand, refers to the amount of task-related schematic knowledge or information that task participants possess and can readily use to complete the task at hand. The idea here is that the higher the processing demand is, the higher task complexity becomes, whereas the higher the familiarity is, the lower task complexity becomes. ‘Complexity’ in Skehan’s model, therefore, appears to be very similar to Sweller’s notion of intrinsic cognitive load (that was reviewed earlier). Lastly, communicative stress refers to complexity challenges other than code and meaning, and it is primarily concerned with the degree of pressure to communicate
spontaneously and effectively. Examples include time pressure, communication modality (e.g., speaking vs. writing), scale (e.g., how many participants are involved in a task), stakes (e.g., how important it is to do the task), and control (e.g., whether task goals can be negotiated).

In discussing the effects that increased cognitive task complexity in these three dimensions will likely have on learners’ task performance, Skehan (1998) argued that a conceptually complex task requires so much online processing of the content that not much of the attentional resources can be devoted to linguistic aspects of task performance. Additionally, he suggested that when learners are able to pay attention to linguistic aspects, the three linguistic aspects of task performance, namely syntactic/lexical complexity, accuracy, and fluency (CALS), cannot simultaneously be attended to (Skehan, 1996). When learners are oriented toward accuracy (e.g., producing target-like grammar and vocabulary), restructuring or complexifying their interlanguage is less likely to occur and fluency is likely to decrease. By contrast, when learners attempt to produce complex language or restructure their interlanguage, a great amount of attention is ostensibly consumed in processing the linguistically complex language forms or the new forms to be acquired, and as a result, little attention is left for focusing on accuracy and fluency. Finally, learners who prioritize fluency, say in spontaneous oral production, are likely to give accuracy and complexity lower priority while they seek to keep up with and maintain online and immediate communication. He further argues that it is linguistic complexity and accuracy that tend to compete for attentional resources (when available attention is limited) rather than with fluency (Skehan, 2009). Given these ‘trade-off’ effects related to increases in cognitive complexity, Skehan (1998) has claimed that tasks should be designed and sequenced in a balanced manner in order to foster the development of various aspects of interlanguage (i.e.,
complexity, accuracy, and fluency) in equilibrium. He has argued additionally that sequencing tasks from simple to complex is ideal in that this task sequence does not force L2 learners to devote all attentional resources to conceptualization of complex ideas from the beginning, but allows them to allocate their attention strategically also to language forms.

1.2.2.4 Robinson’s Cognition Hypothesis and Skehan’s Limited Attentional Capacity approach: How are they different?

In a sense, one of the biggest discrepancies between Robinson’s CH and Skehan’s LAC lies in whether linguistic complexity and accuracy as performance goals actually compete for attentional resources or can be attended to simultaneously (Skehan, 2009). On this issue, Skehan has made an interesting observation in Applied Linguistics in 2009. He identified a few cases in previous studies on cognitive task complexity where L2 learners were able to pay attention to both language complexity and accuracy at the same time, and as a result, both dimensions of learner task performance were simultaneously improved. The first study which found such an effect was Foster and Skehan (1999). It investigated the effects of different types of pre-task planning on L2 learners’ task performance, in terms of the CALF indices. Among individual, group-based, and teacher-led planning conditions, learners in the teacher-led planning condition were able to produce more complex and more accurate language than were their counterparts. What this study finding seems to be suggesting is that when some support is available to L2 learners (e.g., through teacher intervention in this case) during pre-task planning time, they are able to allocate their attention more strategically to complexification of ideas and associated language use, as well as to rehearsal, which leads to more accurate use of the language during performance (Skehan, 2009).
The next set of studies, Tavakoli and Skehan (2005), Tavakoli and Foster (2008), and Foster and Tavakoli (2009), found that a structured task that required information integration (e.g., a picture-based narrative task whose storyline was straightforward, clear, and easy to understand and which required L2 learners to integrate the main story with what is going on in the background) was successful in eliciting both complex and accurate language production. Here, it seems to be the case that different aspects of task characteristics have encouraged task participants to draw their attention to accuracy or complexity (Skehan, 2009). Task structure is likely to be closely related to accuracy of language production, because attentional resources left from understanding the storyline can be devoted to producing accurate speech; whereas the need for information integration is more intimately relevant to syntactic complexity, because connecting foreground and background information of a story requires complex language structures (e.g., subordination), such as ‘when’ and ‘while’ (Skehan, 2009). This case is particularly interesting in that it seems to suggest that simultaneous attention to linguistic accuracy and complexity can be made possible if we reduce the extraneous cognitive load of figuring out a storyline (by making the task structure clear) and yet pose a high functional (or facilitative) demand of integrating multiple pieces of information (such as foreground and background information). This idea of increasing functional demands, which in this context can perhaps be related to germane cognitive load in Sweller’s model, while reducing the extraneous cognitive load, seems to be compatible with CLT’s ultimate goal of optimizing cognitive load of a given task for better performance (and learning).

Lastly, Foster and Skehan (2013) have found that L2 learners who were given a post-task activity of transcribing their own utterances produced more complex and more accurate language
than did learners in the control group (with no transcription as a post-task activity) when engaging in a decision-making task. To this end, Skehan (2009) argued that “the change in task performance seems to be the result of an interpretation by the task participants of what they should do,” hence “the participants seem to be the ones … who shift their goals and prioritize certain performance areas” (p. 523). Given that both language complexity and accuracy are increased at the same time in the studies mentioned above, at first glance, it seems to provide empirical evidence for Robinson’s CH and discredit Skehan’s LAC. However, it is not increased cognitive task complexity, Skehan (2009) argues, that motivated this simultaneous attention to language complexity and accuracy, but it is a pedagogical intervention (e.g., teacher-led planning), optimal levels of conceptual and linguistic demands (e.g., a structured task design that requires information integration), and learner choice induced by a task condition (e.g., a post-task activity of transcribing one’s utterances) that fueled their ability to pay simultaneous attention to linguistic complexity and accuracy. It is important to note here that Skehan’s argument laid out in this 2009 article emphasizes the influence of factors in addition to inherent cognitive task complexity, such as pedagogical intervention and learner choice, on L2 task performance. This is in contrast to Robinson’s model that portrays cognitive task complexity in itself as the primary factor in thinking about the relationship between tasks and learner L2 performance.
1.3 Problem statement

Likely due to their contrasting proposals (and of course in addition to their importance in the field), Robinson’s CH and Skehan’s LAC have inspired a remarkable body of research on the effects of increased or decreased cognitive task complexity on L2 task performance and language acquisition over the past two decades. Sasayama, Malicka, and Norris (2013, 2014, 2015), for example, identified in their research synthesis at least 129 studies published in refereed journals (never mind those in other publication venues, like book chapters and dissertations) on the topic of cognitive task complexity over the past 25 years. Probably the first study of this kind within SLA was published in 1987 (an article by Ellis on the effect of planning time), predating any substantial theorization of the sort described above, and clear growth and sustained interest in the topic have been observed since the mid-2000s (i.e., after the efforts of Robinson and Skehan to provide theoretical accounts of cognitive task complexity). However, despite these widespread interests, it has proven to be relatively challenging to identify straightforward answers to questions regarding the effects of cognitive task complexity on L2 performance, interaction, and learning (though within quite specific task complexity operationalizations, task types, and research methodologies, greater consistency has emerged, e.g., Robinson, 2011a; Skehan, 2009, 2014). This challenge is apparent in fundamentally inconsistent findings across studies in the way that presumed cognitive demands lead to observable changes in performance or learning. There are, of course, several possibilities for this lack of consensus, including the influence of learner variables (e.g., L2 proficiency) and task design variables, such as task types or task modality; but above all, a major concern has to do with the actual versus designed influence of the independent variable, that is, cognitive task complexity, as has been noted by Robinson
In other words, even though tasks might have been designed to differ in complexity, it is possible that there may not have been actual cognitive difference between the designed-to-be simple and designed-to-be complex conditions for the given tasks under investigation and for the given group of learners in the many studies carried out so far.

Research into the effects of cognitive task complexity is important not only for the sake of advancing our understandings of the phenomena involved and clarifying theoretical positions, but also in that, if carried out properly, it may have considerable potential for offering useful implications about L2 educational practices. Indeed, the desire for such implications (e.g., syllabus design, pedagogic task selection and sequencing) was the starting point for much of the related theory and research in the first place. Furthermore, we have already seen attempts at utilizing theories of cognitive task complexity for educational decision making in TBLT-informed language programs. For example, in designing a TBLT syllabus, the Korean Program at the University of Hawai‘i chose tasks based on a thorough needs analysis and then sequenced a group of tasks (used over a few lessons) from simple to complex by increasing the number of elements involved in each task (i.e., [+/- few elements]) and making other demands (e.g., interactional) gradually closer to the real-world, target task as the sequence progressed (Chaudron, Doughty, Kim, Kong, Lee, Lee, Long, Rivers, & Urano, 2005). A commercial task-based textbook, Widgets (Benevides & Valvona, 2008), is another example that has adapted the idea of cognitive task complexity for sequencing instructional materials. Starting with a self-introduction stage contextualized in the business setting, where students have just started working at a company and are attending the first employee orientation (Stage 1: “Welcome
aboard”), tasks become increasingly cognitively complex. The second stage, “Eureka!,” requires learners to brainstorm new product ideas, and in Stage 3: “Decisions, decisions”, they engage in a debate and choose the best products for each group. Stage 4, “The customer is always right,” then asks learners to conduct a market research on a new product (chosen in Stage 3), whereas Stage 5, “As seen on TV,” is a task where students create and present an infomercial for the new product. Stage 1 is claimed to be the simplest of all stages, followed by Stage 2, Stage 3, and Stage 4/5 sequentially (Pearson Education Asia, n.d.). Although in-depth explanations of what makes each stage more complex than the previous are not provided, we do see clear evidence to support that the theoretical idea of cognitive task complexity is already making its way into practice. For a third example, in the language testing context, Norris, Brown, Hudson, and Yoshida (1998) (see also Brown, Hudson, Norris, & Bonk, 2002; Norris, Brown, Hudson, & Bonk, 2002) developed and investigated task-based performance assessment instruments for English as a second language learners, adapting Skehan’s framework of cognitive task complexity as a basis for predicting performance demands and likelihood of success for L2 learners.

As these, among other, examples demonstrate, illuminating issues related to the actual cognitive complexity of pedagogic, assessment, and other tasks is important to the TBLT field and instructed SLA: Researchers and practitioners alike need a thorough accounting of what, exactly, makes a task more or less complex for L2 learners. At a minimum, providing such an account should help us (a) begin to resolve heretofore uninterpretable or ambiguous study findings (i.e., due to uncertainties about the amount and type of cognitive complexity that given tasks pose); (b) thereby advance our understandings of the effects of cognitive task complexity
on L2 performance, interaction, and learning; and ultimately (c) provide more accurate evidence-based suggestions on task-based L2 educational practices, such as syllabus design, task selection, and assessment, in particular. Accordingly, this dissertation explores possibilities for clarifying what aspects of tasks influence actual (i.e., as realized by the learners) versus designed cognitive complexity, highlighting procedures that should allow us to detect the effects of inherent or manipulated dimensions that constitute tasks, and in turn enabling the more accurate interpretation of L2 learners’ task performance (and learning) as a result of engaging in tasks.

1.4 Definitions of constructs

In this dissertation, a variety of methods are used to attempt to tease apart the cognitive complexity of a task as it is realized by learners engaging in it. Accordingly, the following constructs play a particularly crucial role in understanding this study: (a) cognitive task complexity, (b) cognitive load, (c) task difficulty, and (d) mental effort. Cognitive task complexity in this paper is defined as a multi-dimensional construct. In the TBLT literature, it is predominantly referred to as ‘task complexity’ without the word ‘cognitive;’ however, it cannot reside merely in the task itself. That is, there always is an interaction between the material (task) and someone’s brain or cognition, and this interaction is exactly what we are interested in. Indeed, proponents of Cognitive Load Theory (e.g., Kalyuga, 2007; Kalyuga, Chandler, & Sweller, 2000) argue that intrinsic cognitive load of a task at hand cannot be determined without taking into account who the learners are. In other words, the exact same task poses distinct degrees of cognitive load on learners with different characteristics (e.g., different levels of knowledge of the task content). Based on this idea prevalent in cognitive psychology, cognitive task complexity in this dissertation is defined as the inherent cognitive demands of a task that are
realized in interaction with learner characteristics. The term *cognitive load* then is used more often in cognitive psychology to mean essentially the same thing: the amount of cognitive processing required for a certain task (Sweller, 1988). The two terms are used synonymously in this paper, though *load* is primarily used in relation to the psychological measures that assess the degree of cognitive task complexity (e.g., dual-task methodology, time estimation, self-ratings), as in measures of ‘cognitive load.’

The construct of cognitive task complexity or cognitive load is contrasted with *task difficulty* in Robinson’s model (2001a, 2005, 2011a), which refers to learners’ perceptions of how difficult a given task is to perform. In Robinson’s model, task difficulty is primarily concerned with learners’ individual difference factors, such as motivation, anxiety, confidence, working memory capacity, intelligence, and aptitude. It is, thus, used to account for why the same task is perceived to be more or less difficult by different individuals. In the current study, task difficulty is treated as purely perceptual and hence irrelevant to the actual attentional resources or mental effort exerted in doing a given task. In that sense, it is principally, if not entirely, referencing the task itself, however *subjective* the estimation might be. *Mental effort*, on the other hand, is a term used in cognitive psychology in relation with Sweller’s (1988) Cognitive Load Hypothesis. It is defined as “the amount of resources actually allocated to accommodate the task demands” and reflects “the amount of controlled processing in which the individual is engaged” (Paas, van Merriënboer, & Adam, 1994, p. 420). Mental effort, then, may be related to perceptions of task difficulty, but it is distinct in reflecting the actual effort engaged by the individual as opposed to the supposed demand imposed by task. The focus here is on the individual, rather than the task itself. A perceived-to-be difficult task does not necessarily
guarantee that increased mental effort will be exerted, and individuals may intentionally devote greater or lesser mental effort despite the inherent task demands. A graphic depiction of these four constructs is found in Figure 1.

![Figure 1. Relationship between ‘cognitive task complexity,’ ‘cognitive load,’ ‘task difficulty,’ and ‘mental effort.’](image)

**1.5 Organization of the dissertation**

This dissertation consists of eight chapters. The first chapter has provided a general introduction to TBLT, and specifically addressed the notion, importance, and application of cognitive task complexity in education in general and instructed SLA in particular. Of special interest is a detailed review of Robinson’s Cognition Hypothesis and Skehan’s Limited Attentional Capacity approach, the two most influential cognitive-interactionist models of TBLT, and the subtle differences that lie between these models. This chapter then introduced a few of the challenges associated with cognitive task complexity research conducted in instructed SLA to date. It finally provided definitions of four important constructs: (a) cognitive task complexity, (b) cognitive load, (c) task difficulty, and (d) mental effort.
Chapter 2 reviews the effects of cognitive task complexity on L2 task performance. It firstly presents distinct operationalizations of cognitive task complexity used in previous studies to better understand the construct as portrayed by other researchers. Then, selected studies that have investigated task complexity effects based on Robinson’s and Skehan’s theoretical frameworks are summarized in an attempt to decipher any patterns in the roles played by cognitive task complexity. However, given largely inconsistent findings, the main argument introduced here revolves around the idea that uncertainties about the actual cognitive complexity difference between tasks used in experiments have to date limited considerably our interpretations and understandings of the real effects of cognitive task complexity on L2 performance. Hence, the chapter argues that the assumption of differences in cognitive task complexity between designed-to-be-simple and designed-to-be-complex tasks must be verified empirically somehow. To further advance our understandings of the nature of cognitive task complexity, it is also argued that an investigation of why a task is simple or complex, too, is an integral part of the endeavor.

Chapter 3 introduces various possibilities for the independent measurement of cognitive load, utilized in the fields of cognitive psychology, multimedia learning, and SLA. Measures introduced include direct and indirect measures of cognitive load as well as objective and subjective measures. It will be argued that these measures can complement each other and that, especially given only recent introduction of the cognitive load measures to the domain of TBLT, it is advisable to triangulate the data by using multiple measures, rather than a single measure, of cognitive load. Data triangulation is critical in correctly interpreting the data at hand and
advancing our understandings not only of the tasks but also of the measures themselves as they are put to use in the domain of TBLT.

Chapter 4 firstly reviews if and how L2 proficiency has been measured in SLA research and in the domain of cognitive task complexity in particular. It then discusses the key moderating role of proficiency in task complexity research by introducing some example studies that have found an important moderating role for proficiency in understanding the effects of cognitive task complexity on L2 performance and learning. The chapter then discusses how L2 proficiency might be measured effectively in the context of L2 research. To this end, as a practical measure of L2 proficiency, two examples of short-cut proficiency estimates (i.e., a cloze test and an elicited imitation test), which are suitable in time-constrained L2 research settings, are introduced.

Chapter 5 sets up the current study by discussing its rationale, what unique contributions the study makes to TBLT, and research questions that this dissertation poses. Then, it describes the research methodology, including participants, materials and instruments, procedures, and measurement and data analysis.

Chapter 6 reports on the quantitative findings from the cognitive load measures for both native and L2 speakers of English. A report of overall patterns in the data will be followed by an analysis by L2 proficiency groups. The chapter then explores participants’ perceptions of each task, indicating what factors contributed to the complexity of task performances. Lastly, the effects of measured cognitive complexity on native speaker and learner task performance will be presented in terms of syntactic complexity, accuracy, lexical variety, and fluency indices.
Chapter 7 discusses the results reported in Chapter 6, by emphasizing important patterns that respond to the original research questions. It is divided into two main sections: (a) validating the assumed relationship between task design and the level and source of cognitive complexity, and (b) interpreting the effects of measured cognitive complexity on task performance. Section 1 highlights the following four important observations: (a) the relationship between learner factors and the level of cognitive complexity, (b) the relationship between the number of elements and the level of cognitive complexity, (c) types of cognitive demands, and (d) differences between the four cognitive load measures themselves. Section 2, then, focuses on: (a) what task design features contributed to improved performance, (b) proficiency effects, and (c) consideration of the contribution of each of the performance measures per se.

Finally, Chapter 8 concludes this lengthy endeavor (i.e., the dissertation) by raising limitations and by offering concluding remarks, theoretical and practical implications, and suggestions for future research.
CHAPTER II: COGNITIVE TASK COMPLEXITY AS A CAUSAL MECHANISM IN SECOND LANGUAGE PERFORMANCE

2.1 Introduction

As introduced in the previous chapter, cognitive task complexity has attracted a lot of theoretical attention, empirical investigation, and pedagogical interest from researchers and practitioners (e.g., teachers, material developers) working not only in the field of instructed SLA, but in education in general. This chapter will focus specifically on cognitive task complexity research conducted in relation to instructed SLA and review briefly how cognitive task complexity is proposed to influence L2 learners’ second language performance. While it is beyond the scope of this chapter (or this dissertation) to synthesize existing research on this topic (an enormous endeavor, given the rapid accumulation of research in the domain; but see Sasayama et al., 2013, 2014, 2015, forthcoming), the focus here will be on illustrating and exemplifying the characteristics of research that has been conducted to date. It also explores challenges that delimit what we know so far about the relationship between cognitive task complexity and L2 performance, based on accumulated empirical evidence in a few example domains of inquiry.

2.2 Predicted effects of cognitive task complexity on L2 performance and learning

As summarized earlier, Robinson’s (2001a) Cognition Hypothesis and Skehan’s (1998) Limited Attentional Capacity approach make distinct predictions about the effects of cognitive task complexity on L2 task performance and potentially on opportunities for L2 acquisition. To recap, Robinson hypothesizes that a cognitively complex task, manipulated along the resource-directing dimension (e.g., [+/- few elements]), should have positive effects on accuracy and
complexity because it poses higher functional demands (e.g., distinguishing among similar characters), and thereby it requires L2 learners to respond somehow, and if successful, to accurately use more complicated sentences, such as relative clauses, for its task completion. On the contrary, fluency is expected to decrease as the degree of cognitive demands increases and attention is diverted to other aspects of performance. An increase in fluency as performance target might be garnered by manipulating the resource-dispersing dimensions of tasks, such as adjusting the amount of planning time available. Skehan, on the other hand, argues that L2 learners are usually unable to pay simultaneous attention to multiple aspects of task performance in general. When cognitive complexity is increased, for example, by depriving time to plan, L2 learners are required to devote more attentional resources to conceptualization of ideas during task performance, and as a result, they will not be able to pay much attention to linguistic elements. Under such a condition, paying simultaneous attention to accuracy and complexity is very difficult, if not impossible, as they tend to compete for attentional resources. As each model adopts a unique theoretical background and leads to distinct predictions for the role of cognitive complexity (as reviewed in Chapter 1), there has ensued an extensive debate in the field on which model receives more empirical support. These two models of cognitive task complexity, as a result, have inspired numerous research studies. However, despite the accumulated interests in cognitive task complexity research over the past two decades or so, we know surprisingly little about the effects of cognitive task complexity on L2 task performance (and even less so about the relationship with L2 learning, given a lack of studies that investigate that particular relationship).

\[ ^2 \text{Note that this dissertation is agnostic to answering this question and instead it is interested in exploring a way forward in researching the core construct at stake (i.e., cognitive task complexity).} \]
2.3 Understanding the status quo

Reviewing previous studies investigating the role of cognitive task complexity in instructed SLA, what stands out the most is that the (supposedly) same construct of cognitive task complexity (i.e., the construct implied within a corresponding theory) has been operationalized in numerous and quite distinct ways. Table 2 below summarizes the most common task complexity operationalizations used in the domain so far, along with some representative studies for each. The operationalizations are presented in the order of frequency, according to Sasayama et al.’s (2013, 2014, 2015) research synthesis (focusing on studies published as refereed journal articles). As can be seen in the table, L2 cognitive task complexity has been operationalized in at least 14 different ways, based on Robinson’s CH and Skehan’s LAC, as well as Sweller’s CLT, and studies often combine these basic features into multi-factorial designs (e.g., learners do tasks that have both a +/- planning time and a +/- few elements dimension as seen in Sasayama & Izumi, 2012). Each of these operational designs leads, presumably, to a specific type of effect on the L2 learner’s cognitive state during task engagement, which in turn leads to putative effects on language use in task performance and greater or lesser degrees of task success. That is, in each such study, more (+) or less (−) of the given feature is presumed to induce a particular cognitive response and subsequent performance effects.
<table>
<thead>
<tr>
<th>Task complexity operationalization</th>
<th>Example studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>+/- few elements</td>
<td>Robinson (2001b), Xi (2010), Sasayama (2011), Sasayama &amp; Izumi (2012), Gilabert (2007b)</td>
</tr>
<tr>
<td>+/- simple storyline</td>
<td>Tavakoli &amp; Foster (2008), Tavakoli (2009)</td>
</tr>
<tr>
<td>+/- no time pressure</td>
<td>Skehan &amp; Foster (1999), Maad (2010)</td>
</tr>
<tr>
<td>+/- no perspective-taking</td>
<td>Iwashita et al. (2001), Elder et al. (2002)</td>
</tr>
<tr>
<td>+/- integrated format****</td>
<td>Al-Shehri &amp; Gitsaki (2010)</td>
</tr>
</tbody>
</table>

Note. The plus (+) indicates lower task complexity (i.e., simpler) and the minus (−) indicates higher task complexity (i.e., more complex). *Manipulated in combination with [+/- few elements]; **A task with clear sequential structure refers to a task with “events that [have] a clear time sequence from beginning to middle and end” (Tavakoli & Foster, 2008, p. 44); ***A dual mode input (e.g., audiovisual) is contrasted with a single mode input (e.g., audio only); **** The integrated format presents learning materials all in one place, whereas the split-attention format requires learners to go to various places to gather information needed to complete a given task.
There are three fundamental challenges to the interpretability of these distinct operationalizations of task complexity and hence to the generalizability of any given study in relation to the theoretical premises of interest. First, each task complexity feature may or may not pose a cognitive demand of a particular kind; for example, having planning time or being familiar with a set of task procedures will likely reduce the complexity of an L2 task in ways that are distinct from having to tell a simple story, with few characters. Hence, the type of cognitive complexity associated with the various operationalizations likely differs, the resulting impact on learners’ cognitive responses will differ, and subsequent L2 performance will differ as well. Although some basic distinctions have been proposed, for example, by Robinson’s CH (in the forms of ‘resource-directing’ dimensions and ‘resource-dispersing’ dimensions), this crucial difference is often lost in the haste by researchers to make claims about greater or lesser task complexity and its effects on performance (i.e., regardless of type). Second, the degree of cognitive complexity associated with a given feature is ambiguous – as the popular [+/-] designation suggests, assumptions regarding amount of cognitive challenge are typically reduced to relative dichotomies (i.e., one task condition has less challenge than the other; a task is ‘simple’ or it is ‘complex’). Such relative dichotomous distinctions within a single study already beg the question of ‘how much complexity difference is enough to result in a salient effect on learner cognition and performance?’; It is all the more ambiguous to attempt comparisons across studies of distinct complexity features with differing degrees of difference between ‘simple’ and ‘complex’ tasks. Third, while the particular feature of interest may be designated as the causal mechanism underlying differences between task conditions, L2 communication tasks are notoriously multi-faceted, and it is unclear to what extent the designed complexity
operationalization is in fact the primary cause of learners’ responses, as opposed to other possible competing factors (e.g., task conditions, learner factors). The upshot of these three challenges is that it has proven extremely difficult to date to compare and combine study findings in attempting to decipher any emerging patterns in the effects of cognitive task complexity on L2 performance and learning.

Now, we might speculate that if we compare studies within the same specific complexity operationalization, a clearer pattern will emerge. Table 3 below summarizes characteristics and results of studies that operationalized task complexity by the feature [+/- few elements] and in turn investigated its effects on L2 performance in the form of measures of linguistic complexity, accuracy, lexis, and fluency (CALF indices). The feature [+/- few elements] has been the most common operationalization along Robinson’s resource-directing dimensions (Sasayama et al., 2013, 2014, 2015), and the CALF indices have been used in the majority of studies that adopted this particular operationalization. This relatively consistent replication of causal and outcome variables was thought to allow for substantial comparability across studies. However, as can be seen in Table 3, even within the same variable of [+/- few elements], study findings have revealed different, and sometimes contradictory indications regarding the effect of increased cognitive task complexity. For some studies, having learners deal with a greater number of elements in the L2 task leads to an increase in accuracy of language use; but in other studies the same phenomenon seems to have no effect on accuracy. For grammatical complexity, there have been positive, null, and negative effects observed for the same feature. The picture for fluency, on the contrary, seems a bit clearer: Increased cognitive complexity (along the [+/- few elements] variable) had negative effects on fluency in all of these studies that measured it. One
possibly interesting pattern from the perspective of theoretical debates mentioned above is that none of these studies has shown a simultaneous increase in syntactic complexity and accuracy when cognitive task complexity is increased, although two studies have shown that accuracy and lexical diversity can go up together. Still, perhaps the main takeaway point is that, as a group, these studies have yielded somewhat inconsistent results on the effects of cognitive task complexity on the CALF indices (especially the first three constructs), with no definitive patterns in the way that presumed cognitive demands lead to qualitative changes in L2 task performance. At stake for the advance of science in this field, then, is to be able to answer the question ‘Why?’. Could it be that differences in task conditions are to blame, with modality differences or monologic/dialogic dimensions of the tasks causing undue influence on performance? Or could it be that learner factors, such as L2 proficiency, interacted with task performance in unpredicted ways? Alternatively, could it be that the presumed types and degrees of cognitive complexity imposed by the operationalizations in each study differed to the extent that they impacted learners’ cognitive responses in such a way as to influence L2 performance differentially? Clearly, these questions must be resolved in order for research within this domain to begin to provide interpretable and generalizable answers to its core theoretical questions.
<table>
<thead>
<tr>
<th>L2</th>
<th>Proficiency</th>
<th>Age</th>
<th>Task type</th>
<th>Individual/pair</th>
<th>Mode</th>
<th>Results (according to authors)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robinson (2001b)</td>
<td>English</td>
<td>Adults</td>
<td>Direction-giving</td>
<td>Pair</td>
<td>Oral</td>
<td>= = + –</td>
</tr>
<tr>
<td>Gilabert (2007b)</td>
<td>English</td>
<td>Adults</td>
<td>Decision-making</td>
<td>Individual</td>
<td>Oral</td>
<td>X = X X</td>
</tr>
<tr>
<td>Michel, Kuiken, &amp; Vedder (2007)</td>
<td>Dutch</td>
<td>Adults</td>
<td>Decision-making</td>
<td>Individual/pair</td>
<td>Oral</td>
<td>= + + –</td>
</tr>
<tr>
<td>Sasayama (2011)</td>
<td>English</td>
<td>Adults</td>
<td>Picture-based narrative</td>
<td>Individual</td>
<td>Oral/written</td>
<td>+ = X X</td>
</tr>
<tr>
<td>Sasayama &amp; Izumi (2012)</td>
<td>English</td>
<td>High school students</td>
<td>Picture-based narrative</td>
<td>Individual</td>
<td>Oral</td>
<td>= – = –</td>
</tr>
</tbody>
</table>

Note. C = Syntactic complexity, A = Accuracy, L = Lexical variety, F = Fluency; + indicates that increased complexity had a positive effect, = indicates that increased complexity had a neutral effect; - indicates that increased complexity had a negative effect; X indicates that the measure was not investigated;
For a second example, where perhaps more stable patterns have begun to emerge, Table 4 summarizes characteristics and results of studies that used a narrative task, operationalized task complexity by providing or not providing pre-task planning time, and investigated its effects on CALF indices. This combination of variables was chosen to allow comparisons of as many studies as possible, all sharing the common causal design feature of [+/− planning time]. The table specifies the length of planning time, type of planning (guided, unguided, or both), age of participants, target L2, interactional requirements of the task (pair or individual), mode of the task (oral or written), and results of each study in terms of CALF indices. These features, I believe, are all straightforward, perhaps with the exception of the type of planning. Guided planning in this context refers to a type of planning where learners receive detailed guidance as to what to focus on during the pre-task planning time. For example, in Mochizuki and Ortega’s (2008) study, participants in the guided planning condition received a handout which explained how to use relative clauses and were told that the use of relative clauses may be helpful for their effective story-telling. On the other hand, in an unguided planning condition, learners are simply asked to plan what they are going to say without any further instructions.

As can be seen in Table 4, these pre-task planning studies, especially the ones that gave 10 minutes of planning time, yielded surprisingly consistent results. None of these studies found a positive effect of pre-task planning on lexical variety, whereas almost all of the studies that operationalized pre-task planning as unguided, 10-minute planning found advantageous effects on syntactic complexity and fluency. The results of accuracy, however, are mixed: some studies have found a positive effect, and others have found a neutral effect on accuracy. Another clear pattern that we can observe is that none of the studies included here showed a negative effect of
pre-task planning on any of the syntactic complexity, accuracy, lexical variety, or fluency indices. The pre-task planning domain provides a window into the possibility that greater control and replication will lead to more consistency across study findings. Here, the challenge of the degree of causal variable can at least be controlled by design, with greater or lesser amounts of planning provided. Still, it is also clear that ambiguities remain. To what extent is this design feature leading to actual differences in cognitive demand on learners? How does planning time interact with the other dimensions of the task in creating particular types of cognitive challenge, deployment of attention, etc.? To what extent does an individual learner’s strategic use of planning time potentially increase versus decrease the complexity of the task, and to what extent is that increase or decrease in complexity related to benefits or detriments for performance? Here again, clarification of the actual impact of this task complexity feature on L2 learners’ cognitive responses seems imperative, if the field is to advance our understandings of the actual causal relationship at play.
<table>
<thead>
<tr>
<th>Study</th>
<th>Length of planning</th>
<th>Type of planning</th>
<th>L2</th>
<th>Proficiency Description</th>
<th>Age</th>
<th>Individual/Pair</th>
<th>Mode</th>
<th>Results (according to authors)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foster &amp; Skehan (1996)</td>
<td>10 min</td>
<td>Unguided</td>
<td>English</td>
<td>Pre-intermediate</td>
<td>Adults</td>
<td>Pair</td>
<td>Oral</td>
<td>+ + X +</td>
</tr>
<tr>
<td>Foster &amp; Skehan (1996)</td>
<td>10 min</td>
<td>Guided</td>
<td>English</td>
<td>Pre-intermediate</td>
<td>Adults</td>
<td>Pair</td>
<td>Oral</td>
<td>+ = X +</td>
</tr>
<tr>
<td>Ortega (1999)</td>
<td>10 min</td>
<td>Unguided</td>
<td>Spanish</td>
<td>Advanced M = 493.1</td>
<td>Adults</td>
<td>Individual</td>
<td>Oral</td>
<td></td>
</tr>
<tr>
<td>Iwashita, McNamara, &amp; Elder (2001); Elder,</td>
<td>3 min</td>
<td>Unguided</td>
<td>English</td>
<td>Institutional TOEFL</td>
<td>Adults</td>
<td>Individual</td>
<td>Oral</td>
<td></td>
</tr>
<tr>
<td>Yuan &amp; Ellis (2003)</td>
<td>10 min</td>
<td>Unguided</td>
<td>English</td>
<td>8 years of previous</td>
<td>Adults</td>
<td>Individual</td>
<td>Oral</td>
<td></td>
</tr>
<tr>
<td>Ellis &amp; Yuan (2004)</td>
<td>10 min</td>
<td>Unguided</td>
<td>English</td>
<td>M = 460.86 - 447.78 TOEFL</td>
<td>Adults</td>
<td>Individual</td>
<td>Written</td>
<td></td>
</tr>
<tr>
<td>Mochizuki &amp; Ortega (2008)</td>
<td>5 min</td>
<td>Unguided</td>
<td>English</td>
<td>Beginner</td>
<td>High school students</td>
<td>Individual</td>
<td>Oral</td>
<td></td>
</tr>
<tr>
<td>Mochizuki &amp; Ortega (2008)</td>
<td>5 min</td>
<td>Guided</td>
<td>English</td>
<td>Beginner</td>
<td>High school students</td>
<td>Individual</td>
<td>Oral</td>
<td></td>
</tr>
<tr>
<td>Tavares (2009)</td>
<td>10 min</td>
<td>Unguided</td>
<td>Intermediate</td>
<td></td>
<td>Adults</td>
<td>Individual</td>
<td>Oral</td>
<td></td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>Length of planning</th>
<th>Type of planning</th>
<th>L2</th>
<th>Proficiency</th>
<th>Age</th>
<th>Individual/ Pair</th>
<th>Mode</th>
<th>Results (according to authors)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Farahani &amp; Meraji</td>
<td>10 min</td>
<td>Unguided</td>
<td>English</td>
<td>Intermediate</td>
<td>Adults</td>
<td>Individual</td>
<td>Written</td>
</tr>
<tr>
<td>Meraji (2011)</td>
<td>10 min</td>
<td>Unguided</td>
<td>English</td>
<td>Intermediate</td>
<td>Adults</td>
<td>Individual</td>
<td>Written</td>
</tr>
<tr>
<td>Mohammadzadeh, Dabaghi, &amp; Tavakoli (2013)</td>
<td>5 min</td>
<td>Unguided</td>
<td>English</td>
<td>Lower intermediate</td>
<td>Adults</td>
<td>Individual</td>
<td>Written</td>
</tr>
</tbody>
</table>

*Note. C = Syntactic complexity, A = Accuracy, L = Lexical variety, F = Fluency; + indicates that pre-task planning had a positive effect; = indicates that pre-task planning had a neutral effect, +/- indicates mixed findings; X indicates that this particular aspect was not investigated; *Local syntactic complexity (i.e., number of relative clauses), but not global complexity, was positively affected by guided pre-task planning;
In sum, based on these observations of how task complexity has been investigated to date, if we are to better understand the role of cognitive task complexity as a causal mechanism, then we need to ensure that a designed-to-be complex task is indeed meaningfully more cognitively complex than a designed-to-be simple task for the particular task under investigation and for the particular group of learners involved. Particularly in light of the diverse operationalizations of cognitive task complexity in Table 2, even though the tasks were designed to differ in complexity, it is possible that there may not have been actual cognitive differences between the designed-to-be simple task and designed-to-be complex task within many of these studies. To resolve these ambiguities and advance our understandings of the phenomenon, the type and degree of cognitive task complexity – that is, the key presumed causal factor in this domain of research – must be verified somehow (Norris, 2010; Robinson, 2001a). Additionally, to fully understand the nature of cognitive task complexity, it is imperative to confirm by what factors cognitive task complexity is caused. To this end, we have to investigate why exactly a measured-to-be a complex task is complex for a given learner group; more specifically, we need to know whether the type and degree of cognitive task complexity are indeed caused by the task design feature(s) per se or they are (also) a product of other task design, instructional, and/or learner factors. Currently, then, it is premature to reach a conclusion about the effects of cognitive task complexity on L2 task performance and learning. Only after we take care of various moderating factors and account for the complex causal mechanism putatively at play (e.g., by carefully measuring the actual level of cognitive task complexity independently of task performance, by better understanding what makes a task cognitively more or less demanding) can we fully investigate the ‘pure’ effects of cognitive task complexity on L2 task performance and learning.
2.4 Chapter summary

Chapter 2 firstly presented numerous distinct operationalizations of cognitive task complexity used in previous studies to better understand the construct as investigated by other researchers. Then, it reviewed the effects of cognitive task complexity on L2 task performance theorized by Robinson’s (2001a) Cognition Hypothesis and Skehan’s (1998) Limited Attentional Capacity approach. These two models of cognitive task complexity make distinct predictions: The former predicts that a complex task has positive effects on accuracy and complexity, while fluency is expected to suffer as a result; whereas the latter predicts that a complex task makes it more difficult than a simpler counterpart to pay simultaneous attention to multiple aspects of task performance, such as accuracy and complexity. When looking at representative studies that have operationalized cognitive task complexity based on [+/- few elements], it was shown that these studies have yielded somewhat inconsistent findings on the effects of cognitive task complexity on the CALF indices (although one exception might be that fluency seems to be negatively affected by an increase in cognitive complexity). On the other hand, representative studies that operationalized cognitive task complexity by provision or deprivation of pre-task planning time have shown that pre-task planning seems to be effective in eliciting syntactically complex as well as fluent language production – a much more consistent result than the previous domain, although the actual impact of this task complexity feature on L2 learners’ cognitive responses remains ambiguous (as is the case for the other domain). It was therefore argued that these ambiguities may be closely related to uncertainties about the designed versus actual cognitive task complexity, in terms of degree and type of actual impact on learners’ cognitive responses to tasks. These uncertainties have severely limited our understandings of not only the effects of
cognitive task complexity on L2 performance and learning but also on the fundamental premise that learners respond cognitively in the ways that researchers presume. Hence, it is as yet premature to try to test any related theories before solving the fundamental issue of the assumed cognitive complexity differences associated with task design. Only after we have accumulated enough studies that have verified, rather than assumed, the cognitive complexity differences among two or more tasks, can we start to interpret the ‘true’ effects of cognitive task complexity.
CHAPTER III: MEASURING COGNITIVE TASK COMPLEXITY

3.1 Introduction

Researchers who are interested in the effects of cognitive task complexity on L2 performance, interaction, and development have generally assumed that differing degrees of cognitive task complexity can be inferred through observations of differing qualities in linguistic production elicited by the tasks. At the same time, however, they have justified that the supposedly complex task actually poses high cognitive demands on learners by arguing that task performance on the complex task is qualitatively different from performance on the simpler task by design. To get beyond the challenges of this somewhat circular reasoning, we need techniques that enable us to measure and verify cognitive task complexity differences (i.e., the levels of the independent variable) independently of linguistic performance measures (i.e., the dependent variable) (Norris, 2010; Norris & Ortega, 2003; Robinson, 2001a; Révész, 2014). Additionally, the use of independent measures of cognitive task complexity should enable us to begin to see the actual degrees of task complexity (i.e., how simple and complex each task is), instead of a relative labeling of simpler or more complex tasks. The use of such measures will also help us take into account participant factors (i.e., who the learners are) and task condition factors (i.e., what tasks are used, how are they implemented), and to investigate the effects of increased cognitive complexity of particular tasks for particular learner populations.

Given the importance of measuring and otherwise investigating cognitive task complexity independently of task performance, a few researchers have begun to address this issue in the field of TBLT. Robinson (2001b) was likely the first SLA researcher to attempt to measure independently the cognitive complexity of two versions of an L2 direction-giving task using a
self-assessment of task difficulty, and since then this measure has been used in a handful of other task complexity studies. However, a recent research synthesis conducted by Sasayama et al. (2013, 2014, 2015) revealed that among 129 studies of L2 task complexity, only 18% of the studies employed independent measures of cognitive task complexity. Among these studies, the majority (70%) used only subjective, self-assessment ratings of perceived task difficulty or mental effort, or holistic estimations of greater/lesser complexity (22%).

Recently, there has been a relatively concerted effort to advance our understandings of the measures and explore possibilities of adapting such measures specifically to TBLT research (e.g., a colloquium on cognitive task complexity measures in TBLT at the American Association for Applied Linguistics 2013 conference). Stemming from this initiative, the measures that have been adapted and utilized in TBLT include: (a) time estimation (e.g., Baralt, 2010, 2013), (b) eye-tracking (e.g., Révész, Sachs, & Hama, 2014), (c) dual-task methodology (e.g., Révéz et al., 2014), and (d) expert judgment (e.g., Brown et al., 2002; Révész et al., 2014). It is important to note, however, that the majority of the studies in TBLT that did use one or more independent measures of cognitive task complexity so far have used these measures to confirm that their task complexity manipulation was successful, rather than to explore the relationship between cognitive task complexity and task design features, task implementation conditions, and learner characteristics and responses to tasks. For this reason, it is as yet largely unknown why a measured-to-be complex task is actually cognitively more complex than a measured-to-be simple task or to what extent and how factors other than task design features may influence the level of cognitive task complexity. To find answers this question, researchers such as Tavakoli (2009) and Kim, Payant, and Pearson (2015) have suggested and demonstrated the use of introspective
methods, such as interviews and stimulated recalls with L2 learners. Focusing on learners’ perspectives is thus another important way to gauge cognitive task complexity and understand what the task really ‘does’ to the learners. In order to set the stage for an investigation into the realities of cognitive task complexity within a given set of L2 communication tasks, this chapter reviews prominent approaches to independently measuring and otherwise investigating cognitive task complexity as utilized within various disciplines and particularly in TBLT research.

3.2 Measures of cognitive load used in cognitive psychology and multimedia learning

Techniques for measuring the cognitive load of a task have been actively explored in several domains, and particularly in the fields of cognitive psychology and multimedia learning. In these fields, interest in measuring cognitive load has stemmed from researchers’ desire to develop instructional materials that induce quantitatively more and better learning, taking into account how different types and degrees of cognitive load relate to the effectiveness of students’ learning of the content knowledge. According to Brünken, Plass, and Leutner (2003), currently available techniques for measuring cognitive load or cognitive complexity of a task can be categorized into four groups, according to objectivity (subjective or objective) and causal relation (direct or indirect). The first group includes subjective, indirect measures, such as a self-report of invested mental effort. These measures are considered indirect because mental effort refers to the cognitive process elicited by executing the task at hand, and the perception of it is likely to involve more than just the task itself (van Gog & Paas, 2008). For example, in the field of cognitive psychology, participants are typically asked to rate their perceived level of mental effort invested in completing a task by using nine-point Likert scales, ranging from 1 (very, very low mental effort) to 9 (very, very high mental effort) after each task (e.g., Paas et al., 1994).
The second group includes subjective but direct measures, such as a self-report of stress level and difficulty of materials. In a study conducted by Kalyuga, Chandler, and Sweller (1999), for instance, participants were asked to “estimate how easy or difficult the instructions [computer-based instructional materials] were to understand by clicking one of the seven choices … from extremely easy (corresponding to the score 1) to extremely difficult (corresponding to the score 7)” (p. 358). As mentioned earlier, a self-assessment of perceived difficulty along these lines has been the most popular measure of task complexity in the field of TBLT. The key difference between these first two categories of measures, according to Brünken et al. (2003), is whether a direct causal link can be established between the results of the measure and the actual cognitive load imposed. Perceived difficulty of materials, for example, is argued to result directly from the actual cognitive load of the materials, whereas low mental effort may be a result of low cognitive load, or it may reflect the learners’ intentional strategy for dealing with excessively high cognitive load (Brünken et al., 2003; van Gog & Paas, 2008). Although task difficulty ratings are claimed to be a direct measure of cognitive load, Brünken et al. (2003) also warn that the differences in difficulty scores could potentially be caused not only by the cognitive load of a given task, but also “individual competency levels of the learners” and/or “different attentional processes” (p. 56). Hence, caution has to be exercised in using and interpreting the data of the task difficulty measure as a truly direct measure of cognitive load. The use of these subjective measures of mental effort and difficulty is, of course, underpinned by the assumption that participants are capable of judging the level of cognitive complexity – or distinguishing the amount of difficulty – imposed on them by different tasks (Révész, 2014).
The third group includes objective, indirect measures, such as learning outcome measures and behavioral and physiological measures. The most popular learning outcome measures used in the domain of multimedia learning are knowledge acquisition scores. Typically, multimedia learning studies compare two or more versions of instructional design of the same materials (e.g., giving an *auditory* explanation of a topic while learners are getting relevant visual input, such as pictures, diagrams, and graphs, versus giving a *written* explanation while they are getting the same visual input). Given that the materials to be learned are exactly the same in the two versions, and hence the task is thought to pose the same degree of intrinsic cognitive load if learners’ knowledge level is kept constant, knowledge acquisition scores are believed to reflect the degree of extraneous cognitive load. High knowledge acquisition scores, for example, supposedly indicate low extraneous cognitive load. This is based on the assumption that learners acquired more knowledge about the topic because they were able to devote more attentional resources to the learning materials due to the low extraneous cognitive load posed by that particular instructional design. This way of measuring cognitive load, however, has been criticized by a few researchers because the degree of knowledge acquisition has been found to be affected not only by cognitive load, but also by measurement methods of knowledge acquisition and learner individual differences (Brünken et al., 2003). The very similar problem of using task performance measures as a direct indication of cognitive task complexity in TBLT has already been pointed out in Chapter 2.

An example of behavioral measurement is learners’ time-on-task. The assumption here is that learners would spend different amounts of time depending on how cognitively complex or demanding a given task is (Brünken et al., 2003). Some examples of physiological techniques
include measures of heart activity (e.g., heart rate) (Paas & van Merriënboer, 1994), eye activity (e.g., pupil dilation, blink rate, eye movement and fixation) (Paas, Tuovinen, Tabbers, & Van Gerven, 2003; Palinko, Kun, Shyrokov, & Heeman, 2010), and skin response (Reimer, Mehler, Coughlin, Godfrey, & Tan, 2009; Shi, Ruiz, Taib, Choi, & Chen, 2007). These measures are considered indirect because “they depend on processes of information storage and retrieval that may [or may not] be affected by cognitive load” (Brünken et al., 2003, p. 56). The primary challenge with using indirect measures, as implied by the name, is that they do not provide a direct indication of cognitive states, and as such may be influenced by moderating variables that intercede between the participant’s cognitive response to a task and performance or behavior on the measures. Yet, behavioral and physiological measures have great potential in that they are free from subjective perspectives and could be useful especially for triangulation purposes. These measures have just started to be utilized in some TBLT studies (see below).

The last group includes objective, direct techniques, such as brain activity measures and the dual-task paradigm. The use of neuro-imaging techniques, such as fMRI for example, enables us to see how actively one’s brain is functioning while engaging in different tasks (Brünken et al., 2003). However, in language research, this technique has as yet only been used to investigate human brain activity when engaging in relatively simple, comprehension tasks, such as an auditory comprehension task (e.g., Lidzba, Schwilling, Grodd, Krägeloh-Mann, & Wilke, 2011; Tagarelli, 2014), a lexical decision task (e.g., Grant, Fang & Li, 2015), and a grammaticality judgment task (e.g., Tagarelli, 2014); therefore, the connection between cognitive load and brain activity, especially for elaborated ‘real-world’ production tasks, such as L2 speaking tasks, is not entirely clear (see also Brünken et al., 2003). Additionally, access to fMRI apparatus can be
extremely difficult and costly. Hence, Brünken et al. (2003) advocate the use of the dual-task methodology to directly and objectively measure the level of cognitive load imposed by a task. For example, Brünken, Steinbacher, Plass, and Leutner (2002) used the dual-task approach to measure extraneous cognitive load posed by multimedia instructional materials. In this study, participants were asked to learn about the human cardiovascular system in their first language (German) by using a multimedia computer-based training program (i.e., the primary task). The training program included 22 screen pages. Among these 22 pages, some were presented with written texts and associated pictures (visual only condition), and others with auditory input and pictures (audiovisual condition), to investigate the modality effect. Participants were simply instructed to read or listen to the passage while looking at the associated pictures. As a secondary task, they were asked to respond as quickly as possible to a color change in the letter ‘A’ presented on the computer screen together with the learning materials. The premise behind this technique is that a primary task with high cognitive demands consumes a lot of attentional resources, and thus participants are left with little attentional capacity to be devoted to the secondary, color-detection task, which in turn results in a slow response rate to the color change (Brünken et al., 2002). Therefore, the longer the reaction time to the secondary task is, the more cognitively complex a primary task is interpreted to be.

In the dual-task paradigm, the importance of an appropriate choice of a secondary task has been emphasized by Schoor, Bannert, and Brünken (2012). Schoor et al. compared two versions of a secondary color-detection task. The primary task was the same as that in Brünken et al. (2002). One version of a color-detection task asked participants to respond to the change of a background color on the computer screen while engaging in the primary task; the other version
asked participants to respond to the color change of a letter displayed in the upper part of the screen. In both cases, the mode of primary task/material presentation was alternated between a visual-only and audiovisual presentation format. They found a statistically significant difference between the two types of presentation format in terms of participants’ reaction time to the color change, but *only* for the visually non-congruent, letter color-detection task. These results suggest that the background color-detection task may have posed too few attentional demands (i.e., ostensibly because it is too overtly obvious to detract very much attention), and as a result participants were able to pay simultaneous attention to the primary learning task and the secondary color-detection task, regardless of the level of cognitive load of the primary task (Schoor et al., 2012). The letter color-detection task, on the other hand, seemed to have required additional attentional resources to focus on a certain area of the input (i.e., the upper part of the screen), and consequently, it successfully detected the difference in the level of cognitive load between the presentation modes (Schoor et al., 2012). These findings underscore the importance of choosing a secondary task that poses optimal processing demands when using the dual-task methodology.

Lastly, although not included in the list of cognitive load measures provided by Brünken et al. (2003), time estimation has also been used as a measure of cognitive load in the field of cognitive psychology. A meta-analysis conducted by Block, Hancock, and Zakay (2010) revealed that when participants are asked prospectively (prior to their task performance) to judge how much time they think they have spent to complete a task at hand, the estimated time tends to decrease as cognitive complexity increases. The attentional-gate model (AGM) proposed by Zakay and Block (1997) is claimed to account for this overall trend. AGM proposes that
attending to time enables what they call the ‘attentional gate’ to open and allows the pulse stream to come in to the cognitive center. The total amount of pulses that reach the cognitive center determines the length of time experienced to have passed. The model, thus, predicts that the more cognitively demanding a task is, the less attentional resources become available to attend to time, which allows less opportunities for the attentional gate to open and in turn decreases the length of perceived time. The prospective time estimation approach, therefore, is considered to be a type of a dual-task methodology, with attending to and estimating time playing the role of a secondary task (Block et al., 2010). On the other hand, in the same meta-analysis, time estimation done retrospectively only (i.e., without advance notification) was shown to display a distinct relationship between the degree of cognitive load and the length of estimated time from the prospective time estimation measure; that is, the more cognitively demanding a task at hand is, the longer, rather than shorter, the time estimation becomes (Block et al., 2010). This pattern cannot be explained by AGM, but rather can be accounted for by another model, the contextual-change hypothesis (Block, 1978). This hypothesis proposes that retrospective judgments of time are about retrieving information stored in memory about contextual changes (e.g., changes in kinds of mental processing) that happened during task performance; hence, the remembered time duration is correlated with the amount of contextual changes experienced and stored in memory (Zakay & Block, 1997). The more complex a task is, therefore, the more varied kinds of mental processing are hypothesized to be experienced, which as a result lengthens the remembered time duration.
3.3 Measures of task complexity used in the field of TBLT

Several researchers have sought to measure cognitive task complexity independently of observations about linguistic performance in the field of TBLT. Robinson (2001b), for example, used a subjective self-rating questionnaire to access participants’ perceived task difficulty. After completing each task, participants in these studies were asked to rate their perceived levels of task difficulty, stress, confidence, interest, and motivation by using nine-point Likert scales. Robinson found that participants perceived a designed-to-be simple, direction-giving task as statistically significantly less difficult and less stressful than a designed-to-be complex direction-giving task. This questionnaire developed by Robinson has been used in a handful of other studies in the field. Gilabert (2007b), for example, used it as an independent measure of cognitive load and indeed found a difference in participants’ perceived difficulty between the simple and complex versions of tasks, but only for certain task types (i.e., instruction-giving, decision-making) and not others (i.e., narrative). Using exactly the same decision-making task from Gilabert’s study, however, Shiau and Adams (2011), found only a very small difference between the two versions of the task in terms of participants’ perceived task difficulty levels (pointing to the possibility that the same task can pose differing degrees of complexity depending on who the learners are, at least when measured by the self-assessment of task difficulty).

Although this measure of task difficulty gives us new insights into how participants perceive the tasks at hand, it falls within the subjective category of measures introduced above, and how learners feel about the tasks may not always accurately reflect the actual cognitive task complexity experienced (as argued by Brünken et al. 2003). Considering the three-way distinctions of task complexity, task difficulty, and task condition (i.e., the Triadic Componential
Framework) proposed by Robinson (e.g., 2001a), this self-rating measure of task difficulty, as the name suggests, may be measuring learners’ perceived task difficulty rather than task complexity that is claimed to be inherent to the task design and distinct from participants’ perceptions (see Robinson, 2001b).

Introspective methods have also been used by a few L2 researchers. For example, Kim et al. (2015) used a stimulated recall and interview (as well as a self-assessment of perceived difficulty) to validate the assumptions of cognitive task complexity. By design, cognitive task complexity was manipulated along the [+/- no reasoning demands] dimension. In stimulated recall sessions, participants watched a video of themselves engaging in the tasks with their partner and commented on their mental states when asked to describe what they were thinking at certain times of their task engagement. Through this technique, Kim et al. found that the designed-to-be more complex, decision-making task (with higher reasoning demands) required comparisons between and evaluations of three phone companies more often than the designed-to-be simpler, information-exchange task (with less reasoning demands), despite the fact that both of these tasks were rated very similarly by the same learners in terms of perceived task difficulty. In other words, in cognitive terms, they actually had to engage in more activities, even though they did not perceive the complex task to be more difficult. Additionally, her interview data revealed that a variety of phenomena, including topic of tasks, language use, and cognitive processes, were mentioned frequently to justify participants’ perceived task difficulty.

Tavakoli (2009) also interviewed teachers and learners, asking “how difficult they found each task, which task they perceived as the most difficult, easiest, or most interesting, why they found some tasks more difficult than others and what factors they thought contributed to this
difficulty” (p. 7). She found that the designed-to-be simpler, tightly structured narrative task was perceived to be easier than the designed-to-be more complex, loosely structured narrative task, and that the cognitive demands of a task (e.g., “understanding the story, following the story up, or fully realizing what was happening in the story” (p. 9)) were the most frequently mentioned reason why participants felt a task to be difficult to perform. The interview also identified linguistic demands, clarity of pictures or storyline, amount of information, task structure (e.g., “concepts such as the way information was organised, whether the story was coherent and whether the sequence of events was natural” (p. 12)), information grounding (e.g., “presence or absence of background information in a picture story” (p. 13)), and affective factors as reasons why the students found – or the teachers thought the students would find – some tasks hard or easy to perform. The findings yielded by Kim et al. (2015) and Tavakoli (2009) point to the importance of accounting for task design features and other task-related factors to fully understand the nature of cognitive task complexity.

Another measure used in the field of TBLT is expert judgments of task complexity (e.g., Brown et al., 2002; Révész, 2011; Révész et al., 2014). In Révész et al.’s (2014) study, for example, participants were asked to read a biography of two famous historical people and then reason about potential causes of the events depicted in the biography. Thus, after reading about Eleanor Roosevelt, they were presented with a sentence “Eleanor helped to pass laws about human rights” (p. 625) with a picture depicting the event. When they hit a computer key to go to the next slide, they saw from left to right the expression ‘if … not’ on the furthest left, two possible causes of the event of Eleanor helping to pass laws about human rights (i.e., “learn about refugees’ problems” and “have good relationship” in the simple version or “care for
soldiers” in the complex version, p.1 of supporting information) next to it, ‘then … not’ in the arrow pointing to the right, and the picture that they saw in the previous slide depicting the actual event that happened in history. Participants were then asked to choose a potential cause that best suited each event and formulate counter-factual sentences, using if … not, then … but (e.g., If Eleanor had not learned about the refugees’ problems, then she had not helped to pass laws about human rights). Task complexity in this study was thus intentionally manipulated by increasing reasoning demands. The potential cause of the event was apparent in the simple version of the task, whereas it was vague (hence requiring more reasoning demands) in the complex version. In this study, as a measure of expert judgment, two doctoral students in Applied Linguistics were asked to rate the complexity of 32 past counter-factual items (i.e., eight sentences for two tasks in the simple and complex versions) on a 5-point Likert scale. The results revealed that the two doctoral students judged items in the supposed-to-be complex version to be considerably more complex than the supposed-to-be simple version of the task. Although these subjective, indirect measures of cognitive load certainly give us new insights into how learners themselves or experts think learners perceive tasks at hand to be, they fall short of determining the reality of the relationship between the task design and learners’ mental responses to it. The following three measures are newly adapted from cognitive psychology for this precise reason, in other words, to tap into learners’ cognitive and mental processes rather than their perceptions.

Baralt (2010, 2013) was the first in the field of TBLT to use subjective time estimation to independently measure cognitive task complexity. Given that the time estimation measure that she used was retrospective, rather than prospective, Baralt hypothesized that increased task demands would lead to estimated time that is longer than the actual time on task. Indeed, her
study revealed that task participants’ estimated time was statistically significantly longer than their actual time on task for the designed-to-be complex task (that required intentional reasoning), and their estimated time was statistically significantly shorter than their actual time on task for the designed-to-be simple task.

Another measure that has been recently utilized in L2 research as an indicator of task complexity is eye-tracking. Révész et al. (2014) were the first to use eye-tracking in TBLT as an independent measure of cognitive load related to task complexity. They measured the number of eye fixations and the length of each fixation for each participant using an unobtrusive, remote eye-tracking device while participants were engaging in two versions of language tasks with differing designed degrees of reasoning demands. The premise behind this technique is that the higher the number of fixations and the longer the fixation, the more cognitively demanding the task of interest is assumed to be. Révész et al. (2014) found that the designed-to-be complex task required statistically significantly more instances of, and longer, eye fixations than did the designed-to-be simple task.

In the same study, Révész et al. (2014) also used the dual-task methodology. Participants in their study were asked to respond to a background color change on the computer screen while simultaneously engaging in a primary task of formulating counter-factual sentences based on input given on the screen. They found that participants’ responses to the color change were statistically significantly slower and less accurate for a designed-to-be complex task (despite small mean differences) compared to a simple task; hence, the designed-to-be complex task was argued to be cognitively more demanding.
3.4 What does data elicited by measures of cognitive load really tell us?

It is a positive sign to see this many measures of cognitive load, originating in cognitive psychology, being used in TBLT in an attempt to tackle the issue of validating the assumption of cognitive task complexity. Our endeavor has only begun, however, and there are several considerations to be made at this juncture. Firstly, these measures of cognitive task complexity have not been validated in the context of TBLT research; hence, the meaning of any single measure may not be entirely clear as yet. In cognitive psychology, the mental effort questionnaire, for example, has been used in a number of studies, and its high validity and reliability have already been established (see Paas et al., 2003). This, however, is not the case with any of the measures utilized in TBLT. Given this lack of validation, the best we can do at this moment is to utilize multiple measures of cognitive load and increase the possibility of correctly interpreting the data that these new measures produce. The more studies that do utilize such measures of cognitive load in the context of TBLT research, the better we can understand how exactly we should interpret them and which measures seem to be more suitable than others in this particular context. Of course, studies that directly validate the measures of cognitive load in relation with typical L2 communication tasks will be a valuable addition to the field as well.

Related to the first issue, it is as yet unclear how much difference on any of these measures is a meaningful difference between tasks, in terms of their complexity. It has been a tradition (not just in TBLT, but also in cognitive psychology and multimedia learning) to rely on statistical significance testing as a way of determining ‘meaningful’ difference. In Révész et al. (2014), for example, the sheer difference in reaction time to the background color change attributable to two different task conditions was only 0.006 seconds on average, yielding a very small effect size of
0.15 (calculated by me). Nevertheless, this difference was found to be statistically significant, which led the authors to argue that their task complexity manipulation was successful in causing a salient cognitive effect. However, whether that amount of difference is in fact salient (i.e., from the point of view of the effect on learners) remains to be determined. This example illustrates that statistical significance does not equal meaningfulness, and it is often times misleading to rely overtly on statistical significance as a proxy for magnitude or importance of effect (see Norris, 2015b). For this reason, it is advisable to use effect sizes rather than statistical significance for data interpretation, as a way to begin uncovering the actual magnitude of difference on any of these measures that might be attributable to task design differences, and to triangulate interpretations via the use of multiple measures, so that we can get at a fuller understanding of what the measures are really telling us.

Lastly, in triangulating data, it is important to be aware of the expected relationship among different measures. A clear case is the relationship between task difficulty and mental effort ratings. Thus, van Gog and Paas (2008) argued that self-ratings of task difficulty and mental effort tap into different constructs and therefore measure possibly related yet distinct phenomena. In CLT, task difficulty is supposedly directly related to the task itself, whereas mental effort is about how much cognitive load learners actually devote to in response to the given task. It has been shown empirically that when a task is extremely difficult, posing overwhelmingly high cognitive load, learners can lose motivation and as a result devote less attentional resources to task completion (van Gog & Paas, 2008). Hence, it is possible that task difficulty and mental effort ratings may elicit non-equivalent results. Another good example is the relationship between prospective and retrospective time estimation. As reviewed above, depending on which
approach is taken, the interpretation of elicited results changes completely. In other words, in the paradigm of prospective time estimation, the more cognitively complex a task at hand is, the shorter the estimated time is supposed to be, whereas retrospective time estimation predicts that a cognitively complex task elicits longer duration of estimated time than does a simple task. In fully interpreting the results elicited by these measures (especially the ones new to the field of TBLT), then, it is highly important to be aware of how each measure should in theory react to different degrees of cognitive load for more accurate interpretation of the elicited data.

### 3.5 Chapter summary

Chapter 3 has provided an overview of various measures of cognitive load, utilized in cognitive psychology, multimedia learning, and SLA (TBLT in particular). Subjective measures, such as mental effort and task difficulty self-ratings, introspective interviews, and expert judgment, give us an opportunity to investigate the nature of task complexity from the perspectives of the participants. Taking into consideration how learners themselves perceive tasks is extremely important because it is the learners who engage in the tasks and it is their L2 abilities that we are trying to foster. Furthermore, having both direct (e.g., task difficulty ratings) and indirect subjective measures (e.g., mental effort ratings) is useful in that it helps us understand in depth whether the task at hand was easy or difficult to handle and whether it had the consequence of encouraging learners to devote more or less attentional resources in attempting to accomplish the task. Objective measures, such as the dual-task methodology and eye-tracking, are free from participants’ interpretations of the tasks at hand and may give us a purer, unbiased indication of cognitive load. When used together, these measures can complement each other. Also, given the recent introduction of cognitive load measures to the
field of TBLT, it is as yet not entirely clear how the data elicited by a single measure of cognitive load should be interpreted. It is crucial, therefore, to use multiple measures of cognitive load, not just one, so that findings can be triangulated into an overall more accurate understanding of the cognitive complexity of a given task.
CHAPTER IV: THE ROLE OF PROFICIENCY IN UNDERSTANDING TASK COMPLEXITY

4.1 Introduction

Understanding the nature and the effects of cognitive task complexity is complex enough in dealing with native speakers, as researchers have done in the fields of cognitive psychology and multimedia learning. Our task becomes even more complex when we are interested in such phenomena for L2 learners. This is because communicative task performance is always mediated by the language involved. In L2 task complexity research, learners’ performances are influenced not only by design features of tasks in which they engage but also by their proficiency levels. L2 learners with higher proficiency levels are typically able to do fundamentally more than learners with lower proficiency levels, and their proficiency may also interact with the putative effects of cognitive task complexity (as we will see in the subsequent section). For these reasons, it is fundamentally essential to take into account this additional factor of proficiency in understanding how and why learners perform L2 communication tasks in the ways that they do. Potentially as well, the independent measures of cognitive load mainly used in cognitive psychology with native speakers may also interact with L2 learners’ proficiency. If cognitive task complexity is inevitably about the interaction of task features with learners’ minds, then the status of the L2 in their minds should play a role in that interaction as well.

Given that in TBLT research, we are mostly interested in L2 learners’ abilities to use a language for communication purposes, rather than simply their explicit knowledge of language forms, L2 proficiency in this paper is defined as L2 learners’ general ability to communicate in the L2. In what follows, I will discuss further why L2 proficiency should be measured, reported,
and treated as a moderating variable in instructed SLA research, and in task complexity research in particular.

4.2 Proficiency as a moderating variable in SLA and TBLT research

Tremblay and Garrison (2010) have rightly argued the importance of documenting and controlling for variables that are likely to affect L2 learners’ performance, interaction, and learning in SLA research, including in particular target-language proficiency. Measuring global L2 proficiency is essential for controlled experimental research, especially to ensure equivalent assignment of participants to groups or conditions and to understand the population to which research findings can be generalized (Norris & Ortega, 2012). Despite its apparent importance, however, the role of L2 proficiency in SLA has been largely ignored by primary researchers, perhaps because of a lack of standardized and efficient estimates of proficiency or researchers’ beliefs that educational distinctions like “curricular level” in a language program are sufficient to determine homogeneity and generalizability of learners in terms of their proficiency (Norris & Ortega, 2012; Robinson, 2005). Thus, it is extremely common to find proficiency treated as an afterthought by researchers in describing the population of a study, with statements like ‘the learners were of intermediate proficiency’ presumed to suffice.

By way of illustration, consider a few studies that have looked into the issue of L2 proficiency in SLA, Thomas (1994, 2006) synthesized research on how target-language proficiency has been measured in the field of SLA during two periods of time (i.e., 1988-1992 & 2000-2004). She identified four types of proficiency measures: (a) impressionistic judgment, (b) institutional status, (c) in-house assessment and research-internal measures, and (d) standardized tests. Impressionistic judgment is a way of estimating participants’ proficiency levels based on
“the experimenter’s unsupported evaluation, or the evaluation of some other (often unspecified) person” (Thomas, 1994, p. 314). The infamous ‘intermediate level learners’ phenomena would be categorized here. Institutional status refers to a type of proficiency measure that relies on program levels, for example, ‘first-semester’ Japanese learners. The idea here is that the higher the program level is the more proficient the participants are supposed to be. In-house assessment and research-internal measures refer to a proficiency test that has been developed and administered locally in educational settings or for research purposes. A placement test is a good example of an in-house assessment measure, and a cloze test is possibly the most common research-internal measure of L2 proficiency (Thomas, 1994). Lastly, standardized tests, such as TOEFL iBT, are typically validated in multiple educational contexts and are available for use to anyone. Thomas found that in both corpora of studies, researchers tended to downplay the role of proficiency in their studies, and either not measure proficiency at all or rely mostly on the first two categories of measures.

Similarly, Tremblay (2011; Tremblay & Garrison, 2010) surveyed three journals (i.e., French Language Studies, Second Language Research, Studies in Second Language Acquisition) to review the extent to which proficiency was measured and how in studies of L2 French. They found that among a total of 144 studies on L2 acquisition, only “a little more than one third of all the studies … assessed L2 learners’ proficiency independently” (Tremblay, 2011, p. 342) by means of in-house assessment instruments, research-internal measures, or standardized tests.

Although similarly uncommon within the specific field of TBLT, several researchers have been interested in the role of proficiency and have investigated how learner proficiency affects or moderates (a) the quantity and quality of learner interactions (e.g., Kim, 2009; Porter, 1983), (b)
the effects of instructional and testing material design on learner task performance (e.g., Ishikawa, 2006; Wigglesworth, 1997; Yoshida, 2012), and (c) learner perceptions of instructional materials (e.g., Chen & Chang, 2011; Kim, 2009).

More specifically within the domain of L2 task complexity, Sasayama et al.’s (2013, 2014, 2015) research synthesis revealed that among the 129 empirical studies published on the topic of cognitive task complexity between 1987 and 2013, 18% of the studies did not report participants’ proficiency levels in any way, 17% of the studies used impressionistic judgments, and 29% of the studies used institutional status to report learners’ global proficiency. Another 12% of the studies used some kind of an in-house assessment, such as a placement test (e.g., Robinson, 1995), cloze test (e.g., Kuiken & Vedder, 2007), or a C-test (e.g., Nemeth & Kormos, 2001). The rest of the studies (24%) used standardized proficiency tests, such as TOEFL (e.g., Barkaoui, Brooks, Swain, Lapkin, 2013), Michigan Test of English Language Proficiency (e.g., Ishikawa, 2006), and the General English Proficiency Test (e.g., Chang, Lei, & Tseng, 2011). The synthesis also revealed that among 129 empirical studies included in the synthesis, only 19% (22 studies in number) treated proficiency as a moderating variable, while the remaining 81% either did not measure or report any information about proficiency or estimated proficiency in some way but did not consider its potential effects on other aspects of the investigation. However, the majority (75% to be precise) of the 22 studies that did look into proficiency as a moderating variable found interaction effects between proficiency and task performance; in other words, depending on their proficiency levels, learners benefited from increased or decreased cognitive complexity of tasks differentially in terms of their task performance (Sasayama et al., 2013, 2014, 2015).
For one concrete example, Ishikawa (2006) investigated the relationship between learner proficiency and the effects of complexity manipulation on two writing tasks. His participants (Japanese high school students) were divided into high and low proficiency groups based on their scores on the Michigan Test of English Language Proficiency. His tasks were picture-based written narratives designed to pose distinct degrees of cognitive complexity. One task was a ‘Here-and-Now’ task where participants were asked to write a story while looking at the picture prompt (simple task). The other was a ‘There-and-Then’ task where participants were asked to look at the picture prompt for five minutes and then write a story without an access to the prompt (complex task). Additionally, participants in the simple and complex task conditions were given the first sentence of a story written in the present tense and in the past tense, respectively. Sasayama et al. (2013, 2014, 2015) calculated Cohen’s $d$ effect sizes between the simple and complex tasks in his study for accuracy, syntactic complexity, lexical variety, and fluency measures, for each proficiency group. As seen in Figure 2, the effect size comparisons revealed that increased cognitive task complexity had (a) a positive effect on accuracy both for low and high proficiency groups, but to varying degrees (i.e., greater effect on low proficiency learners); (b) identical magnitude of positive effects on syntactic complexity and fluency for both groups; but (c) a very high positive effect on lexical variety for low proficiency participants, but a negative effect for high proficiency participants. In other words, high proficiency learners produced less lexically varied narratives under the complex condition compared with the simple condition, while low proficiency learners produced much more lexically varied narratives under the complex condition compared with the simple condition.
Figure 2. Effect sizes between the simple and complex tasks for low and high proficiency groups in Ishikawa’s (2006) study. Adapted from “Task complexity in TBLT research and beyond: A research synthesis,” by S. Sasayama, A. Malicka, and J. Norris, 2013, Paper presented at the EUROSLA 23 conference. Copyright 2013 by Sasayama et al. Adapted with permission.

To provide another example, Yoshida (2012) investigated the relationship between the effects of readability (i.e., a source of cognitive task complexity) on learners’ reading comprehension and their L2 proficiency. Yoshida also used the Michigan Placement Test, although only the reading section, to divide her participants (university students in Japan) into high and low proficiency groups. In the main experiment, the participants read two texts, a simpler, expository text with higher readability scores and a more complex, narrative text with lower readability scores. Again, Sasayama et al. (2013) calculated Cohen’s $d$ (effect size indicator) between the simple and complex texts on total number of ideas recalled and total number of main ideas recalled. The results revealed that the more complex reading text had equally negative effects on learners’ total recall regardless of their proficiency levels; by contrast, it had differentially positive effects on their recall of main ideas, with high proficiency
learners benefiting from the increased cognitive load of the reading task more substantially than low proficiency learners (see Figure 3).

Figure 3. Effect sizes between the simple and complex tasks for low and high proficiency groups in Yoshida’s (2012) study. Adapted from “Task complexity in TBLT research and beyond: A research synthesis,” by S. Sasayama, A. Malicka, and J. Norris, 2013, Paper presented at the EUROSLA 23 conference. Copyright 2013 by Sasayama et al. Adapted with permission.

Another study that demonstrates the important role of L2 proficiency in understanding the effects of task complexity, in this case with attention to the phenomenon of learner-learner interaction, is Kim (2009). She investigated the effects of cognitive task complexity, manipulated by increasing reasoning demands in one task and number of elements in another, on language related episodes (LREs) for different proficiency groups. In her study, proficiency was operationalized by participants’ TOEFL scores, and participants who scored between 340-420 (on the paper-based version) were assigned to a low proficiency group and those who scored between 440-490 were assigned to a high proficiency group. In the experiment, participants
engaged in four tasks in pairs: simple and complex versions of a picture-based narrative task (manipulated by [±/− no reasoning demands]) and of a picture difference task (manipulated by [±/− few elements]). Cohen’s d (effect size indicator) calculated by Sasayama et al. (2014, 2015) between the simple and complex tasks in terms of resulting LREs revealed interesting trends. When manipulated by increasing reasoning demands, the complex version of the narrative task had a strong positive effect on high proficiency learners, whereas the exact same task had a strong negative effect on low proficiency learners in terms of lexical LREs. In contrast, when manipulated by increasing the number of elements, the complex version of the picture difference task had a strong positive effect on low proficiency learners and a small positive effect on high proficiency learners in terms of lexical LREs (see figure 4). This study is particularly interesting because it demonstrates that complex versions of a task that are operationalized in distinct ways interact differentially with learners’ L2 proficiency.
These three example studies illustrate the importance of treating target-language proficiency as a key moderating variable in terms of a likely differential effect on L2 task performance, and thereby they point to a definitive requirement that proficiency be investigated persistently in all task complexity research. To be clear, however, these and other studies to date have all investigated the possibility of an interaction effect between learners’ L2 proficiency and L2 performance on tasks that are presumed to differ in terms of cognitive complexity; what has yet to be considered within the task complexity research domain is the possibility that
proficiency interacts with the actual cognitive complexity of the task per se, in addition to the resulting performance.

When it comes to the relationship between L2 proficiency and the actual degree of cognitive complexity that a task poses to each learner, there is a true lack of investigation. It has been argued that cognitive task complexity is inherent to the task itself and therefore should be considered independent of learner variables, such as L2 proficiency (e.g., Robinson, 2001a). As illustrated in Chapter 1, in the domain of cognitive psychology, however, learner variables (e.g., learners’ knowledge level of the task content) are often presumed to have a critical effect on the degree of intrinsic cognitive load (e.g., Kalyuga, 2007). As such, it is hard to imagine that a task poses exactly the same amount of cognitive complexity regardless of learners’ L2 proficiency and other learner factors. For instance, a picture-based narrative task with as many as nine characters (i.e., [− few elements]) is argued to be cognitively more demanding than a similar task with only one character to be referred to (i.e., [+ few elements]) for any L2 learner, because the complex task requires them to distinguish among multiple similar characters and thereby pay more attention to the use of the L2 (in this case, noun modifiers in particular). In reality, however, it is possible that the complex task poses higher actual cognitive demands on lower proficiency learners due to their lack of vocabulary and/or grammatical knowledge than on higher proficiency learners. That is, the need to engage linguistically with the content of the task may itself serve as a source of increased cognitive load, that increase being relative to the proficiency of the learner. On the other hand, it is also possible that the complex task and the simple task pose similar amounts of cognitive demands on higher and lower proficiency learners, or even lower cognitive demands on the lower proficiency learners; that is, if the lower
proficiency learners are so overwhelmed by the complex task that they are unable to even think of the need to distinguish among similar characters to successfully complete the task at hand, then it may in reality induce relatively less cognitive demand as the learners simply do whatever they can to get through it with the linguistic resources they have. Given these uncertainties, it is important to investigate the possible moderating role of L2 proficiency, not only in interaction with task performance but also in potentially mediating the degree of cognitive task complexity that is realized by a given learner on a given task. The remaining issue to consider, then, is how learner proficiency can and should be measured.

4.3 Measurement of L2 proficiency

The review studies by Thomas (1994, 2006), Tremblay (2011; Tremblay & Garrison, 2010), and Sasayama et al. (2013, 2014, 2015) clearly illustrate a lack of attention to L2 proficiency as moderating variable in the field of SLA, including the domain of task complexity research, as well as researchers’ heavy reliance on easy and perhaps inaccurate estimates of proficiency (i.e., impressionistic judgment, institutional status). The use of such estimates is problematic because they tell us very little about actual L2 proficiency of the learners and they severely limit our ability to generalize study findings (although it is still better to report some kind of a proficiency estimate than ignoring it entirely). In fact, in an effort to develop a practical and yet accurate estimate of proficiency in French, Tremblay found that beyond the first year of instruction at a university French program, learners categorized in the same instructional level were distinctly heterogeneous, rather than homogeneous, in terms of their proficiency levels, when measured by a French cloze test. Hence, it has been empirically shown that simply using institutional status by itself is not enough to precisely capture learners’ proficiency levels.
Standardized testing would be an ideal option, especially because it allows us to compare results across multiple studies in constant terms, and as a result helps us decipher new knowledge by the use of research synthesis techniques. However, standardized tests are typically quite costly and time-consuming to be administered in experimental settings. It is certainly possible to collect participants’ self-reported scores of standardized proficiency tests, and it would be useful to have this piece of information. However, it is only rarely that all participants have taken the same standardized proficiency test uniformly outside of the experimental context, and hence their standardized test scores are not always available to be used in the experiment. For these reasons, researchers have worked on development and validation of short-cut estimates of language proficiency in various languages to be used in the investigation of language instructional studies (just as Tremblay did).

4.4 Developing short-cut estimates of L2 proficiency

In order to better understand various phenomena in SLA, it is fundamental to take into account the role of proficiency as a moderating variable. For this to occur, a valid and accurate estimate of proficiency that is suitable in the context of L2 research is crucial. Researchers’ attempts to develop such short-cut estimates of proficiency have a long tradition, though perhaps only with selected languages and without any sustained program of validation. Short-cut proficiency estimates, such as a cloze test or an elicited imitation test, have been used as a global measure of proficiency since the 1970s, with much greater attention being paid recently to their development for diverse L2s (Mozgalina, 2015; Tremblay, 2011; Tremblay & Garrison, 2010; Yan, Maeda, Lv, & Ginther, 2015). These tests are called reduced redundancy tests and are developed “to test a subject’s ability to function with a second language when noise is added or
when portions of a test are masked” (Spolsky, 1969, p. 10). This type of proficiency assessment has a number of advantages (Norris, 2015a). Firstly, reduced-redundancy tests can be developed relatively easily and administered in a short period of time – the latter is particularly helpful in a research setting where we already have a number of tasks and tests to administer. Secondly, they are capable of capturing learners’ integrated language skills. In order to successfully complete a cloze test, for example, learners need not only to know grammatical forms, but also to use their knowledge of the forms to understand a passage and to fill in the blanks (see below for more descriptions of how cloze-tests are typically developed). Lastly, many studies have shown that these tests show high internal consistency or reliability and high positive correlation with standardized test scores. The following section then reviews two of the reduced redundancy tests that have been used extensively in SLA research and which are relevant for the current study: (a) a cloze test and (b) an elicited imitation test.

4.4.1 Cloze test

One example of an efficient L2 proficiency measure is a cloze test. A cloze test is developed by deleting every $n^{th}$ word from a reading passage and providing a blank instead of each deleted word for test-takers to fill in. Brown (1980) reviewed nine validation studies of cloze tests and found that each cloze test was correlated highly with a criterion measure, such as TOEFL, the American University Beirut test, the UCLA ESL Placement Examination, and other placement tests, ranging from $r = .63$ to $.89$. The cloze test that he himself developed yielded a high correlation of $r = .88$ to $.91$ (depending on scoring methods) with the UCLA ESL Placement Examination. Its reliability was found to range from $K-R20 = .89$ to $.95$ (again depending on scoring methods). This means that most of the variance in learners’ scoring
patterns, and presumably their global proficiency, could be explained by the test itself. Similarly, Caulfield and Smith (1981) found that a Spanish cloze test that they developed was correlated highly with a standardized Spanish test, the Modern Language Association Cooperative Foreign Language Test (which includes all four skills) at $r = .90$. The cloze test was highly correlated not only with the written part of the test (i.e., reading and writing) but also with the listening and speaking parts of the test at $r = .71$ and $.77$, respectively. More recently, Tremblay (2011; Tremblay & Garrison, 2010) found that as participants’ program level (at a university setting) increased, they scored differentially higher on the French cloze test, and the corresponding K-R20 coefficient was found to be .94. These studies, then, seem to suggest that a cloze test may be a practical and reliable measure that can be efficiently used to estimate global L2 proficiency among learner populations.

4.4.2 Elicited imitation test

Another example of a short-cut estimate of L2 proficiency is an oral elicited imitation test. This test entails asking learners to listen to a series of stimulus sentences spoken in the target language and repeat them as precisely and as much as possible (Bley-Vroman & Chaudron, 1994; Erlam, 2006; Tracy-Ventura, McManus, Norris, & Ortega, 2013). The elicited imitation test is claimed to tap into a learner’s interlanguage system, because a successful completion of the test requires them to process a stimulus that they hear, rather than simply parroting it. If the target form included in the stimuli is not yet part of learners’ interlanguage, it will be difficult, if not impossible, for them to repeat the stimuli accurately (Bley-Vroman & Chaudron, 1994). A meta-analysis conducted by Yan, Maeda, Lv, and Ginther (2015) found that the average weighted effect size (Hedge’s $g$) of elicited imitation tests included in the analysis was 1.34 for
between-proficiency-level comparisons. This indicates that elicited imitation tests in these studies were able to distinguish among participants with different proficiency levels (as estimated by impressionistic judgment or institutional status, or as measured on criterion tests) highly consistently. Additionally, 11 studies that investigated the relationship between scores of elicited imitation tests (as a global measure of proficiency) and other proficiency measures revealed that the correlation ranged from $r = 0.46$ to $0.82$ (Yan et al., 2015). Similarly, in her review of validation studies of elicited imitation tests, Zhou (2012) found that reliability of the instrument ranged from .78 to .96 and that they correlated with other oral proficiency tests highly, ranging from .45 (guided oral interview) to .87 (SOPI). It is interesting to note that the elicited imitation test has also been found to be more sensitive in distinguishing between different proficiency levels when (a) it is used to measure global L2 proficiency rather than discrete grammar knowledge, (b) the sentences included in the test increase in length rather than remaining constant, and (c) the scoring is done as interval or ordinal scales rather than a dichotomous right or wrong (Yan et al., 2015). Indeed, an elicited imitation test in Russian developed by Mozgalina (2015) showed high psychometric values precisely for these reasons. Based on these findings, overall an elicited imitation test is argued to be another effective measure of global language proficiency.

4.5 Chapter summary

Chapter 4 illustrated the importance of measuring L2 proficiency and treating it as a moderating variable in task complexity research (and beyond). The majority of the studies on cognitive task complexity that investigated the role of L2 proficiency found interaction effects between proficiency and effects of cognitive task complexity. In other words, it is very likely
that the effects of cognitive task complexity are influenced and mediated by learners’ L2 proficiency levels. Similarly, although there has been a lack of research on the relationship between L2 proficiency and the amount of cognitive demands that a task poses, intuitively and also based on theories of cognitive psychology, it would make more sense to believe that L2 proficiency has a considerable role to play in determining the level of cognitive demands of a given task. Given the importance of L2 proficiency in cognitive task complexity research, the chapter concluded with two examples of a short-cut estimate of L2 proficiency, a cloze test and an elicited imitation test, that are suitable in the context of L2 research.
CHAPTER V: THE CURRENT STUDY

5.1 Rationale for the current study

In an attempt to deal with the issues surrounding L2 cognitive task complexity research reviewed above, this study utilizes multiple independent measures of cognitive load to attempt to validate assumed differences in four narrative tasks designed to be cognitively more or less complex, and it then investigates the relationship between the established complexity differences and Japanese-L1 English learners’ task performance in terms of CALF indices. It additionally explores the possible moderating effect of L2 proficiency on cognitive complexity and on L2 task performance. In what follows, I summarize unique contributions that the study makes to existing research on L2 cognitive task complexity:

1. Using multiple independent measures of cognitive load for triangulation

   As several measures, especially dual-task methodology and time estimation, are new to the field of SLA, interpretation of findings is not straightforward. When used individually, we may not know whether data elicited by a measure are reflective of the actual cognitive load of a task, or simply that the measure is not appropriate for the kinds of tasks we typically use in L2 research (e.g., oral communication tasks). Thus, it is extremely important to triangulate interpretations with multiple measures of cognitive load.

   Additionally, the present study not only investigates cognitive task complexity independently from task performance with multiple measures, but also attempts to deepen our understandings of the nature of the measures themselves. To this end, the current study calculated effect sizes for each measure in an attempt to gauge the degree to which each of the
measures is sensitive to cognitive complexity differences; more specifically, the study investigated whether any of the measures tends to exaggerate or minimize task differences.

2. Promoting in-depth understandings of the nature of tasks through the use of open-ended learner data

Although measurement of cognitive task complexity is becoming a new trend in the recent L2 task complexity research, the question of ‘what makes a task more or less cognitively complex’ has been explored only in a handful of studies (e.g., Kim et al., 2015, Tavakoli, 2009). Finding answers to this question is important in that it allows us to explore whether the manipulated task design element (e.g., number of elements) indeed makes a task more or less complex, as intended by a task-designer, and what other task-related factors we need to attend to, which in turn helps researchers, language teachers, and materials developers design more or less complex tasks in a more evidence-based manner.

3. Measuring cognitive complexity of four, rather than two, tasks

Sasayama et al. (2013, 2014, 2015) revealed that 85% of studies on cognitive task complexity adopted a dichotomous approach. In other words, researchers used two tasks, one being simpler and the other being more complex, to investigate the effects of cognitive complexity. Using four tasks in the current study allows for a range of complexity to be investigated, which will help us (a) better understand the relationship between task design and cognitive complexity and (b) begin to understand how much difference in task design may be needed to bring about complexity differences detected by measures of cognitive load.
4. Treating proficiency as a moderating variable

There has been a lack of attention to proficiency as a key moderating variable in instructed SLA in general, and in TBLT in particular. When researchers do pay attention to this issue, they have done so in investigating the role of proficiency as a moderator of task performance. Hence, no previous studies have yet examined the relationship between L2 proficiency and cognitive task complexity. The current study thus makes the first attempt in task complexity research to investigate how learners’ L2 proficiency might affect with the type and degree of cognitive task complexity, in addition to its effects on L2 task performance. Additionally, this investigation sheds new light on the interaction between L2 proficiency and various measures of cognitive load. It makes important and unique contributions, given that the measures of cognitive load newly introduced to TBLT have been used (and validated) primarily with L1 populations, and their potential interaction with L2 proficiency is still unknown.

5. Comparing results of L2 learners with a native-speaker baseline

In order to investigate the effects of cognitive task complexity on task performance without interference from L2 proficiency, it is important to examine how native speakers of English would react to the four tasks in terms of both the degree of cognitive complexity and their performances on these tasks. Although not typically done, the need for including a native-speaker baseline in task complexity research (and TBLT research in general) has been emphasized by a few researchers (e.g., Foster & Tavakoli, 2009).³

³ Importance of collecting a native-speaker baseline was repeatedly argued for in the 2015 Biennial International Conference on TBLT as well.
5.2 Research questions

The current study consists of two parts. The first part explores solutions to the issue of supposed versus actual cognitive task complexity, attempting to validate the assumed relationship between task design and the level of cognitive complexity of four picture-based narrative tasks for university-level Japanese-L1 learners of English as well as for native speakers of English. Accordingly, Part 1 of the study poses the following three research questions (RQs):

RQ1. What is the relationship between presumed task complexity, as implied by the design of four narrative tasks, and cognitive load, as measured by (a) dual-task methodology, (b) time estimation, (c) task difficulty self-assessment, and (d) mental effort self-assessment, for university-level Japanese-L1 learners of English as well as for native speakers of English?

RQ2. What task design or other factors do the English learners perceive and report in relation to presumed increases in cognitive task complexity, as they engage in the four narrative tasks?

RQ3. Is there a relationship between learners’ target language proficiency and the four measures of cognitive load as they are applied to four levels of presumed task complexity?

Building on the first part of the study, Part 2 investigates the relationship between measured, rather than assumed, cognitive task complexity and task performance of the same learner (i.e., university-level Japanese-L1 learners of English) and native speaker (i.e., English-L1 university students in the US) populations, in terms of syntactic complexity, accuracy, lexical variety, and fluency indices. Accordingly, it poses the following two additional research questions:
RQ4. What are the effects of increased cognitive task complexity on task performance in terms of syntactic complexity, accuracy, lexical variety, and fluency indices, for both university-level Japanese-L1 learners and native speakers of English?

RQ5. Is there a relationship between learners’ target language proficiency and their task performance?

5.3 Methods

In this section, the methodology of the current study, including participants, materials and instruments, procedures, and measurement and data analysis, will be described in detail.

5.3.1 Participants

Participants in this study were sampled to represent two distinct groups: English-L1 university students in the US and university-level Japanese-L1 learners of English. Care was taken to recruit participants from the same educational settings (and even the same courses when possible) in the first and second parts of the study to allow comparison to the extent possible. They were all recruited on a voluntary basis (i.e., not pressured to participate). Native speaker participants were recruited from two courses on linguistics at some US university. The researcher visited their classes to make an announcement about the experiment and interested students signed up on the spot. They received extra credits for their enrolled linguistics course, as compensation for their participation in this research. English-L2 participants were recruited with help of five professors at three Japanese universities. They circulated a flyer in their English classes and interested participants contacted me via email. I also visited some of the classes to make further announcements. This group of participants was paid 1,000 yen or 1,500 yen
(approximately $10 or $15), as compensation for their participation, depending on the length of the experiment.

5.3.1.1 Native speakers

30 university students whose L1 was English participated in the first part of the dissertation study (i.e., Study Part 1) on a voluntary basis, in Fall 2014. All, but two, were enrolled in the same course on linguistics (although they belonged to different sections). The remaining two participants were enrolled in another linguistics course. 23 were females, and the average age of all the participants was 18.77 ($SD = 1.43$), ranging from 18 to 24. All of them, except for two, were undergraduate students who were majoring in various fields (see Table 5). The remaining two participants were graduate students in TESOL (Teaching English to Speakers of Other Languages). All were living in the US at the time of the experiment.

Table 5

<table>
<thead>
<tr>
<th>Major</th>
<th>Number of participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguistics</td>
<td>7</td>
</tr>
<tr>
<td>Psychology</td>
<td>3</td>
</tr>
<tr>
<td>Business</td>
<td>2</td>
</tr>
<tr>
<td>English</td>
<td>2</td>
</tr>
<tr>
<td>French</td>
<td>2</td>
</tr>
<tr>
<td>German</td>
<td>2</td>
</tr>
<tr>
<td>Anthropology</td>
<td>1</td>
</tr>
<tr>
<td>Classics</td>
<td>1</td>
</tr>
<tr>
<td>Computer Science</td>
<td>1</td>
</tr>
<tr>
<td>International Politics</td>
<td>1</td>
</tr>
<tr>
<td>Undeclared</td>
<td>6</td>
</tr>
</tbody>
</table>

Similarly, participants in the second part of the study were sampled from the same populations of participants who took part in Study Part 1. L1 participants were recruited from courses on linguistics from which the participants in Part 1 were also recruited, to increase the
likelihood of homogeneity between the two participant groups. 31 university students whose L1 was English participated on a voluntary basis in Spring and Fall 2014. Among these, 20 were females, and the average age of all the participants was 19.06 ($SD = 1.26$), ranging from 17 to 21. All of them were undergraduate students who were majoring in various fields (see Table 6). All were living in the US at the time of the experiment.

<table>
<thead>
<tr>
<th>Major</th>
<th>Number of participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business</td>
<td>4</td>
</tr>
<tr>
<td>Linguistics</td>
<td>3</td>
</tr>
<tr>
<td>Psychology</td>
<td>3</td>
</tr>
<tr>
<td>Government</td>
<td>2</td>
</tr>
<tr>
<td>French</td>
<td>2</td>
</tr>
<tr>
<td>Arabic</td>
<td>1</td>
</tr>
<tr>
<td>Biology</td>
<td>1</td>
</tr>
<tr>
<td>German</td>
<td>1</td>
</tr>
<tr>
<td>History</td>
<td>1</td>
</tr>
<tr>
<td>Marketing</td>
<td>1</td>
</tr>
<tr>
<td>Mathematics</td>
<td>1</td>
</tr>
<tr>
<td>Undeclared</td>
<td>8</td>
</tr>
</tbody>
</table>

5.3.1.2 English-L2 speakers

Fifty-three Japanese-L1 university students (32 females) participated in Study Part 1 of the dissertation study on a voluntary basis. Their average age was 21.25 ($SD = 3.57$), ranging from 18 to 35. 27 were undergraduate students majoring in English; 17 were undergraduate, non-English majors; and nine were graduate students in Applied Linguistics. All were living in Japan at the time of the experiment. They were divided into low, mid, and high proficiency groups on the basis of a cloze test developed by Brown (1980) that was administered at the end of the experiment (see subsequent section on instruments). Table 7 summarizes participants’ self-
reported scores on standardized English tests where available. The tests of which the participants self-reported their scores varied widely, including TOEFL iBT, TOEFL ITP, TOEIC, and Eiken⁴. As these test scores are not comparable with each other, no analysis of them will be reported, except for the range of test scores for each test for high, mid, and low proficiency groups (divided based on the cloze test) found below.

Table 7
Participants’ Self-Reported Scores of Standardized English Tests in Study Part 1

<table>
<thead>
<tr>
<th></th>
<th>TOEFL iBT</th>
<th>TOEFL ITP</th>
<th>TOEIC</th>
<th>Eiken</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low (20)</td>
<td>NA</td>
<td>488 (1)</td>
<td>325-520 (11)</td>
<td>3-Pre1 (5)</td>
</tr>
<tr>
<td>Mid (14)</td>
<td>74-85 (4)</td>
<td>493-567 (5)</td>
<td>700-940 (4)</td>
<td>1 (1)</td>
</tr>
<tr>
<td>High (19)</td>
<td>73-114 (8)</td>
<td>567-695 (3)</td>
<td>765-990 (3)</td>
<td>Pre1-1 (4)</td>
</tr>
</tbody>
</table>

Note. The number in parenthesis indicates the number of participants who reported their scores for each test.

Sixty-six Japanese-L1 university students (38 females) participated in Study Part 2 on a voluntary basis. Their average age was 20.56 (SD = 1.99), ranging from 18 to 30. 43 were undergraduate students majoring in English; 17 were undergraduate, non-English majors; and six were graduate students in Applied Linguistics. All were living in Japan at the time of the experiment. As was the case with the first study, the participants in the second study were also divided into low, mid, and high proficiency groups on the basis of the Brown’s (1980) cloze test. Table 8 below summarizes participants’ self-reported scores on standardized English tests where available.

⁴ A standardized English proficiency test developed and administered by the Eiken Foundation of Japan)
Table 8  
*Participants’ Self-Reported Scores of Standardized English Tests in Study Part 2*

<table>
<thead>
<tr>
<th></th>
<th>TOEFL iBT</th>
<th>TOEFL ITP</th>
<th>TOEIC</th>
<th>Eiken</th>
<th>IELTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low (13)</td>
<td>NA</td>
<td>460-490 (2)</td>
<td>450-860* (8)</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>Mid (30)</td>
<td>70-96 (6)</td>
<td>470-598 (23)</td>
<td>690-945 (13)</td>
<td>2-1 (16)</td>
<td>6 (1)</td>
</tr>
<tr>
<td>High (13)</td>
<td>82-113 (7)</td>
<td>583-590 (3)</td>
<td>830-990 (7)</td>
<td>Pre1-1 (9)</td>
<td></td>
</tr>
</tbody>
</table>

Notes. The number in parenthesis indicates the number of participants who reported their scores for each test. *The second highest score on TOEIC in this low proficiency group was 670.

As can be discerned from the partial data available on standardized test scores, it is possible that the low proficiency group in Study Part 2 may have been slightly higher in L2 proficiency than that in Study Part 1, but overall each proficiency group in the two studies show a similar range of scores on the self-reported standardized test scores.

5.3.2 Materials and instruments

In this section, a detailed description of the primary L2 tasks, the secondary task, time estimation, self-assessment questionnaire, background questionnaire, and the proficiency test is given.

5.3.2.1 Primary tasks

Among different models of cognitive task complexity introduced above, cognitive task complexity in this study was initially operationalized along the resource-directing dimension in Robinson’s (2001a) model. This decision was made in an effort to consider how the measures adopted here might contribute to resolving challenges in this particular domain of research (as described in Chapter 2 above). Within this domain, given a large number of previous studies that investigated monologic, narrative tasks, a set of monologic, narrative tasks was selected for the current study as well, to maximize comparisons of the results of the current study with previous studies. As described earlier, there are many possible ways of intentionally manipulating the
level of cognitive task complexity in L2 communication tasks. One of the purposes of this study was to compare multiple tasks with varying degrees of cognitive complexity (rather than a dichotomous approach, simple versus complex task) and to examine the relationship among task design features, the level and type of cognitive task complexity, and, in Study Part 2, learners’ task performance. To create tasks with multiples levels of complexity, the use of the [+/- few elements] dimension was thought to be most natural and appropriate (as well as most common in the domain), as it presumably allows for a gradual increase in the level of cognitive complexity, rather than other variables that are inherently dichotomous (e.g., [here-and-now] versus [there-and-then]). Accordingly, five oral narrative tasks were selected for use in the present study: one practice task and four main tasks. For each of these monologic tasks, participants were instructed to tell a story in English based on a six-framed picture set. To encourage learners’ engagement with the tasks, prior to completing the tasks, they were told that their performances would be assessed based on completeness (i.e., whether all six pictures were included in the story), effectiveness (i.e., whether the order of the pictures was clear to the listener), and creativity.

A picture set for the practice narration task was adapted from Heaton (1966). It contained four characters: a son, a daughter, a mom, and a dog. The story was about the two children going on a picnic without knowing that their dog was hiding in their picnic basket and eating their lunch. The four main tasks included: Alarm Clock (Task 1), Going Fishing (Task 2), Baby (Task 3), and Dog (Task 4). The cartoon picture sets for the first three tasks were adapted from Hill (1960), and the picture set for Task 4 was adapted from Elder and Iwashita (2005). In order to minimize the possible effects of extraneous variables, the number of picture frames was equalized to be six. To this end, two pictures were eliminated from the original sets of Tasks 1
Following Robinson’s (2001a) Cognition Hypothesis, the primary design feature of the tasks was the number of elements (i.e., [+/- few elements]), operationalized in this study according to the number of characters in a story, with the four tasks containing respectively one (a man), two (a man and a girl), four (a mother, a baby, a girl, and a boy), or nine (a dog, four men, a woman, two ambulance attendants, and a policeman) characters. A storyline for each task is found in Table 9.

To be clear, this is not to claim that the number of characters was the only difference among the four picture sets. Pilot studies with 33 learners and 13 native speakers of English revealed that these picture sets also differed in terms of other features, such as storyline clarity, code complexity, and changes in the setting/background (see Table 9). However, it is worth emphasizing that these factors (including the number of elements) may or may not influence the level of cognitive task complexity: Without measuring the level of cognitive load independently from task performance, there is no way of verifying that (or which among) these factors increase or decrease the level of complexity in reality. As a starting point in task design, then, four picture sets were chosen to differ obviously in the number of characters involved in each story, in an attempt to maximize the possible differences in cognitive complexity among the four tasks (as other researchers have done), while keeping in mind that other task features may also affect the level of cognitive complexity assessed by various independent measures. It was also decided that attempting to hold constant other factors within the narrative tasks, and only varying the number of characters, would not have proved particularly insightful: Addition of new characters to what is otherwise the same story would have led to an inevitable practice effect (regardless of order of presentation), never mind challenges to maintaining participants’ interest and effort in
completing the four tasks. Accordingly, four distinct stories were selected (much as a teacher might do), each differing in number of characters involved as a starting point for presuming likely task complexity differences. Of interest to this study, then, was not the verification of \([+/-\) elements\] as the sole causal factor in determining cognitive task complexity; rather, tasks that differed in number of characters were selected on the assumption that there would be detectable differences in measurable cognitive load, and the study set out to utilize distinct measurement (and other) approaches to investigating whether that was in fact the case, as well as what the sources of cognitive task complexity might actually be.
Table 9
Summary of Task Features

<table>
<thead>
<tr>
<th>Task</th>
<th>Source</th>
<th>Number of Characters</th>
<th>Storyline</th>
<th>Storyline Clarity</th>
<th>Code Complexity</th>
<th>Changes in Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Hill (1960) pp. 48–49</td>
<td>1</td>
<td>A man sets the alarm, closes the curtains, and goes to bed at night. In the morning, he wakes up when the alarm goes off, but he is so annoyed that he throws his pillow at the clock. In the end, he goes back to bed again.</td>
<td>Moderate clarity (e.g., “The story seemed straightforward and easy to understand.” “I didn’t notice what happened at first time.”)</td>
<td>Vocabulary: Relatively easy, related to daily life (e.g., light, pillow, alarm clock, sleep, throw)</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>Hill (1960) pp. 10–11</td>
<td>2</td>
<td>A man with a professional outfit and fishing gear is fishing in a river. A girl with a makeshift fishing rod comes along and starts fishing on the riverbank across from him. She keeps catching fish, while the man can only catch trash. Then, the man gets frustrated and ends up catching his own boot. The girl breaks up laughing and loses a fish that she was about to catch.</td>
<td>Relatively unclear storyline (e.g., “I wasn’t sure what the storyline really was.” “I couldn’t see the ‘point’ of the sequence.” “It was hard to recognize the object in Panel 3.”)</td>
<td>Vocabulary: Some specific fishing terms (e.g., fishing, fishing pole, catch a fish); otherwise relatively easy (e.g., man, girl, river, trash, boot)</td>
<td>No</td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>Source</th>
<th>Number of Characters</th>
<th>Storyline</th>
<th>Storyline Clarity</th>
<th>Code Complexity</th>
<th>Changes in Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 3</td>
<td>Hill (1960) pp. 60–61</td>
<td>A mom is reading a book on a chair, while her new-born baby is sleeping by her side. She eventually falls asleep. Then, her two older children come in to the room and decide to play a prank on her. A girl picks up the baby, and a boy puts a ball with a face in the baby’s bed. When the mother wakes up, she is startled to find that the baby turned into a ball.</td>
<td>Clear storyline and ending. (e.g., “It was easier to see the sequence of events in this story.” “You knew exactly what the kids were up to.”)</td>
<td>Vocabulary: Some low frequency words (e.g., crib, prank); otherwise, relatively easy (e.g., baby, brother, sister, ball, exchange, surprise) Function: Describing, narrating</td>
<td>No</td>
</tr>
<tr>
<td>Task 4</td>
<td>Elder &amp; Iwashita (2005) p. 238</td>
<td>In a downtown area, there are two people having coffee outside of a restaurant, one man working on a ladder, and a man walking with his big dog. Suddenly, the dog runs into the street where a car is coming. The car swerves to avoid the dog but instead runs onto the sidewalk. It ends up hitting the man on the ladder. An ambulance comes and carries the injured man away. When the police come, they almost arrest the driver, but he insists that it was the man and the dog’s fault. In the end, the old man is put into prison with his dog.</td>
<td>Relatively unclear storyline. (e.g., “I couldn’t quite figure out the sequence of events. I was still trying to decide what was happening as I was telling the story.” “In Panel 5 it was difficult to tell who the character was.”)</td>
<td>Vocabulary: Related to specific events, in this case accidents (e.g., swerve, sidewalk, ambulance, arrest, jail, police); low frequency words (e.g., ladder, leash) Function: Describing, narrating, cause and effect</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Note. Storyline clarity is based on pilot participants’ retrospective comments for each task. The quotes in parentheses are direct quotations of their comments. The Corpus of Contemporary American English (Davies, 2008) was used to gauge word frequency.
Note that all of picture sets were presented with 12 cm x 19 cm dimensions and were shown on a 18 cm x 29 cm computer screen to make sure that participants could see the details of each picture in each set.

**5.3.2.2 Secondary task**

A secondary, color-detection task was developed following the methodological recommendations in Brünken et al. (2002). Given findings in Schoor et al. (2012) that the letter color-detection task was a more sensitive measure of cognitive load, a letter color-detection task was chosen over a background color-detection task (cf. Révész et al., 2014). Participants were instructed to press the space bar when they detected a color change in the letter ‘A’ presented in the upper part of the computer screen together with the primary task pictures (presented in the center of the screen). Using PsyScope X Build 57 (2010), the letter A was programmed to change from black to red and back to black. The duration of the red letter was set for 5,000 milliseconds, whereas the duration of the black letter was set to be various (from 3,000 to 60,000 milliseconds) so that participants would not be able to predict when the color might change. Both the primary (story-telling) and secondary (color-detection) tasks were presented on the computer screen simultaneously (see hypothetical example in Figure 5). Participants were instructed to respond to the letter-color change only when the color changed from black to red, and ignore the color change from red to black (i.e., they actually had to pay attention, rather than simply react to any color change in either direction). Reaction times in response to the color changes (i.e., pressing the space bar) were recorded by PsyScope X Build 57 in milliseconds.
Figure 5. Screenshot of a hypothetical primary narrative task and the secondary color-detection task.

5.3.2.3 Time estimation

After completing each story-telling task (together with the secondary, color-detection task for the dual-task group), participants were asked to estimate the time they had spent on the task. They were asked, “How long do you think your story was?” in their L1 and filled in _______ minutes _______ seconds. This question was included in the paper-based self-assessment questionnaire (See Appendix A for an English version and Appendix B for a Japanese version). Although participants were not explicitly told to estimate the length of time on task prior to their engagement in the tasks, they very likely knew that they would be asked to do so when completing the four main tasks, given that they had already experienced the whole procedure (i.e., narrate a story – the dual-task group also experienced engaging in the color detection task
simultaneously – then answer questions about time estimation and task difficulty/mental effort) in the practice phase. Hence, I would consider this time estimation approach to be prospective, rather than retrospective (see discussion in Chapter 3).

5.3.2.4 Self-assessment of perceived task difficulty and mental effort

To measure perceived task difficulty and mental effort exerted, participants responded to the following questions, presented in participants’ L1, on a paper-based questionnaire: (a) Was the story-telling task easy? (b) How much mental effort did you put in to complete this task? (See Appendix A & B). They used nine-point Likert scales, ranging from 1 (*very easy/not at all*) to 9 (*very difficult/intensively*) to indicate their responses. These questions were developed following similar questionnaires in Paas et al. (1994) and Robinson (2001b). Although the task difficulty questionnaire originally developed by Robinson (2001b) contained questions not only about task difficulty but also anxiety, self-assessment of task achievement, interest, and motivation, in the current study, only the question that inquires explicitly about difficulty of tasks was used; this is the only question that directly taps into the construct of task difficulty versus other phenomena not in focus for the current study. Additionally, participants were asked to write explanatory comments for their ratings, to further investigate what factors made certain tasks easy or hard to perform. They were given the following prompts:

– What made the story-telling task easy/difficult? Please provide details
– What made the task require low/high mental effort? Please provide details.

Once again, these questions were asked in participants’ L1, and they provided comments in their L1.
5.3.2.5 Proficiency tests

As a quick estimate of L2 learners’ global English proficiency, a cloze test developed by Brown (1980) was administered to all L2 English learners (see Appendix C). The test text was about ‘Man and his progress,’ and every seventh word starting from the third sentence was deleted (for more details, see Brown, 1980). Participants were instructed to read the passage quickly to get the general meaning first, and then to fill in the blanks with the most appropriate English words. They were also told that only one word could be filled in for each blank, and that they should do their best in filling out as many blanks as possible. The cloze test included 50 blanks, and participants were given 25 minutes to complete it. As reviewed earlier, the criterion-related validity coefficient with the UCLA ESL Placement Examination for the cloze test was found to be \( r = .88 \) for the scoring method of exact scoring (Brown, 1980), which was adopted in the current study. Cronbach’s alpha reliability estimate was found to be \( \alpha = .89 \) in Study Part 1 and \( \alpha = .76 \) in Study Part 2 in the current study, suggesting that it would prove consistent in distinguishing among participants.

In addition, participants in the second part of the study also engaged in an oral elicited imitation test in English. This new measure of proficiency was introduced in response to a concern that the cloze test might be more closely associated with receptive skills rather than productive skills given its written modality (e.g., Caulfield & Smith, 1981). The oral elicited imitation test that focuses on learners’ spoken English abilities was thought to be more compatible with the skills needed for completing the narrative tasks (e.g., Ellis, 2005, 2006; Wu & Ortega, 2013). An oral elicited imitation test developed by Ortega, Iwashita, Norris, and Rabie (2002) was used in this experiment. In this test, participants were asked to repeat each in a series
of sentences that they heard in English as precisely and as completely as possible (See Appendix D for a list of prompts). They first read and then heard the instructions aurally in their first language, Japanese (see Appendix E for the written instructions in Japanese and Appendix F for its translation in English). They then practiced repeating five sentences in Japanese. The main part of the test then included 29 sentences in English for participants to repeat.\footnote{The original elicited imitation included 30 items. Due to the poor quality of the recording, however, one item was eliminated for use in this study.} The test as a whole lasted for approximately 10 minutes. In this elicited imitation test, the sentences presented increased in length and the ordinal scoring method was used. These are two characteristics that Yan et al. (2015) identified to be important to make a more discriminating elicited imitation test. Cronbach’s alpha reliability estimate was found to be $\alpha = .97$ in the current study, suggesting that it would prove quite consistent in distinguishing among participants.

5.3.2.6 Background questionnaire

A background questionnaire for L1 speakers included questions about the following factors: gender, age, educational background, and first language (see Appendix G). A questionnaire for L2 English speakers asked questions including: gender, age, educational background, first language, and most recent TOEFL/IELTS/TOEIC/Eiken scores (see Appendix H).

5.3.3 Procedure

Prior to the main study, two pilot studies were conducted in order to examine what worked and what did not, and hence what needed improvements, as well as to explore the nature of the four tasks from the participants’ perspectives. In what follows, the details of each pilot study will
be described, as they contributed substantially to the improvement of subsequent data collection procedures for the main study.

5.3.3.1 Pilot study 1

The first pilot study was conducted with three L2-English speakers and three L1-English speakers, who were living in the US at the time of the experiment. L2 English speakers were all female graduate students with various L1s, and their ages ranged from 25 to 32 ($M = 31.67$, $SD = 6.51$). L1 English speakers included one male, undergraduate student and two female, graduate students, and their ages ranged from 21 to 29 with the average being 24.67 years old ($SD = 4.04$). In this first pilot study, participants were given written task instructions that included paragraphs explaining the task procedures and emphasized the importance of focusing on the primary, narrative task, as opposed to the secondary, color-detection task. Subsequently, they engaged individually in the four narration tasks along with the secondary, color-detection task in a sound-proof booth. After each task, they estimated the duration of their story and gave ratings on perceived difficulty and mental effort.

Based on this pilot study, modifications were made to the task procedures and instructions. First, some pilot participants pointed out that the very complex nature of the tasks (i.e., simultaneous engagement in the primary, narrative task and the secondary, color-detection task) made the experiment hard to handle; therefore, to reduce the extraneous cognitive load, a practice task was inserted at the beginning of the experiment to allow participants to experience the task procedures before they engaged in the four main tasks. Second, giving task instructions in a paragraph format seemed to have contributed to participants’ confusion. Therefore, instructions were revised to show task procedures in bullet points and a graphic chart, instead of
paragraphs. The pilot study also revealed that simply telling participants to focus on the primary, story-telling task, as opposed to the secondary, color-detection task, may not have been effective enough to prevent them from focusing on the color-detection task. Based on this observation, the revised task instructions introduced a motivation for focusing on the primary, story-telling task. Specifically, the revised task instructions included the following sentence: “Your story will be assessed on its effectiveness, completeness, and creativity; therefore, please do your best in telling the stories.”

5.3.3.2 Pilot study 2

The second pilot study was conducted with 33 L2-English speakers and 13 L1-English speakers. L2 English speakers included 12 males and 21 females with various L1 backgrounds. Their average age was 25.75 ($SD = 7.09$), ranging from 19 to 50. L1 English speakers, on the other hand, included 7 males and 6 females. They had an age range of 19 to 39 with the average of 25.54 ($SD = 6.88$).

These pilot participants were given the revised task instructions in written English and engaged in the practice task as well as the main four narration tasks. As in the first pilot study, after they engaged individually in each version of the story-telling task with the color-detection task in a sound-proof booth, they estimated the length of their story and gave ratings on perceived difficulty and mental effort. This time, they were also asked to provide explanatory comments for their ratings to find out what they perceived to make the tasks easy or hard to perform.

Observations of this second pilot study and the analysis of the participants’ task performances identified a few new issues that needed to be dealt with. Firstly, it was hard to
ensure that participants were correctly following the task instructions without monitoring them during their engagement in the tasks. Some L2 English participants explicitly said after the experiment that they did not fully understand the instructions or they forgot about the color-detection task. Given this observation, all participants in the subsequent main study were monitored throughout the experiment. They were also given additional instructions if needed, after the practice task, to ensure their understandings about what to do during the experiment. Secondly, two additional revisions were made to the task instructions. The second pilot testing revealed that giving task instructions in participants’ L2 (English) was not an effective way to communicate complex task procedures. In the second, as well as first, pilot studies, participants’ L1s were various; hence, I had no choice but to give the task instructions in English (the target language). However, some of them had a hard time understanding the task procedures and failed to follow them correctly in the experiment (due to their relatively low proficiency in English). Based on these observations, in the main study, participants who share a single L1, Japanese, were recruited and the task instructions were given in their L1. The second issue occurred when one pilot participant improvised and narrated stories in his L1 (rather than English). Given this observation, in the newly revised task instructions, participants were explicitly told to tell stories in English and the phrase ‘in English’ was highlighted.

Additionally, these pilot studies revealed that nobody told a story that is longer than five minutes when no time limit was posed. For this reason, participants in the main study were given up to five minutes to complete each task (although they could have gone on up to six minutes, after which PsyScope was automatically terminated).
5.3.3.3 Procedures in the main study

In the main study, each participant took part in the experiment one by one in a quiet room with only the researcher present but seated behind the participant (so as not to be a distraction). Participants first read a consent form (see Appendix I & J for Part 1 participants in English and Japanese, respectively, and Appendix K & L for Part 2 participants in English and Japanese) and agreed to participate in the study, and then they answered the background questionnaire. Subsequently, participants in the dual-task condition (i.e., speakers who participated in Part 1) were given the task instructions sheet, which included instructions about the secondary color-detection task and was carefully revised based on the two pilot studies as described above (see Appendix M for an English version and Appendix N for a Japanese version). By contrast, participants in the single-task condition (i.e., participants in Part 2) were given the same task instructions sheet, but without the instructions on the secondary color-detection task (see Appendix O for an English version and Appendix O for a Japanese version).

After reading the task instructions, all participants were asked whether there was anything unclear about the instructions, and once their understandings were confirmed, participants engaged in the practice story-telling task. (They were explicitly told that this one was a practice task). While they were engaging in the practice story-telling task (together with the color detection task for participants in the dual-task condition), the researcher monitored them to make sure that they indeed understood the task instructions and were following them appropriately, and they were given additional instructions following the practice task as needed. Participants in the single task condition rarely needed additional instructions; on the other hand, the most common additional instructions given to participants in the dual-task condition were about the
color detection task. For example, some participants were told not to forget hitting the space bar when the color change happens, to hit the space bar when the color of the letter changes from black to red, not the other way around, and so on. All participants filled out the questionnaire, which included questions about time estimation, task difficulty, and mental effort, as soon as they finished telling the story.

Subsequently, participants engaged in the four main narrative tasks. Participants who were assigned to the dual-task condition simultaneously worked on the secondary, letter color-detection task while telling the stories. The researcher took observation notes throughout the experiment to identify any remaining anomalies or unexpected interruptions to the procedure. The order of the four main tasks was randomized by PsyScope X Built 57. Before each narration (including the practice task), participants were given 30 seconds to familiarize themselves with the pictures and figure out the storyline, and they were given up to five minutes to complete each task. After each story-telling task, participants estimated how long they thought their story had been (i.e., time estimation), completed the self-assessment questions about perceived task difficulty and mental effort, and provided explanatory comments for their ratings.

After participants finished all five tasks, all L2 English participants also took the cloze test in English (for 25 minutes). L2 English participants in Study 2 additionally engaged in the oral elicited imitation task in English (for 10 minutes). The whole experiment took about half an hour for each L1 participant, and approximately an hour for each L2 participant.

5.3.4 Measurement and data analysis

In order to address the research questions posed for the first part of the dissertation study (i.e., validation of the relationship between task design and cognitive task complexity), diverse
measures of cognitive load, adopted from cognitive psychology, were used. On the other hand, measures of linguistic quality were used to address the research questions of Study Part 2 (i.e., investigation of the relationship between cognitive task complexity and task performance).

5.3.4.1 Measures and data analysis for Study Part 1

Independent measures of cognitive load used in this study included: (a) dual-task methodology, (b) time estimation, (c) task difficulty questionnaire, and (d) mental effort questionnaire. For specific measures extracted, see Table 10 below.

Table 10
Specific Measures Extracted

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Specific measures extracted for each technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dual-task methodology</td>
<td>Reaction time of accurate responses Error Rate 1 Error Rate 2</td>
</tr>
<tr>
<td>Time estimation</td>
<td>Difference</td>
</tr>
<tr>
<td>Task difficulty</td>
<td>Likert-scale rating Open-ended responses</td>
</tr>
<tr>
<td>Mental effort</td>
<td>Likert-scale rating Open-ended responses</td>
</tr>
</tbody>
</table>

5.3.4.1.1 Dual-task methodology

Participants’ reaction times to the secondary, color-detection task were recorded every time they pressed the space key to indicate their noticing of the color changes. Their responses to the color-detection task were analyzed in two ways: (a) reaction time of accurate responses, and (b) two types of error rates, following previous studies (e.g., Baddeley & Logie, 1999; Logie, Baddeley, Mane, Donchin, & Sheptak, 1989). Accurate responses are defined as participants’ responses to the color change when it changed from black to red. The average reaction time of accurate responses for each participant on each task was calculated by dividing the summed
reaction time of accurate responses to the color change by the number of accurate response instances. All inaccurate responses (i.e., responses to the color change when it changed from red to black or responses made the second time or more to the color change when it changed from black to red) were ignored in this calculation, as erroneous responses are captured by the error rates (described below), and the rate of these erroneous responses does not reflect the amount of cognitive load imposed by the tasks in the same way as the rate of accurate responses does. In terms of error rates, the first type was a rate of errors made by not responding to the color change when required (hereafter Error Rate 1). It was calculated by adding the number of instances where each participant missed the color change, dividing it by the number of color changes s/he had encountered in total, and multiplying it by 100.6 The second error rate was a rate of errors made by responding to the color change when not required (when it changed from red to black) (hereafter Error Rate 2). It was calculated by dividing the number of instances where each participant erroneously hit a computer key (i.e., hitting a computer key when the color of the letter A was black) by the number of occurrences of the black letter, and multiplying it by 100.

5.3.4.1.2 Time estimation

Following Baralt (2010, 2013) and previous studies in cognitive psychology, in order to calculate the average difference (in seconds) between participants’ estimated time and actual time on task, each participant’s actual time on a task (recorded by PsyScope X Build 57) was subtracted from their estimated time. For example, if a participant estimated the time on task to

---

6 Depending on how long their narratives went on, the participants encountered different numbers of instances of color change across different tasks, ranging from two to 29 (for English-L2 speakers) and four to 42 (for native speakers of English). The longer their narratives went on, the more instances of color change they encountered as a result.
be five minutes when in fact his/her actual time on task was six minutes, s/he would get a score of −1.

5.3.4.1.3 Self-assessment of task difficulty and mental effort

Self-assessment of task difficulty and mental effort consisted of two parts: numeric ratings and open-ended responses.

5.3.4.1.3.1 Ratings

The questionnaire included one question each for perceived task difficulty and mental effort (see the Materials section). For each question, participants gave 9-point Likert-scale ratings to indicate their responses. These ratings were analyzed separately from each other, for each task.

5.3.4.1.3.2 Open-ended responses

Participants’ explanatory comments were analyzed qualitatively for emerging categories, separately for task difficulty and mental effort. I first looked through all open-ended responses and identified emerging themes (e.g., number of elements, storyline clarity/complexity, code complexity, task instruction-related issues). I then categorized all responses into these themes, adding new ones (e.g., change of background, familiarity, clarity of pictures, learner-generated factors) when necessary. Finally, I identified and combined related themes to create superordinate categories that were as mutually exclusive as possible. As a result, I ended up with three superordinate categories: (a) conceptual input, (b) code complexity, and (c) performance factors. Conceptual input is concerned with the task prompt (i.e., picture sets), and it is further divided into (a) storyline/picture quality, (b) simple input making task difficult, (c) number of characters, (d) logic, (e) familiarity/relatability, and (f) background. Code complexity is about
linguistic aspects of task performance. Performance factors are concerned with issues related to participants’ own overall performance quality, rather than the conceptual input or the language required to tell a story, and they are further divided into two sub-categories: (a) task-instruction-related factors and (b) learner-generated factors. Finally, I also added a category ‘other’ which included comments related to (a) the fact that participants’ got used to the format of the narrative task as they progressed through the experiment, (b) the fact that they were not allowed to write down what they would say during the 30-second planning time, and (c) affective factors (e.g., panicked, nervous), as well as comments (d) that could not be grouped into any of the three main superordinate categories (e.g., “難しい。とても頭を使った。 [Difficult. I used my brain a lot.]”)

Table 11 below gives definitions of each subordinate category with some concrete example comments.
<table>
<thead>
<tr>
<th>Superordinate categories</th>
<th>Subordinate categories</th>
<th>Definitions</th>
<th>Example comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conceptual input</td>
<td>Storyline/picture</td>
<td>quality</td>
<td>Comments about (a) clarity/complexity of storyline and (b) clarity of input (pictures)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>● 絵がシンプルで説明しやすかった [The pictures were simple, and it was easy to describe them.] (Task 1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>● 絵が細かくて、3 コマ目で男の人が何を釣り上げているのか判別できなくて焦った [The picture was so fine that I couldn’t figure out what the man was catching in the third frame and I got panicked.] (Task 2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>● 4 枚目と5 枚目のイラストを違うように説明できなかった [I could not describe the fourth and fifth frames differently.] (Task 3)</td>
</tr>
<tr>
<td>Simple input making</td>
<td>A special case of the</td>
<td></td>
<td>Comments where learners argue that a simple input made the task difficult to deal with</td>
</tr>
<tr>
<td>task difficult</td>
<td>first category;</td>
<td></td>
<td>● 絵が淡々としたものだったため、自分で話を作らなければならない気持ちがした [As the pictures were simple and plain, I felt that I had to create a story on my own.] (Task 1)</td>
</tr>
<tr>
<td></td>
<td>Comments where</td>
<td></td>
<td>● ただ男の人が寝てるだけの絵など、あまり変化がないコマが多かったため、瞬時にストーリーを理解ことができなかった [There were a lot of frames that didn’t entail much action, like the frame where a man is simply sleeping; so, it was difficult to understand the storyline right away.] (Task 1)</td>
</tr>
<tr>
<td>Number of characters</td>
<td>Comments about the</td>
<td></td>
<td>登場人物が多く、それを区別するのが難しい [There are a lot of characters, and it is difficult to distinguish among them.] (Task 4)</td>
</tr>
<tr>
<td></td>
<td>number of elements (e.g., characters, events) involves in the story, and consequences for having multiple characters</td>
<td></td>
<td>● 少年と少女が母親の子供であると仮定したときに、彼らをどう表現したらいいか迷った [If we assume that the boy and the girl are the mother’s children, I wondered how I should address them.] (Task 3)</td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>Superordinate categories</th>
<th>Subordinate categories</th>
<th>Definitions</th>
<th>Example comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conceptual input</td>
<td>Logic</td>
<td>Comments about knowing or not knowing why things happen the way they do in the picture strips</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>・男性が外を覗いた理由を思いつけなかった [I couldn’t figure out why the man looked outside.] (Task 1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>・ボーリングボールがなぜ出てきたのか、どこから出てきたのか分かりなくて説明に悩んだ [I didn’t know where the bowling ball came from and why, so I didn’t know how to describe the pictures.] (Task 3)</td>
</tr>
<tr>
<td></td>
<td>Familiarity/relatability</td>
<td>Whether the storyline depicted by each task is familiar or relatable to them. This category does not include word familiarity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Background</td>
<td>Whether the background changes throughout the picture strips</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Code complexity</td>
<td>Comments about linguistic aspects (e.g., general proficiency, grammar, lexis, complexity, accuracy, fluency)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Performance factors</td>
<td>Comments that are related to (a) creativity, (b) completeness, (c) effectiveness</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Task instruction-related issues</td>
<td></td>
<td>・わからない単語を言い換えようとしたから [Because I tried to paraphrase the words that I didn’t know] (Task 2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>・創造性を考えるのが難しい [I was difficult to make my story creative.] (Task 1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>・イラストの順序を的確に説明しようとしたら [I tried my best to make the order of the pictures clear.] (Task 3)</td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>Superordinate categories</th>
<th>Subordinate categories</th>
<th>Definitions</th>
<th>Example comments</th>
</tr>
</thead>
</table>
| Performance factors     | Learner-generated factors | Comments that reflect (a) participants’ strategies to deal with high cognitive activity and (b) their attempts to perform well | • 分からない単語は考えても出てこないので、半分あきらめが入っている [I’m half giving up because I cannot think of words that I don’t know.] (Task 1)  
• うまく表現しようと思って、余計に負担がかかった [I gave myself extra burden for trying to describe well.] (Task 2)  
• 一つ前の課題では、時制を現在形にしてとても混乱したので過去形で統一しようとしたのが良かった [I was really confused when I was telling the previous story because I decided to use the present tense, so I decided to stick to the past tense this time, and this strategy worked well.] (Task 3) |
| Other                   | (No subordinate category) | Any comments other than the above | • 少しずつだが、この形式に慣れてきたから [Although little by little, I am getting used to this (task) format.] (Task 1)  
• 紙に書いて準備せずに作るのは難しい [It is difficult to create a story without preparing it on paper.] (Task 2)  
• 準備の時に考えていたことを言う前に次の絵の説明を始めてしまい、少し動揺してしまった [I was panicked a bit because I started describing the next picture before telling all that I have prepared (about the previous picture).] (Task 1)  
• 難しい。とても頭使った。[Difficult. I used my brain a lot.] (Task 4) |
Based on these definitions, I counted the number of occurrences of comments that belonged to each superordinate and subordinate categories. All these comments were also color-coded for ease of distinction. Comments used to justify why the task at hand was easy or required little mental effort were coded in blue, whereas comments given to explain why the task was difficult or required much mental effort were coded in red. At the time of the data collection, I went through each participant’s comments quickly to see if it was clear whether they were supposed to explain why the task was easy or difficult (or required little or a lot of mental effort). When not clear, I asked participants to clarify and took notes. For this reason, this color-coding was relatively straightforward. When still in doubt, the participant’s rating for that particular comment was referred to and compared with the rest of his/her ratings. If the rating was lower relative to the other ratings (for the other tasks), the comment was judged to be a way of justifying why the task was easy/less mental effort consuming. A rating higher than the other ratings, in turn, was interpreted to be provided to justify why the task was difficult/more mental effort consuming. In reality, however, I rarely had to refer to these ratings, as the comments were typically clear in this regard.

5.3.4.1.4 Inter-coder reliability

Among the measures used in Study Part 1, only the coding of open-ended responses required high inference. One native speaker of Japanese, who is a researcher in the field of Applied Linguistics, served as a second coder. The second coder went through a half-an-hour training, where he and I reviewed the coding guidelines (see Appendix Q) with some concrete coded examples. He was encouraged to ask questions along the way, and any uncertainties were resolved in this training session. Subsequently, he coded 10% of the open-ended responses data.
(i.e., 12 participants’ responses for the four tasks). Based on these codings, inter-rater reliability was calculated to be 78.42 for the content and 94.24 for the color-coding (i.e., whether the comment is provided to justify that the given task was easy/less mental effort consuming or difficult/more mental effort consuming). This inter-rater reliability was calculated by dividing the number of items where a discrepancy in category or color-coding was observed by the total number of items coded and multiplying it by 100. Inspecting his codings, it became apparent that the distinction between code complexity and performance factors was sometimes tricky. For example, a comment like “より詳しく絵を説明しようと思うと、文法的にめちゃくちゃな言い回しかできなかった [When I tried to explain the pictures in details, I could only use inaccurate grammar],” is a result of learners’ attempt to tell a more detailed story, but the consequence lies in the participant’s concern about the language use. For this reason, code complexity was further divided into (a) task-induced code complexity, (b) effort for a better performance (language-related), and (c) strategy to deal with a given task (language-related). Additionally, performance factors were divided into (a) effort for a better performance (performance-related), and (b) strategy to deal with a given task (performance-related), instead of task instruction-related issues and learner-generated factors, because the latter distinction was rather hard to make and the new distinction was thought to yield a more internally consistent category. In the end, the coding scheme for the open-ended responses was finalized as shown in Table 12.
### Table 12

**Superordinate/Subordinate Categories and Their Definitions and Examples (Revised)**

<table>
<thead>
<tr>
<th>Superordinate categories</th>
<th>Subordinate categories</th>
<th>Definitions</th>
<th>Example comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conceptual input</td>
<td>Storyline/picture quality</td>
<td>Comments about (a) clarity/complexity of storyline and (b) clarity of input (pictures)</td>
<td>- 絵がシンプルで説明しやすかった [The pictures were simple, and it was easy to describe them.] (Task 1)&lt;br&gt;- 絵が細かくて、3コマ目で男の人が何を釣り上げているのか判別できなくて焦った [The picture was so fine that I couldn’t figure out what the man was catching in the third frame and I got panicked.] (Task 2)&lt;br&gt;- 4枚目と5枚目のイラストを違うように説明できなかった [I could not describe the fourth and fifth frames differently.] (Task 3)</td>
</tr>
<tr>
<td>Simple input making task difficult</td>
<td>A special case of the first category (i.e., storyline/picture quality) Comments where learners argue that a simple input made the task difficult to deal with</td>
<td>- 絵が淡々としたものだったため、自分で話を創らなければならない気がした [As the pictures were simple and plain, I felt that I had to create a story on my own.] (Task 1)&lt;br&gt;- ただ男の人が寝てるだけの絵など、あまり变化がないコマが多くかったため、瞬時にストーリーを理解することができなかった [There were a lot of frames that didn’t entail much action, like the frame where a man is simply sleeping; so, it was difficult to understand the storyline right away.] (Task 1)&lt;br&gt;- 登場人物が多く、それを区別するのが難しい [There are a lot of characters, and it is difficult to distinguish among them.] (Task 4)</td>
<td></td>
</tr>
<tr>
<td>Number of characters</td>
<td>Comments about the number of elements (e.g., characters, events) involves in the story, and consequences for having multiple characters</td>
<td>- 少年と少女が母親の子供であると仮定したときに、彼らをどう表現したらいいか迷った [If we assume that the boy and the girl are the mother’s children, I wondered how I should address them.] (Task 3)</td>
<td></td>
</tr>
<tr>
<td>Superordinate categories</td>
<td>Subordinate categories</td>
<td>Definitions</td>
<td>Example comments</td>
</tr>
<tr>
<td>--------------------------</td>
<td>------------------------</td>
<td>-------------</td>
<td>-----------------</td>
</tr>
</tbody>
</table>
| Conceptual input         | Logic                  | Comments about knowing or not knowing why things happen the way they do in the picture strips | • 男性が外を覗いた理由を思いつけなかった [I couldn’t figure out why the man looked outside.] (Task 1)  
• ボーリングボールがなぜ出てきたのか、どこから出てきたのか分からなくて説明に悩んだ [I didn’t know where the bowling ball came from and why, so I didn’t know how to describe the pictures.] (Task 3) |
|                         | Familiarity/relatability | Whether the storyline depicted by each task is familiar or relatable to them. This category does not include word familiarity | • 魚の大きさも大きいのか小さいのか、知識がないためにストーリーを作るのが大変だった [With regards to the size of the fish, I didn’t have enough knowledge to judge whether they were big or small, so it was hard to create a story.] (Task 2)  
• 同じ場面でストーリーが組み立てられていた [The story took place in the same scene.] (Task 1) |
|                         | Background              | Whether the background changes throughout the picture strips | |
|                         | Task-induced            | Comments about linguistic aspects (e.g., general proficiency, grammar, lexis, complexity, accuracy, fluency) | • 英語ができない [I cannot speak English.] (Task 3)  
• 大がつっこむ、犬が飼い主を振りまわすの表現がうまくでこなかったから [I couldn’t remember how to say “the dog ran into something” or “the dog gave his owner lots of troubles.”] (Task 4)  
• わからない単語を言い換えるようとしたから [Because I tried to paraphrase the words that I didn’t know] (Task 2)  
• 頭の中でストーリーを考えるのはできたが、いざそれを伝えようとすると、主語だったり動詞の時制だったりを合わせるのが難しかった [I could think of a story in my head, but when I tried to actually tell the story, it was difficult to conjugate verbs appropriately and use the right verb tense.] (Task 1)  
• 同じ事は説明するために違った言い方を考えることができなかったが、それを考えようとして頭を使った [I could not think of phrases to explain the same actions but I used much mental effort to try to come up with them.] (Task 2) |
<p>|                         | Effort for a better performance (language-related) | Comments about linguistic aspects that reflect participants’ attempts to perform well | |
|                         |                        |             | (continued) |</p>
<table>
<thead>
<tr>
<th>Superordinate categories</th>
<th>Subordinate categories</th>
<th>Definitions</th>
<th>Example comments</th>
</tr>
</thead>
</table>
| Code complexity          | Strategy to deal with a given task (language-related) | Comments about linguistic aspects that reflect participants’ strategies to deal with a given task | - 分からない単語は考えても出てこないので、半分あきらめが入っている [I’m half giving up because I cannot think of words that I don’t know.] (Task 1)  
- 一つ前の課題では、時制を現在形にしてとても混乱したので過去形で統一しようとしたのが良かった [I was really confused when I was telling the previous story because I decided to use the present tense, so I decided to stick to the past tense this time, and this strategy worked well.] (Task 3)  
- 全てのイラストについて説明するのが難しく、抜かしてしまった所もあった（ので、頭を使わなかった） [It was difficult to describe all the pictures, and I had to skip some parts (so I didn’t use much mental effort).] (Task 4) |
| Performance factors      | Effort for a better performance (performance-related) | Comments about performance quality that is related to (a) creativity, (b) completeness, (c) effectiveness  
Comments about performance quality that reflect participants’ attempts to perform well  
Comments about performance quality that reflect participants’ strategies to deal with a given task | - 創造性を考えるのが難しい [I was difficult to make my story creative.] (Task 1)  
- イラストの順序を的確に説明しようとしたから [I tried my best to make the order of the pictures clear.] (Task 3)  
- うまく表現しようと思って、余計に負担がかかった [I gave myself extra burden for trying to describe well.] (Task 2)  
- 最初の 30 秒の内に、ストーリーこそ考えつきましたが、ふくらまずのは無理だと思い余計な事は考えないようにしました。[I could come up with a bare story within 30 seconds (during the preparation time), but I thought it would be impossible to elaborate on it, so I decided not to think of anything else.] (Task 2)  
- シンプルな内容の絵で、やろうと思えばいくらでも創造性のあるストーリーができただろうに、ただ状況を説明するだけになってしまったら（頭を使わなかった） [The pictures depicted a simple story, so I could have made the story creative if I tried to, but I could only describe the situations (and hence I did not put in much mental effort).] (Task 1) |
|                         | Strategy to deal with a given task (performance-related) | Comments about performance quality that is related to (a) creativity, (b) completeness, (c) effectiveness  
Comments about performance quality that reflect participants’ attempts to perform well  
Comments about performance quality that reflect participants’ strategies to deal with a given task | |
<table>
<thead>
<tr>
<th>Superordinate categories</th>
<th>Definitions</th>
<th>Example comments</th>
</tr>
</thead>
</table>
| Other                    | Any comments other than the above | ● 少しずつだが、この形式に慣れてきたから [Although little by little, I am getting used to this (task) format.] (Task 1)  
● 紙に書いて準備せずに作るのは難しい [It is difficult to create a story without preparing it on paper.] (Task 2)  
● 準備の時に考えていたことを言う前に次の絵の説明を始めてしまい、少し動揺してしまった [I was panicked a bit because I started describing the next picture before telling all that I have prepared (about the previous picture).] (Task 1)  
● 難しい。とても頭を使った。[Difficult. I used my brain a lot.] (Task 4) |

*Note.* Revised parts are surrounded by a dotted line.
5.3.4.2 Measures and data analysis for Study Part 2

With the exception of the dual-task measures of reaction times, all measures and analyses reported above for Study Part 1 were also utilized for Study Part 2. In addition, L2-English participants’ task performances were assessed by: syntactic complexity, accuracy, lexical variety, and fluency (CALF) indices. L1-English participants’ performances were assessed by the same indices, except for accuracy, given the assumption that L1 speakers should be able to use grammatical structures accurately in performing the narrative tasks, and the fact that grammatical accuracy of L1 speakers is indeed seen as the target in analyzing learners’ use of the language.

The recordings of participants’ task performances were first transcribed by the researcher and her colleague, who was a master’s student in TESOL. In what follows, I will describe each measure of accuracy, fluency, and complexity (including both syntactic and lexical measures).

5.3.4.2.1 Unit of analysis

In using CALF indices, T-unit was chosen as a primary unit of analysis. Following Hunt (1970), it was defined as “a main clause plus all subordinate clauses and non-clausal structures attached to or embedded in it” (p. 4). There are other units of analysis available in SLA research, such as C-unit and AS-unit. These units of analysis were created in order to deal with the elliptical nature of spoken data. In this study, the T-unit was chosen over C-unit or AS-unit, because the narrative tasks used in this study were one-way, monologic tasks and hence, only a few elliptical utterances were expected, and there was no interruption of the narration by interaction with other interlocutors.
5.3.4.2.2 Accuracy

Accuracy is argued to be the “most straightforward and internally consistent construct” (p. 4) among the three constructs of CAF (Housen, Kuiken, & Vedder, 2012) (the construct of lexical variety is included in the construct of complexity). In this study, it refers to grammatical and lexical accuracy and concerns how error-free someone’s language production is. To this end, the error-free T-unit ratio was chosen for the analysis of global, grammatical/lexical accuracy. This measure of accuracy has been used in other task complexity studies, such as Gilabert (2007b), Ishikawa (2007, 2008), and Robinson (1995). Grammatical accuracy concerns how accurate the use of language structures is and includes structural phenomena, such as articles, verb tenses, conjugations, and so on. Grammatical accuracy was judged only within the level of a given sentence, and suprasentential factors (e.g., cohesive ties, pragmatics) were not taken into account. For this reason, it was not counted as an error when different verb tenses were used within a single task performance (between sentences) or when common nouns were substituted for pronouns, as long as the language use was deemed accurate within a single T-unit. Lexical accuracy, on the other hand, is about the extent to which the choice of lexis is accurate in the given context. Preposition errors are included as part of lexical errors. For both types of accuracy, self-corrections were coded as accurate if learners were able to correct their errors accurately and obviously, that is, immediately following the error. A native speaker of English, who was a highly experienced SLA researcher and coder of this kind of data, conducted all of the accuracy analyses. To calculate the overall ratio of error-free T-unit to the total number T-units, the number of error-free T-units was divided by the total number of T-units and multiplied by 100.
5.3.4.2.3 Fluency

Fluency is concerned with the flow of a stream of speech (or sometimes writing) and refers to how quickly one can access or deploy his/her L2 knowledge. Fluency is a multi-dimensional phenomenon and can be divided into “breakdown (dys)fluency, indexed by pausing,” “repair (dys)fluency, indexed by measures such as reformulation, repetition, false starts, and replacements,” and “speed, with measures such as syllables per minute” (Skehan, 2009, p. 512-513). In the domain of language testing, there is a new movement to treat pauses and dysfluency markers not as evidence of dysfluency, but as communicational strategies (e.g., Liyange & Gardner, 2013). This is based on the argument that pauses and dysfluency markers used by L1 speakers are a manifestation of the strategy use by the speaker in an attempt to deal with new formulation of speech and high cognitive activity, rather than of his/her inability to produce fluent speech (Goldman-Eisler, 1968). Given these new understandings of breakdown and repair fluency, fluency in this study was analyzed simply in terms of speed. Following Freed, Segalowitz, and Dewey (2004), it was operationalized as the number of words per second. In counting the total number of words, all dysfluency markers, namely reformulations (i.e., “phrases or clauses that are repeated with some modification either to syntax, morphology, or word order,” (Skehan & Foster, 1999, p. 107)), repetitions (i.e., “immediate and verbatim repetition of a word or phrase”), false starts (i.e., “utterances that are abandoned before completion”), and replacement (i.e., “lexical items that are substituted for another”), were excluded. A fluency ‘score’ was calculated by dividing the total number of words produced by an individual on a given narrative by time on task in seconds. Note that only actual time used during the available five minutes per task was involved in the calculation. Each participant’s
time on task was tracked and recorded by PsyScope. Additionally, the total number of words produced was also used as a quantity-based fluency measure (e.g., Adams & Nik, 2014).

5.3.4.2.4 Complexity

Complexity is a complex phenomenon and concept (Bulté & Housen, 2012; Housen & Kuiken, 2009; Norris & Ortega, 2009; Pallotti, 2009). In the field of SLA, it has been treated in two different ways: syntactic complexity and lexical complexity (or variety).

Syntactic complexity refers to how complicated or developmentally advanced one’s structures or a unit of analysis (e.g., T-unit, clause) are considered to be. Norris and Ortega (2009) argued for measuring syntactic complexity multi-dimensionally, especially when analyzing utterances produced by L2 learners with a variety of proficiency levels. They contended that (a) length (e.g., how many words on average does one T-unit include) and (b) amount of subordination (e.g., how many clauses on average does one T-unit include) differentially capture L2 learners’ ability to complexify their utterances at distinct levels of proficiency. Mean length of clauses, which reveals learners’ syntactic complexity at the subclausal or phrasal level, is considered to be an appropriate measure to assess advanced language learners’ syntactic complexity. Studies have shown that advanced language learners are able to complexify their sentences within a clause or a phrase, mostly through nominalization (i.e., creating noun phrases), and hence tend to use subordination proportionally less.

Subordination measures, on the other hand, are more likely to be suitable for less proficient (e.g., intermediate proficiency) learners, because at this stage of proficiency, the amount of subordination is expected to grow as proficiency goes up. Within the length-based measures, Norris and Ortega also argued that different measures tap into distinct dimensions of the
“Any length-based metric with a potentially multiple-clausal unit of production in the denominator” (p. 561) is claimed to measure global complexity, whereas ‘mean length of clause’ is a measure that can capture subclausal or phrasal complexity (Norris & Ortega, 2009).

Following Norris and Ortega’s (2009) suggestions, syntactic complexity was analyzed by the following measures: (a) mean length of T-unit (MLT), (b) mean length of clause (MLC), and (c) clauses per T-unit (C/T). Mean length of T-unit, a length-based measure of global syntactic complexity, was chosen to assess speakers’ overall syntactic complexity. The number of words included in each T-unit was averaged across all T-units to calculate the mean length of T-unit. Mean length of clauses, a length-based measure of syntactic complexity at phrasal level, was chosen in particular to assess to what extent higher proficiency learners complexify their sentences typically through the use of nominalizations. Lastly, clauses per T-unit, a measure of subordination-based complexity, was chosen to capture especially lower level learners’ attempt to complexify their sentences, and calculated by counting the number of clauses per T-unit and dividing it by the total number of T-units.

Clauses are defined in this study to include ‘independent clause’ and ‘dependent clause.’ An independent clause, as the name suggests, can function on its own and is essentially a sentence. A dependent clause, by contrast, cannot stand alone and includes ‘coordination’ (e.g., I went to a store and bought bread), ‘subordination’ (e.g., The dog barked when the criminal broke into the house, He said that he would be late), and ‘embedding’ (e.g., Jane likes the restaurant where she had her birthday party last year) (Dependent clauses underlined; Celce-Murcia & Larsen-Freeman, 1999, p. 20).
Lexical complexity, or variety, was assessed by: (a) TTR and (b) $D$. The $D$ measure was chosen, given that it is the least sensitive measure of lexical variety to the length of speech. The $D$ measure has been used in a few previous studies, such as Malicka (2014), Révész (2011), and Tavakoli (2009), and TTR in Robinson (1995) among other studies. In the current study, it was presumed that different tasks would elicit distinct lengths of speech and that speakers with different proficiency levels would produce a variety of lengths of speech. For this reason, the length of speech needed to be taken into consideration when interpreting lexical phenomena. $D$ was calculated automatically by the Computerized Language Analysis (CLAN) programs (MacWhinney, 2000). To calculate $D$ on CLAN, the following procedures were taken:

1. Dysfluency markers, namely reformulations, repetitions, false starts, and replacement, were all excluded from the transcription.

2. Each participant’s task performance for each task was saved separately as a simple text file.

3. Each simple text file was converted into a ‘.cha’ document, so that CLAN can execute a number of automated functions.

4. Finally, to calculate the $D$ value, type in ‘vocd’ followed by a file name (e.g., NS1T1.txtin.cha)

The advantage of $D$ is that it is calculated in a way that is not as affected by text length as are many of the other lexical measures, such as Type-Token Ratio (TTR) and Guiraud’s Index. $D$ is calculated by:

Plot[ting] the TTR versus token curve…., deriving each point from an average of 100 trials on sub samples of words of the token size for that point” and “find[ing] the best fit between
the ideal curves of theory and the curves drawn from empirical data by a curve-fitting procedure which adjusts the value of the parameter … until a match is obtained between the actual curve for the transcript [of the language sample] and the closest number of the family of curves represented by the mathematical model. (Malvern, Richards, Chipere, & Durán, 2004, p. 55-56)

One disadvantage to the $D$ measure, however, is that it requires 50 words at minimum to allow its calculation. Some participants, especially L2-English speakers with low proficiency levels, were expected not to produce enough words on each of the four tasks for $D$ to be calculated. To compensate for this, TTR, which does not require a certain number of words for its calculation, was also measured.

5.3.4.2.5 Inter-coder reliability

Measures that require high inferences were dual-coded by a second coder. These measures included: (a) identifications of T-units and clauses and (b) accuracy scoring. A second coder (i.e., an English-L1 PhD student in Linguistics) coded 10% of the L2 speaker data (i.e., 12 participants’ performances on the four tasks). The second coder participated in an hour-long coder training. In this training, we went through a second-coder training manual (see Appendix R) and practiced coding a handful of learner task performances. Any questions, uncertainties, and discrepancies in coding were resolved immediately in the training session. After the training, he coded 12 participants’ performances, and an inter-rater reliability was calculated based on this round of coding to be 93.89 for T-units, 92.97 for clauses, and 87.31 for accuracy. This inter-rater reliability was calculated by dividing the number of items where a discrepancy in coding was observed by the number of items coded and multiplying it by 100.
5.3.4.3 Proficiency test

In this section, how the two proficiency tests, a cloze test and an oral elicited imitation test, were scored will be discussed in detail.

5.3.4.3.1 Cloze test

The cloze test was scored by using an exact scoring method. In other words, participants’ responses were scored as correct only when they were exactly the same as the original text. One point was given for the correct answer and zero points were given otherwise. Although the acceptable-answer scoring method may increase test reliability (Brown, 1980), given the already high reliability of the cloze test ($\alpha = .89$ in Study Part 1 and $\alpha = .76$ in Study Part 2), the exact scoring method was chosen for ease and reliability of scoring. Correct answers were summed to create a total proficiency test score ranging from 0 to 50 points.

5.3.4.3.2 Elicited imitation test

The elicited imitation test, on the other hand, was scored by using a five-point scoring rubric, following Ortega et al. (2002):

- 4 = Perfect repetition
- 3 = Accurate content repetition with some (un-)grammatical changes
- 2 = Changes in content or changes in form that affect content
- 1 = Repetition of half or less of the stimulus
- 0 = Silence, only one word repeated, or unintelligible repetition

The scoring guidelines developed by Ortega et al. were slightly modified to fit the context of the present research (see Appendix S). The changes made to the original guidelines are illustrated below:
1. The original scoring guidelines indicated that minimal repetition, for example only one content word plus function word(s), should be scored as zero. Given that the length of a sentence in the elicited imitation test grows as the item difficulty increases, there included short items that only contain a few content words (or idea units). For instance, if the prompt is “I have to get a haircut,” a response of “I have to get” would be scored as zero, according to the original scoring guidelines, which seems quite unfair because only the words missing are an article and ‘haircut.’ For this reason, I have revised this part of the guidelines as follows: If the item is short and contains only 2-3 idea units, do not apply the above rule. For these short items, as long as 2 or more words (except for articles) are repeated, score 1.

2. From the criteria of Score 1, I deleted the following statements:
   - When barely half of lexical words get repeated and meaningful content results that is unrelated (or opposed) to stimulus, frequently with hesitation markers
   - Or when string doesn’t in itself constitute a self-standing sentence with some (targetlike or nontargetlike) meaning (This may happen more often with shorter items, where if only 2 of 3 content words are repeated and no grammatical relation between them is attempted, then score 1)

Both of these criteria were thought to be covered by the overarching criterion of Score 1: Score one “when only about half of idea units are represented in the string but a lot of important information in the original stimulus is left out.” The deletion of these additional criteria was also thought to make the scoring easier to do, and perhaps less various.
3. For the Score 3 criteria, I added four instances of acceptable substitutions, to make the guidelines clearer:
   - Omission of ‘at all’ in Item 30 (“There are a lot of people who don’t eat anything at all in the morning”) is acceptable
   - Omission of adverbs that only express emphasis (e.g., always) is acceptable
   - That/the substitution on Item 12 (“That restaurant is supposed to have very good food”) is acceptable
   - Good/nice substitutions are acceptable

4. I have clearly specified that a contraction (e.g., I’ve > I have, it’s > it is) and its non-contracted form should be treated as synonymous (Score 3); however, if the use of a contraction makes the meaning of the sentence ambiguous (e.g., ‘The red book’s on the table’ or ‘The red books on the table’?), is should be scored as 2.

5. Lastly, I added the following two notes:
   - If a participant starts repeating a sentence before the beep and do not restart after the beep, score as N/A.
   - Even if a pronunciation of one sound is not native-like, no points will be taken off, unless it severely affects the meaning of the sentence (e.g., light < right).

Typically in the Japanese-L1 participants’ responses, these sounds include: /s/ vs. /ʃ/, /t/ vs. /ʈ/, /h/ vs. /ɻ/, /j/ vs. /ɻ/, /w/ vs. /o/.

The second note is added, given that this elicited imitation test is more meaning oriented than pronunciation, and the judgment of native-likeness in pronunciation would make the scoring more subjective than necessary.
In making these revisions, I consulted with a second coder of the elicited imitation test (i.e., an English-L1 PhD student majoring in Linguistics) to see whether these revisions would make the coding scheme more comprehensible and coding easier to do. The test had 29 items and hence the total score of this test ranged from 0 to 116.

5.3.4.3.3 Inter-rater reliability

Scoring the cloze test was straightforward and required no inferences, because there was only one correct answer and no partial scores were given. Scoring of the elicited imitation test, on the other hand, required subjective judgments, given a range of possible scores (from zero to four). Hence, a second rater, who also served as the second coder of T-unit/clause identification and accuracy, rated 23% of the data (i.e., 15 participants’ responses to the test). Given that participants’ proficiency levels were intentionally various, I made sure to include participants with various proficiency levels in these 15 responses to be double-coded. As was the case with the coding of T-units, clauses, and accuracy, he went through a one-hour second rater training. In this training, we went through the revised scoring guidelines so that he could familiarize himself with the coding protocol and practice coding five responses. Any questions, uncertainties, and discrepancies were resolved at this point. He then rated 15 participants’ responses to the elicited imitation test. Based on these ratings, an inter-rater reliability estimate was calculated to be 78.62. This inter-rater reliability was calculated by dividing the number of items where a discrepancy in scoring was observed by the number of items coded and multiplying it by 100. The inter-rater reliability of 78.62 is not extremely high; however, it must be noted that the majority of these discrepancies, except for five items, were realized as a difference in just one score point on the 5-point scale.
5.3.4.4 Statistical analysis

In making comparisons between tasks and proficiency groups, I focused on descriptive statistics, graphic depictions of patterns, and effect sizes, as well as my interpretations of them. My decision not to rely upon inferential statistics, in the form of significance testing, was based on insightful suggestions made by statisticians (e.g., Tabachnick & Fidell, 2013), psychologists (e.g., Cohen, 1994; Cumming, 2012), and L2 researchers (e.g., Larson-Hall, 2010; Norris & Ortega, 2000, 2006; Norris, 2015b; Plonsky, 2013) about the danger of focusing on null hypothesis statistical testing and p-values as the primary or singular approach to interpreting quantitative data. To summarize their arguments, null hypothesis statistical testing encourages dichotomous thinking and treats phenomena as if they were black or white. In other words, it inappropriately encourages researchers to focus on the p-value and whether it exceeds or does not exceed an arbitrary value (typically 0.05) instead of investigating the actual, descriptive patterns in the data and what they might mean in the given context. Additionally, p-values tell us nothing about the magnitude of the effect or how likely it is to identify the same range and distribution of scores when an experiment is replicated. Another reason for focusing on descriptive statistics, instead of inferential statistics, is that participants in the present study were recruited on a voluntary basis rather than being sampled randomly from each population of Japanese-L1 university students in Japan or of university-level native speakers of English. This approach to data collection violates the assumption of random sampling from a known population for the use of inferential statistics (Tabachnick & Fidell, 2013), and although it is a nearly ubiquitous practice in L2 research, strictly speaking, it is inappropriate to utilize inferential statistics in this situation for anything but highly exploratory data ‘snooping’.
Additionally, this study included two independent variables, one (i.e., degrees of cognitive task complexity) with four levels and the other (i.e., proficiency) with two levels, as well as four dependent variables (i.e., reaction time from the dual-task methodology, time estimation, task difficulty ratings, mental effort ratings) in Study Part 1 (i.e., validation of assumptions about cognitive task complexity). Study Part 2 (i.e., investigation of the relationship between cognitive task complexity and task performance) included two independent variables, one (i.e., degrees of cognitive task complexity) with four levels and the other (i.e., proficiency) with two levels, and as many as eight dependent variables (i.e., task performance measures). Given this large number of variables, in order to control for type I error, the experiment-wise alpha level would need to be adjusted for each pair-wise comparison on each dependent variable, resulting not only in a very stringent criterion for finding statistical significance but also in a considerable reduction in the overall power of the inferential tests. Lastly, by comparing the mean and the upper 95% confidence interval for each condition (i.e., task in this case), we can obtain essentially the same information that significance testing would provide: If the mean exceeds the upper 95% confidence interval, the difference is likely statistically significant. For all of these reasons, and in light of the cutting-edge (i.e., exploratory) nature of the research methodology employed in this study, the main focus here was on understanding the findings through analysis of descriptive statistics and graphs, as well as effect sizes to gauge the magnitude of difference.

Data were first screened to detect any sever outliers (scores more than three standard deviations away from the mean), and when found, their data were deleted list-wise for the particular measures. Subsequently, means (or medians when data were not normally distributed) and standard deviations, as well as 95% confidence intervals, were calculated for all measures on
all levels of each factor (i.e., task and proficiency group). Additionally, in order to make
comparisons between the different measures of cognitive load (with distinct scales) possible and
to investigate which measure may be most affected by or sensitive to the manipulation of
cognitive complexity of tasks, effect sizes (Cohen’s $d$) were calculated for each measure between
Task 1 (the designed-to-be simplest task) and Task 4 (the designed-to-be most complex task).
These two tasks were chosen because of their greatest difference in cognitive complexity by
design. Similarly, in investigating the effects of cognitive task complexity on task performance,
Cohen’s $d$ was calculated for each of the outcomes measures again between Task 1 and Task 4.
Cohen’s $d$ is simply the standardized mean difference between the average scores of the two
groups on the dependent variable. However, in the current study, because differences between
tasks were of primary interest, the measures of cognitive load were repeated for each task.
Hence, mean differences had to be calculated on the same measure between two tasks, following
a repeated-measures design. There has been disagreement regarding the most appropriate
approach to the standardization of effect sizes in repeated-measures designs (e.g., Lipsey and
Wilson, 2001). Recently, Cumming (2012) argued that in calculating Cohen’s $d$ for a repeated-
measures design, “the best choice of standardizer for $d$ is usually $s_{av}$ [averaged standard deviation
of the two conditions]” as it is “the best estimate of the population SD [standard deviation]” (p.
292). Following his recommendation, Cohen’s $d$ was calculated by dividing the mean difference
(of Tasks 1 and 4) by the standard deviation averaged between the two tasks.
CHAPTER VI: RESULTS

6.1 Introduction

This section firstly reports on the quantitative findings from the cognitive load measures for native speakers of English and for English-L2 speakers. With regards to the English-L2 speaker data, the overall trends are presented first; then the results are broken down and reported by proficiency groups. Secondly, results of the English-L2 participants’ explanatory comments for their task difficult and mental effort ratings are explored overall and then by proficiency groups. Lastly, the relationship between measured cognitive task complexity and task performance in terms of the syntactic complexity, accuracy, lexical variety, and fluency (CALF) measures are reported. For this part as well, the overall trends are presented first followed by the proficiency groups analysis.

6.2 Measured cognitive load: Quantitative results

In what follows, I will report on the quantitative results of the cognitive load measures (i.e., dual-task methodology, time estimation, task difficulty, and mental effort self-ratings) for the four main narrative tasks. The data gathered from the native speakers of English will be presented first, followed by English-L2 speakers’ data. As for the latter, the overall trends as well as the patterns by proficiency groups will be reported.

6.2.1 Native speakers of English

First, all individual data for all four measures were inspected for univariate outliers, with a cut-off set at three standard deviations away from the mean. In Study Part 1, one outlier was found in the dual-task methodology analysis, three in time estimation, and one from the task difficulty rating. Hence, all of these cases were eliminated from further analysis. In Study Part 2, no outliers were found in any sectors of the data set. In Part 1, once the outliers were removed,
the data for the reaction time of the dual-task methodology for each task was inspected for normality of the distributions; the data sets were normally distributed with no severe skewness or kurtosis, except for Task 2. Task 2 was shown to be positively skewed, with most of the participants responding to the color change quite quickly. In addition, the two error rates were shown to be positively skewed and leptokurtic for all four tasks, indicating that overall the native speaker participants made very few errors in responding to the color changes. Given the considerable skewness of distributions for the reaction time analysis for Task 2 and the two error rates of the dual-task methodology, the decision was made to use the medians, instead of their means, as the more accurate representation of central tendency for these particular measures (for Task 2 only with the reaction time). The data for the rest of the measures (i.e., time estimation, task difficulty and mental effort self-ratings) were all shown to be distributed normally with no overt skewness or kurtosis. In Part 2, data showed normal distributions on all three measures. Tables 13 and 14 below summarize means (or medians for the reaction time analysis of the dual-task methodology for Task 2 and for error rates) and standard deviations for each measure on the four tasks in Part 1 and Part 2 for the English native speaker participants. Each task is numbered in sequence from cognitively simplest (1) to cognitively most complex (4) according to the original design.
Table 13
Means/Medians (and Standard Deviations) and Effect Sizes ($d$) for Each Cognitive Load Measure on the Four Tasks in Part 1 (for Native Speakers of English, $n = 30$)

<table>
<thead>
<tr>
<th>Task</th>
<th>Dual-Task Methodology</th>
<th>Time Estimation (sec)</th>
<th>Task Difficulty****</th>
<th>Mental Effort****</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reaction Time (sec)**</td>
<td>Error Rate 1 (%)***</td>
<td>Error Rate 2 (%)***</td>
<td></td>
</tr>
<tr>
<td>Task 1</td>
<td>0.85</td>
<td>0.00</td>
<td>0.00</td>
<td>2.83</td>
</tr>
<tr>
<td></td>
<td>(0.25)</td>
<td>(11.41)*</td>
<td>(14.10)*</td>
<td>(26.78)</td>
</tr>
<tr>
<td>Task 2</td>
<td>0.82</td>
<td>0.00</td>
<td>0.00</td>
<td>-5.46</td>
</tr>
<tr>
<td></td>
<td>(0.46)*</td>
<td>(11.11)*</td>
<td>(8.82)*</td>
<td>(40.32)</td>
</tr>
<tr>
<td>Task 3</td>
<td>0.87</td>
<td>0.00</td>
<td>12.50</td>
<td>4.58</td>
</tr>
<tr>
<td></td>
<td>(0.27)</td>
<td>(6.48)*</td>
<td>(12.83)*</td>
<td>(24.70)</td>
</tr>
<tr>
<td>Task 4</td>
<td>0.93</td>
<td>0.00</td>
<td>0.00</td>
<td>-6.65</td>
</tr>
<tr>
<td></td>
<td>(0.27)</td>
<td>(8.86)*</td>
<td>(10.69)*</td>
<td>(38.43)</td>
</tr>
<tr>
<td>$d$</td>
<td>0.31</td>
<td>N/A</td>
<td>N/A</td>
<td>-0.29</td>
</tr>
</tbody>
</table>

Note. The $d$ value was calculated between Task 1 and Task 4. *The median, instead of the mean, is provided. **The reaction times were measured in milliseconds but are reported in seconds here for ease of interpretation. ***Error Rate 1: Not responding to the color change when required; Error Rate 2: Responding when not required. ****$k = 9$. 

135
Table 14
Means (and Standard Deviations) and Effect Sizes (d) for Each Cognitive Load Measure on the Four Tasks in Part 2 (for Native Speakers of English, n = 31)

<table>
<thead>
<tr>
<th>Time Estimation (sec)</th>
<th>Task Difficulty*</th>
<th>Mental Effort*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>24.96 (39.09)</td>
<td>3.35 (1.80)</td>
</tr>
<tr>
<td>Task 2</td>
<td>12.78 (38.24)</td>
<td>4.61 (1.93)</td>
</tr>
<tr>
<td>Task 3</td>
<td>26.44 (39.85)</td>
<td>3.77 (1.73)</td>
</tr>
<tr>
<td>Task 4</td>
<td>26.60 (44.70)</td>
<td>4.77 (1.63)</td>
</tr>
<tr>
<td>d</td>
<td>0.04</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Note. The $d$ value was calculated between Task 1 and Task 4. *$k = 9$.

6.2.1.1 Dual-task methodology

The native speaker participants reacted to the color change (i.e., the secondary task) quite quickly, on average at the rate of under one second on all four tasks. As can be seen in Figure 6 and Table 13 above, their average reaction times were similar across tasks, with Task 2 (the designed to be second simplest task) eliciting slightly faster reaction times and Task 4 (the design to be most complex task) eliciting slightly slower reaction times than the other three tasks. However, the difference between Task 2 and Task 4 was a mere 0.11 seconds. Note that the 95% confidence intervals for the four tasks are all overlapping with each other. These results suggest that, for this group of native speaker participants, the four tasks posed relatively similar degrees of cognitive demands according to this measure. Cronbach’s alpha reliability estimate for this measure of dual-task methodology (reaction time) was found to be $\alpha = .64$, suggesting that it was relatively consistent in distinguishing among participants, and quite possibly the alpha was not higher due to the lack of variability among the participants. In order to investigate the magnitude of difference between Task 1 and Task 4, the two tasks designed and shown to be maximally different in terms of their cognitive complexity overall (taking into account both native and non-native speaker data, as reported below). The Cohen’s $d$ for dual-task methodology (reaction time) was found to be $d = .31$ between Task 1 and Task 4, indicating a small to medium effect.
size (Cohen, 1988). In terms of error rates, most of the participants made no errors at all on any of the four tasks, except for the Type 2 Error Rate on Task 3. Yet, the average error rate, to this end, was only 12.5%. These results indicate that participants’ responses to the color change were extremely accurate in general. This finding further supports the validity of the reaction time data.

Figure 6. Reaction time of the dual-task methodology for the four tasks with 95\% confidence intervals in Part 1 (for native speakers of English).

6.2.1.2 Time estimation

Turning our attention to the next cognitive load measure, time estimation, the picture appears to be quite different from the pattern that we saw in the dual-task methodology. In Study Part 1, the native speaker participants estimated their time on task to be on average three to five seconds longer than they had actually spent when engaging in Task 1 (the designed-to-be simplest task) and Task 3 (the designed-to-be second most complex task), whereas their average estimated times were five to seven seconds shorter than their actual time on task for Task 2 (the
designed-to-be second simplest task) and Task 4 (the designed-to-be most complex task). As can be seen in Figure 7, the 95% confidence intervals for the four tasks are once again all overlapping. These results suggest that Tasks 1 and 3 posed somewhat lower cognitive demands than Tasks 2 and 4, but only to a certain (non-statistical) extent. Similarly in Study Part 2, Task 2 elicited the shortest, positive difference on average between the actual time on task and the estimated time. The other three tasks elicited longer, positive differences than Task 2, with the 95% confidence intervals for all four tasks overlapping with each other (see Figure 8). These results suggest that for the Part 2 participants, Task 2, but not Task 4, posed somewhat higher cognitive demands than did the other three tasks, although not to a meaningful extent.

Cronbach’s alpha reliability estimates for this measure of time estimation were found to be $\alpha = .87$ in Part 1 and $\alpha = .93$ in Part 2, suggesting that it was highly consistent in distinguishing among participants. Finally, the Cohen’s $d$ was found to be $d = -.29$ in Part 1 and $d = .04$ in Part 2 between Task 1 and Task 4, indicating a small effect size in Part 1 and a very small effect size in Part 2 (Cohen, 1988).
Figure 7. Time estimation for the four tasks with 95% confidence intervals in Part 1 (for native speakers of English).

Figure 8. Time estimation for the four tasks with 95% confidence intervals in Part 2 (for native speakers of English).
6.2.1.3 Self-ratings of perceived difficulty and mental effort

Lastly, the measures of task difficulty and mental effort elicited similar trends in both parts of the study. Overall, the tasks were perceived to be on the easy and less effortful half of the scale (i.e., lower than 4.5) for the native speaker participants; they did seem to distinguish among the four tasks nonetheless. Note, however, that on average Task 4 crossed just into the higher half of the scale (i.e., above 4.5) for both difficulty and effort ratings and Task 2 did so for effort, but not for difficulty. Notice also that the mental effort ratings for all four tasks were always about .30 points higher than the task difficulty ratings. Looking into patterns for each of the four tasks, Task 1, the designed-to-be simplest task, was perceived to be least difficult and least mental effort consuming on average in both Study Part 1 and Part 2. Task 4, the designed-to-be most complex task, was rated to be most difficult and most mental effort consuming of all the four tasks in Part 1, although in Part 2, Task 2 was rated to be slightly more effort consuming (.10 difference than Task 4 on a 9-point scale). Generally, then, Task 2 and Task 3 fell between the other two tasks, though with Task 2 being perceived as more difficult and to have required more mental effort than Task 3. As can be seen in Figures 9-12, the mean values (indicated by a circle) of Tasks 2 and 4 exceed the upper 95% confidence intervals for Tasks 1 and 3 on both task difficulty and mental effort ratings in both study parts. All these findings indicate that, in terms of the self-rating measures, Tasks 2 and 4 posed meaningfully higher cognitive demands than did Tasks 1 and 3. Cronbach’s alpha reliability estimates for task difficulty and mental effort ratings were found to be $\alpha = .70$ and $\alpha = .81$ in Part 1 and $\alpha = .61$ and $\alpha = .76$ in Part 2, suggesting that they were generally consistent in distinguishing among participants. Finally, the Cohen’s $d$ was found to be $d = .69$ and $d = .83$ for task difficulty and $d = .68$ and $d = .89$ for
mental effort in Part 1 and Part 2, respectively. Note the much larger effect sizes (between Task 1 and Task 4) for these self-assessment measures than for the other two measures.

Figure 9. Task difficulty ratings for the four tasks with 95% confidence intervals in Part 1 (for native speakers of English).
Figure 10. Mental effort ratings for the four tasks with 95% confidence intervals in Part 1 (for native speakers of English).

Figure 11. Task difficulty ratings for the four tasks with 95% confidence intervals in Part 1 (for native speakers of English).
6.2.1.4 Summary of complexity measures between tasks for native speakers

To summarize the native speaker data, the dual-task methodology and time estimation indicated no meaningful difference in the degrees of cognitive load among the four tasks, whereas Task 1 and Task 3 were shown by the self-rating measures to be meaningfully less difficult and less effort consuming than Task 2 and Task 4. Note that these findings do not entirely line up with the predicted order of cognitive complexity, with Task 1 designed to be simplest among the four tasks, followed by Task 2, Task 3, and Task 4 sequentially. That is, Task 2 seems to have behaved and been perceived similarly to Task 4, as the most difficult and effortful, while the other tasks fell in the order predicted. Comparing Task 1 and Task 4, the two tasks that were designed to be maximally different in the number of elements (and hence in cognitive complexity), task design manipulation had overall small effects on the measures of
dual-task methodology and time estimation (objective measures) and large effects on the two self-assessment measures. Hence, although to a small extent, Task 1 and Task 4 were found to pose somewhat distinct degrees of cognitive load to the native speaker participants.

6.2.2 English L2 speakers

Turning to English L2 speakers’ data, findings from the four measures of cognitive load (i.e., the dual-task methodology, time estimation, task difficulty self-ratings, and mental effort self-ratings) in relation to the four tasks as performed by English L2 speakers are reported here. The overall trends are explored first, and then the analysis is presented by proficiency groups.

6.2.2.1 Overall trends

Once again, all data were first inspected for outliers, with a cut-off set at three standard deviations from the mean. As a result, one participant’s data were excluded from the analysis of the dual-task methodology, one from time estimation, and one from task difficulty in Study Part 1. In Study Part 2, two participants’ data were excluded from time estimation. In both parts of the study, once outliers were removed, distributions for all measures, except for the two error rates (only used in Part 1), were found to be normal across tasks, with no severe skewness or kurtosis. For the error rates of the dual-task methodology (used only in Part 1), distributions were positively skewed and leptokurtic, reflecting the generally quite low average error rates for most participants. Given the skewness of distributions for the two error rates of the dual-task methodology, the decision was made to use the medians, instead of their means, for these particular measures when investigating their central tendencies. Tables 15 and 16 below summarize means (or medians for the error rates) and standard deviations for each measure on the four tasks in Part 1 and Part 2.
Table 15
Means/Medians (and Standard Deviations) for Each Cognitive Load Measure on the Four Tasks and Effect Sizes (d) in Part 1 (for English-L2 Speakers, n = 53)

<table>
<thead>
<tr>
<th>Task</th>
<th>Dual-Task Methodology</th>
<th>Time Estimation (sec)</th>
<th>Task Difficulty****</th>
<th>Mental Effort****</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reaction Time (sec)**</td>
<td>Error Rate 1 (%)***</td>
<td>Error Rate 2 (%)***</td>
<td></td>
</tr>
<tr>
<td>Task 1</td>
<td>1.18 (0.39)</td>
<td>8.55 (19.78)*</td>
<td>0.00</td>
<td>1.36 (37.24)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(18.03)*</td>
<td></td>
<td>5.67 (1.98)</td>
</tr>
<tr>
<td>Task 2</td>
<td>1.27 (0.51)</td>
<td>15.48 (20.82)*</td>
<td>0.00</td>
<td>-12.36 (44.83)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(11.97)*</td>
<td></td>
<td>6.58 (1.85)</td>
</tr>
<tr>
<td>Task 3</td>
<td>1.22 (0.50)</td>
<td>16.03 (17.71)*</td>
<td>5.88</td>
<td>-8.57 (48.01)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(12.73)*</td>
<td></td>
<td>6.25 (1.70)</td>
</tr>
<tr>
<td>Task 4</td>
<td>1.33 (0.48)</td>
<td>14.29 (20.80)*</td>
<td>10.00</td>
<td>-14.20 (58.18)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(17.75)*</td>
<td></td>
<td>7.33 (1.35)</td>
</tr>
<tr>
<td>d</td>
<td>0.34</td>
<td>NA</td>
<td>NA</td>
<td>-0.32 (1.00)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.98 (1.00)</td>
</tr>
</tbody>
</table>

**Note.** The d value was calculated between Task 1 and Task 4. *The median is provided. **The reaction times were measured in milliseconds but are reported in seconds here for ease of interpretation. ***Error Rate 1: Not responding to the color change when required; Error Rate 2: Responding when not required. ****k = 9.
Table 16
Means (and Standard Deviations) for Each Cognitive Load Measure on the Four Tasks and Effect Sizes (d) in Part 2 (for English-L2 Speakers, n = 67)

<table>
<thead>
<tr>
<th></th>
<th>Time Estimation (sec)</th>
<th>Task Difficulty*</th>
<th>Mental Effort*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>−15.32 (34.91)</td>
<td>5.13 (2.03)</td>
<td>5.40 (1.99)</td>
</tr>
<tr>
<td>Task 2</td>
<td>−20.70 (43.99)</td>
<td>6.73 (1.71)</td>
<td>6.58 (1.69)</td>
</tr>
<tr>
<td>Task 3</td>
<td>−23.06 (41.20)</td>
<td>6.27 (1.43)</td>
<td>6.30 (1.54)</td>
</tr>
<tr>
<td>Task 4</td>
<td>−33.49 (55.63)</td>
<td>6.94 (1.76)</td>
<td>6.93 (1.39)</td>
</tr>
<tr>
<td>d</td>
<td>−0.39</td>
<td>0.95</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Note. The d value was calculated between Task 1 and Task 4. *k = 9.

6.2.2.1.1 Dual-task methodology

It can be observed that on average, participants responded to the color change faster when engaging in Task 1 (the designed-to-be simplest task) than on the other three tasks. Additionally, participants were slowest at responding to the color change correctly when engaging in Task 4, the designed-to-be most complex task. These results suggest that for this group of learners, Task 1 posed the least amount of cognitive demand, whereas Task 4 was cognitively most complex and demanding. As can be seen in Figure 13, the mean value of Task 4 exceeds the upper 95% confidence interval for Task 1, which indicates a meaningful difference between the two tasks. While Task 2 and Task 3 (the two medium difficulty tasks by design) fell between Task 1 and Task 4, Task 2 elicited slightly slower reaction times than did Task 3. Cronbach's alpha reliability estimate for this measure of dual-task methodology (reaction time) was found to be α = .69, suggesting that it was relatively consistent in distinguishing among participants. In order to investigate the magnitude of difference, an effect size was calculated for the difference between Task 1 and Task 4 once again. The Cohen’s d for dual-task methodology (reaction time) was found to be d = .34 between Task 1 and Task 4, indicating a small to medium effect size (Cohen, 1988) in favor of the complex task. In terms of error rates, only small differences were found among the four tasks for both error rates, on average never higher than 17% in responding
to the color change. Hence, participants’ responses to the color change were relatively accurate in general, which in turn supports the validity of the reaction time data.

Figure 13. Reaction time of the dual-task methodology for the four tasks with 95% confidence intervals in Part 1 (for English L2 speakers).

6.2.2.1.2 Time estimation

In Study Part 1, on average, all tasks except Task 1, were judged to have taken shorter than the actual time spent (see Table 15 and Figure 14). Task 4 elicited the longest negative difference (−14.20), while Task 1 elicited a very small positive difference (1.36) between the estimated time and the actual time on task. Tasks 2 and 3 (the middle two tasks by design) fell between the other two tasks. Note that the 95% confidence intervals are all overlapping with each other. Cronbach’s alpha reliability estimate for the measure of time estimation was found to be $\alpha = .85$. Cohen’s $d$ for this measure of time estimation was found to be $d = -.32$ between Task 1 and Task 4, which indicates a small to medium effect size (Cohen, 1988) in favor of the complex task. In
Study Part 2, we can observe a similar trend to Part 1. All tasks were estimated to have taken shorter than their actual time on task on average (see Table 16 and Figure 15). Once again, Task 4 elicited the biggest difference (−33.49) and Task 1 elicited the smallest difference (−15.32), but this time the mean value of Task 1 exceeds the upper 95% confidence interval for Task 4 – hence the difference between the two tasks was shown to be meaningful. Much like Part 1, the middle two tasks, Task 2 and Task 3, fell between the other tasks; however, the order of these two tasks was in the expected direction. In other words, Task 2 elicited a smaller difference between the estimated time and the actual time on task than did Task 3. Note that Cronbach's alpha reliability estimate for the measure of time estimation was found to be \( \alpha = .88 \). Cohen’s \( d \) for this measure of time estimation was found to be \( d = -.39 \) between Task 1 and Task 4, which indicates a small to medium effect size (Cohen, 1988). These findings are in line with the prediction made on the prospective, rather than retrospective, time estimation approach in the field of cognitive psychology: The more complex a task at hand becomes, the shorter the time passed is perceived to be.
Figure 14. Time estimation for the four tasks with 95% confidence intervals in Part 1 (for English L2 speakers).

Figure 15. Time estimation for the four tasks with 95% confidence intervals in Part 2 (for English L2 speakers).
6.2.2.1.3 Self-ratings of perceived difficulty and mental effort

Overall, the English-L2 informants found the four tasks to be quite difficult and effortful. Their ratings fell in the upper half of the scale (i.e., above 4.5) on average for all four tasks. This pattern clearly contrasts with the pattern found for the native speaker informants who rated the same four tasks to be much easier and much less effort consuming. In addition, for the L2 groups, there is almost no difference observed between difficulty and effort, on average, which is also in contrast with the native speaker groups who constantly rated mental effort to be .30 points higher than difficulty. Comparing the four tasks, then, in both Part 1 and Part 2, on average, Task 1 was rated as least difficult and least mental effort consuming, while Task 4 was rated most difficult and judged to require the greatest amount of mental effort (see Tables 15 & 16). These ratings suggest lower cognitive demands required to complete Task 1 and higher cognitive demands required to complete Task 4. As can be seen in Figures 16 through 19, the mean values of Task 4 exceed the upper 95% confidence intervals for Task 1, which indicates a meaningful difference between the two tasks on both self-rating measures. Once again, Tasks 2 and 3 fall between Tasks 1 and 4. Comparing these two tasks, Task 2 received slightly higher difficulty and mental effort ratings. Cronbach's alpha reliability estimates for the measures of task difficulty and mental effort self-assessment were found to be $\alpha = .78$ and $\alpha = .81$, respectively in Part 1 and $\alpha = .69$ and $\alpha = .68$, respectively in Part 2. Cohen’s $d$ for these measures was $d = .98$ and $d = .95$ (for task difficulty ratings in Part 1 and Part 2, respectively) and $d = 1.00$ and $d = .89$ (for mental effort ratings in Part 1 and Part 2) between Task 1 and Task 4, all indicating quite large effect sizes (Cohen, 1988) in favor of the complex task. Note that the magnitude of $d$ is similar for the L2 and native speaker groups, indicating that both groups distinguished to a similar extent between Task 1 and Task 4 on these self-assessment measures.
Figure 16. Task difficulty ratings for the four tasks with 95% confidence intervals in Part 1 (English L2 speakers).

Figure 17. Mental effort ratings for the four tasks with 95% confidence intervals in Part 1 (English L2 speakers).
Figure 18. Task difficulty ratings for the four tasks with 95% confidence intervals in Part 1 (English L2 speakers).

Figure 19. Mental effort ratings for the four tasks with 95% confidence intervals in Part 2 (English L2 speakers).
6.2.2.1.4 Summary of complexity measures between tasks for English L2 speakers

In general, the tasks were shown to be cognitively more demanding by all four measures for the English L2 speaker group than for the native speaker group. Comparing the four tasks within the English L2 speaker group, on the basis of four distinct and independent measures of cognitive complexity, Task 1, the designed-to-be simplest task, was consistently shown to be the simplest among the four tasks in terms of cognitive load. Similarly, Task 4, presumed to be most complex by task design, was consistently shown to be cognitively most complex. Task 2 and Task 3 – the two medium-difficulty tasks by design – also fell consistently between Task 1 and Task 4 in terms of their measured cognitive demands; however, differences in the level of cognitive complexity found between these two middle tasks tended to be small most of the time, and contrary to the intended task design, all measures, except for time estimation in Part 2, showed Task 2 to be slightly more cognitively demanding than Task 3. Comparing Task 1 and Task 4, task complexity manipulation had small to medium effects as reflected by the two objective measures of cognitive load, the dual-task methodology and time estimation, while it had much larger effects in terms of the two self-assessment measures. That is, although to varying degrees, all these measures indicated that Task 1 was indeed measured to be the simplest task and Task 4 to be the most complex task in cognitive terms.

6.2.2.2 Role of proficiency

The results were further explored to investigate the possible effects of participants’ L2 proficiency on cognitive load measures. In order to separate the participants into proficiency groups, their cloze test scores were analyzed and reported first. Then, the section moves onto summarizing the results of each of the cognitive load measures for distinct proficiency groups.
6.2.2.2.1 Proficiency tests

On average, participants in Study Part 1 scored 14.06 (SD = 8.14) on the cloze test (k = 50). The minimum observed score was zero and the maximum was 30. The scores were normally distributed with no substantial skewness or kurtosis. No severe outliers were detected. After inspecting the score distribution, participants were divided into three groups: low, mid, and high proficiency groups, according to their test scores. The lowest-scoring 20 participants (ranging from zero to 10) were classified into a low proficiency group, and the highest-scoring 19 participants (ranging from 19 to 30) were categorized as a high proficiency group. The twentieth potentially “high proficiency” participant scored 18, and four other participants scored the exact same score. Inclusion of the twentieth participant, then, would have implied inclusion of the other four participants in the high proficiency group. A decision was made not to include the participants who scored 18 to keep the number of participants in the low and high proficiency groups as similar as possible. The rest of the participants, then, were classified as mid-proficiency.

In Study Part 2, participants’ average score on the cloze test was 14.36 (SD = 5.64), a very similar average score as Part 1 participants’. The minimum score observed was four and the maximum was 30. As was the case in Part 1, the scores were normally distributed with no substantial skewness or kurtosis, and no severe outliers were detected. In this part of the study, participants were divided into the low, mid, and high proficiency groups based on the cut points set in Part 1 to allow comparisons between the two parts of the study. Hence, 16 participants who scored from four to 10 were assigned to the low proficiency group, and 13 participants who scored from 20 to 30 (there was no one who scored 19) were grouped into the high proficiency group. The rest of the 37 participants were then assigned to the mid proficiency group.
Recall that Part 2 participants not only took the cloze test, but also the oral elicited imitation test. On the elicited imitation test, participants scored on average 56.17 (SD = 20.83) with \( k = 116 \). The lowest observed score was 24 and the highest was 102. As was the case with the cloze test, the scores on the elicited imitation test were normally distributed with no substantial skewness or kurtosis, and no severe outliers were detected. When the elicited imitation scores were analyzed by the proficiency groups (assigned based on the cloze test scores), no severe outliers were detected, and the scores on the elicited imitation increased as participants’ proficiency went up (see Table 17). Additionally, the cloze test scores and the elicited imitation test scores were correlated at \( r = 0.64 \) (\( p < .05 \)). These findings provide some criterion-related validity support for using the cloze test as the test of global proficiency and perhaps counter the concern that the cloze test may not be an appropriate measure of proficiency (i.e., given that its modality is incongruent with the main oral narrative tasks, in this study context).

Table 17
The Mean Score, Standard Deviation, and Range on Elicited Imitation Test for Each Proficiency Group

<table>
<thead>
<tr>
<th></th>
<th>Low proficiency</th>
<th>Mid Proficiency</th>
<th>High proficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (SD)</td>
<td>41.69 (14.01)</td>
<td>54.11 (17.47)</td>
<td>79.85 (16.99)</td>
</tr>
<tr>
<td>Range</td>
<td>26-69</td>
<td>24-89</td>
<td>49-102</td>
</tr>
</tbody>
</table>

Finally, in order to maximize the proficiency difference, the mid-proficiency group was excluded from comparative analyses, and only the low and high proficiency groups were investigated in the subsequent sections.  

---

7 An independent samples t-test revealed that the mean scores on the cloze test and the elicited imitation test of the high proficiency group were statistically significantly higher than the mean scores of the low proficiency group (Part 1, cloze test: \( t (35) = -19.21 \), \( p = 0.00 \); Part 2, cloze test: \( t (17.88) = -14.83 \), \( p = 0.00 \); Part 2, elicited imitation test: \( t (27) = -6.63 \), \( p = 0.00 \)).
6.2.2.2 Cognitive load measures by proficiency

To explore the possible relationships between cognitive task complexity and proficiency differences, the results of the four independent measures of cognitive task complexity were each analyzed by comparing the means between the proficiency groups overall and on each task. To maximize the possibility of difference in proficiency, the analysis here will focus mostly on the comparison between the low and high proficiency groups. However, the descriptive statistics (as well as effect sizes) for the mid proficiency participants will be reported to give a fuller picture of patterns found in the data. Given small sets of data for each proficiency group, outliers will not be eliminated from the analysis at this time, but if the data are found to be skewed, the median is used, instead of the mean, as a more trustworthy indicator of central tendency. Tables 18 through 20 below summarize means and standard deviations for each measure on the four tasks for each proficiency group in Part 1 and Part 2.
Table 18
Means/Medians (and Standard Deviations) and Effect Sizes (d) for Dual-Task Methodology and Time Estimation on the Four Tasks by Proficiency in Part 1

<table>
<thead>
<tr>
<th>Task</th>
<th>Dual-Task Methodology</th>
<th>Time Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reaction Time (sec)</td>
<td></td>
</tr>
<tr>
<td>Task 1</td>
<td>1.13</td>
<td>1.36</td>
</tr>
<tr>
<td></td>
<td>(0.26)</td>
<td>(0.41)</td>
</tr>
<tr>
<td>Task 2</td>
<td>1.22</td>
<td>1.26</td>
</tr>
<tr>
<td></td>
<td>(0.49)</td>
<td>(0.46)</td>
</tr>
<tr>
<td>Task 3</td>
<td>1.27</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td>(0.51)</td>
<td>(0.43)</td>
</tr>
<tr>
<td>Task 4</td>
<td>1.10</td>
<td>1.46</td>
</tr>
<tr>
<td></td>
<td>(0.47)*</td>
<td>(0.44)</td>
</tr>
<tr>
<td>d</td>
<td>-0.08</td>
<td>0.24</td>
</tr>
</tbody>
</table>

Note. L = Low Proficiency Group (n-size); H = High Proficiency Group (n-size). The d values are calculated between Task 1 and Task 4. * The median is provided.

Table 19
Means/Medians (and Standard Deviations) and Effect Sizes (d) for Task Difficulty and Mental Effort on the Four Tasks by Proficiency in Part 1

<table>
<thead>
<tr>
<th>Task</th>
<th>Task Difficulty</th>
<th>Mental Effort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L (20)</td>
<td>M (14)</td>
</tr>
<tr>
<td>Task 1</td>
<td>6.40</td>
<td>5.57</td>
</tr>
<tr>
<td></td>
<td>(1.88)</td>
<td>(1.56)</td>
</tr>
<tr>
<td>Task 2</td>
<td>6.65</td>
<td>6.93</td>
</tr>
<tr>
<td></td>
<td>(2.18)</td>
<td>(1.59)</td>
</tr>
<tr>
<td>Task 3</td>
<td>7.15</td>
<td>6.43</td>
</tr>
<tr>
<td></td>
<td>(1.50)</td>
<td>(1.28)</td>
</tr>
<tr>
<td>Task 4</td>
<td>8.00</td>
<td>7.07</td>
</tr>
<tr>
<td></td>
<td>(1.23)*</td>
<td>(0.73)</td>
</tr>
<tr>
<td>d</td>
<td>1.01</td>
<td>1.23</td>
</tr>
</tbody>
</table>

Note. L = Low Proficiency Group (n-size); H = High Proficiency Group (n-size). The d values are calculated between Task 1 and Task 4. * The median is provided.
Table 20

Means (and Standard Deviations) and Effect Sizes (d) for Each Cognitive Load Measure on the Four Tasks by Proficiency in Part 2

<table>
<thead>
<tr>
<th></th>
<th>Time Estimation (sec)</th>
<th>Task Difficulty</th>
<th>Mental Effort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>−23.18</td>
<td>−18.03</td>
<td>1.25</td>
</tr>
<tr>
<td></td>
<td>(40.46)</td>
<td>(34.13)</td>
<td>(26.34)</td>
</tr>
<tr>
<td>Task 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>−18.24</td>
<td>−23.37</td>
<td>−16.12</td>
</tr>
<tr>
<td></td>
<td>(49.51)</td>
<td>(46.64)</td>
<td>(30.09)</td>
</tr>
<tr>
<td>Task 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>−15.10</td>
<td>−29.19</td>
<td>−15.28</td>
</tr>
<tr>
<td></td>
<td>(49.13)</td>
<td>(40.72)</td>
<td>(31.66)</td>
</tr>
<tr>
<td>Task 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(58.94)</td>
<td>(61.89)</td>
<td>(25.15)</td>
</tr>
<tr>
<td>d</td>
<td>−0.33</td>
<td>−0.38</td>
<td>−0.70</td>
</tr>
</tbody>
</table>

*Note. L = Low Proficiency Group (n-size); M = Mid Proficiency Group (n-size); H = High Proficiency Group (n-size). The d values are calculated between Task 1 and Task 4. *The median is provided.
6.2.2.2.1 Reaction time in dual-task methodology by proficiency

First, inspection of the data distribution for each proficiency group for this measure of reaction time in the dual-task methodology revealed that the data for Task 3 for the high proficiency group and the data for Task 4 for the low proficiency group were positively skewed and leptokurtic. Otherwise data were normally distributed. Overall, high proficiency participants were as slow as or sometimes slower at detecting the color change than low proficiency participants (see Table 18 and Figure 20), a possibly surprising result. Looking at participants’ reaction times by tasks, surprisingly, low proficiency participants’ reaction times for accurate responses to the color change during Tasks 1 and 4 were quickest among the four tasks. High proficiency participants, on the other hand, reacted to the color change most quickly while engaging in Task 1 and slowest on Task 4, around 277 milliseconds slower than on Task 1. In fact, participants in the low proficiency group reacted to the color change around 266 milliseconds faster on Task 4 than did the high proficiency group. For these low proficiency learners, the task that elicited the slowest responses to the color change was Task 3. It is also interesting to note that high proficiency participants’ reaction times during their engagement in Task 2, which was designed to be cognitively less complex than Task 3, were almost identical to Task 3, while this trend was not observed with low proficiency participants. Cohen’s $d$ for the measure of dual-task methodology (reaction time) was calculated separately for each proficiency level between Task 1 and Task 4, and found to be $d = -0.08$ (very small, negative effect size) for the low proficiency group and $d = 0.58$ (medium effect size) for the high proficiency group (Cohen, 1988).
Figure 20. Reaction time of the dual-task methodology for the four tasks with 95% confidence intervals by proficiency in Part 1.

6.2.2.2 Time estimation by proficiency

The data for time estimation were all found to be normally distributed, except that the data for Task 2 (in Part 1) and Task 3 (in Part 2) for the mid proficiency group were leptokurtic, with Task 2 also being negatively skewed. The measure of time estimation revealed very similar patterns as did the dual-task methodology. As shown in Table 19 and Figure 21, in Study Part 1 on average, low proficiency participants perceived all tasks to have taken shorter than they actually required. They also perceived the time on task to be increasingly shorter than the actual time on task up until Task 3, as tasks became cognitively more complex by design (ranging from −8.67 seconds for Task 1 to −30.06 seconds for Task 3). High proficiency participants, by
contrast, perceived all tasks to have taken longer than these tasks actually took to complete, except for Task 4. Task 4 was perceived to have taken 10.69 seconds shorter than the actual time on task, on average. Cohen’s $d$ for the measure of time estimation was calculated separately for each proficiency level between Task 1 and Task 4. It was found to be $d = -.27$ for the low proficiency group and $d = -.37$ for the high proficiency group, both indicating small effect sizes (Cohen, 1988).

![Figure 21. Time estimation for the four tasks with 95% confidence intervals by proficiency in Part 1.](image_url)

The picture for the time estimation measure in Study Part 2 was somewhat different. Both the low and high proficiency learners perceived the time to have taken shorter, on average, than their actual time on task for all the four tasks, except on Task 1 for the high proficiency group
(see Table 20 and Figure 22). For the low proficiency learners, Task 1 yielded the second shortest negative difference between the actual time of task and their estimated time (−23.18), and it was Task 3 that elicited the shortest negative difference between the low proficiency participants’ estimated and actual times on task (−15.10), followed closely by Task 2 (−18.24). Task 4, then, elicited the longest negative difference (−39.80). The high proficiency group, on the other hand, estimated their time on task to be longer than their actual time on task for Task 1 on average (1.25). The differences between their estimation and the actual time they spent on doing the other three tasks were similar in average terms, raging from −15.28 (Task 3) to −16.71 (Task 4). Cohen’s $d$ for the measure of time estimation was calculated separately for each proficiency level between Task 1 and Task 4. It was found to be $d = −.33$ for the low proficiency group (indicating a small to medium effect size) and $d = −.70$ for the high proficiency group (indicating a medium to high effect size, Cohen, 1988).
For the measure of task difficulty, the data for Task 4 for the low proficiency group in Part 1 and the data for Task 2 for the mid proficiency group in Part 2 were found to be negatively skewed. Task 4 in Part 1 for the low proficiency group was also found to be negatively skewed and leptokurtic for the measure of mental effort. Other data were normally distributed. In Study Part 1, overall, high proficiency learners perceived each task to be easier and less mental effort consuming than did low proficiency learners. Looking at their perceived difficulty and mental effort by tasks, on average, Task 1 was rated as least difficult and requiring least mental effort by both proficiency groups (see Table 19 and Figures 23-24) although to varying degrees. Similarly,
Task 4 was rated as being most difficult and requiring greatest mental effort by both low and high proficiency groups. As all the other measures revealed, Task 2 was cognitively more demanding than Task 3 for high proficiency participants. In other words, their ratings for Task 2 were higher than Task 3 (by 1.11 points for task difficulty and 1.05 points for mental effort). This pattern, however, was only the case with high proficiency participants, and not with low proficiency participants. Consequently, the differences between the low and high proficiency groups on average task difficult and mental effort are minimal for Task 2 (different by 0.43 and by 0.19, respectively). Cohen’s $d$ for the measures of task difficulty and mental effort was calculated separately for each proficiency level between Task 1 and Task 4 in Part 1, and found to be $d = 1.01$ (task difficulty) and $d = 1.21$ (mental effort) for the low proficiency group and $d = 0.97$ (task difficulty) and $d = 1.23$ (mental effort) for the high proficiency group, all indicating very large effect sizes (Cohen, 1988).
Figure 23. Task difficulty for the four tasks with 95% confidence intervals by proficiency in Part 2.
In Study Part 2, all tasks, but Task 2, were perceived to be less difficult and less effort consuming by the high proficiency group than the low proficiency group on average (see Table 20 and Figures 25-26). Task 2 was rated as more difficult and more effort consuming by the high proficiency learners. Comparing their ratings for each task, similar to the findings of Part 1, Task 1 was perceived to be least difficult and mental effort consuming by both high and low proficiency learners. Task 2, then, was rated as second easiest/least effort consuming by the low proficiency learners, whereas it was perceived to be the most difficult and to require most amount of mental effort of all the four tasks by the high proficiency participants. Among the 13 participants assigned to the high proficiency group, nine gave a higher score(s) on either task difficulty or mental effort, or both of them for Task 2 than for Task 4. Curiously, the majority of
their explanatory comments for the high ratings on Task 2 were related to their concern and efforts for a better task performance (see further results along these lines below). Here are some illustrative comments:

#70 ストーリーをつくるためにどうすればおもしろく、うまく説明できるか、不安に思いながら話していた [I was telling a story anxiously thinking of how I can make the story interesting and describe it well.]

#19 順序を正しく説明しようとしたから [I was trying to get the order of the pictures right.]

#42 違いが少ない6枚のイラストについて、どうやって違いをつけて説明するかを考える必要があったから [I had to consider how to describe each of the six pictures with little difference differentially.]

#53 女の子と男性は親子関係にあるのかがはっきりしなかった [It was not clear whether the girl and the man were a daughter and a father.]

For the high proficiency participants in Study Part 2, their attempts to fulfill the task performance assessment criteria (i.e., completeness, effectiveness, creativity) and produce what they thought would be a good narrative performance seemed to have made Task 2 somewhat more cognitively demanding than Task 4. Cohen’s $d$ for the measures of task difficulty and mental effort in Part 2 was calculated separately for each proficiency level between Task 1 and Task 4 in Part 1, and found to be $d = .83$ (task difficulty) and $d = 1.05$ (mental effort) for the low proficiency group and $d = 1.38$ (task difficulty) and $d = 1.05$ (mental effort) for the high proficiency group, all indicating large effect sizes (Cohen, 1988).
Figure 25. Task difficulty for the four tasks with 95% confidence intervals by proficiency in Part 2.
Figure 26. Mental effort for the four tasks with 95% confidence intervals by proficiency in Part 2.

6.2.2.2 Summary of proficiency differences

In general, the four tasks were observed to be cognitively less demanding for high proficiency participants than for low proficiency participants. Comparing the four tasks, although there were some differences in the measured level of cognitive complexity for the four tasks among cognitive load measures and between Study Part 1 and Study Part 2, what is clear is that Task 1 and Task 4 were unanimously shown to be distinct from one another by all measures of cognitive load (except for the dual task methodology used with the low proficiency group – see Chapter 7 for possible explanations) for both high and low proficiency learners in both parts of the study. For the high proficiency participants, the picture seems to be less ambiguous than that
for the low proficiency participants. That is, all measures, except for time estimation, showed the order of complexity to be Task 1, followed by Task 3, with Tasks 2 and 4 being the most complex (but they change places depending on the measures), and the effect sizes between T1 and T4 were noticeably larger for the high proficiency participants on all but two instances (i.e., task difficulty in Study Part 1, mental effort in Study Part 2). The results for low proficiency participants appear to be a bit more complicated. The two self-rating measures showed the order of complexity to be as expected (based on the task design feature) in both study parts, whereas the dual-task methodology and time estimation (in Part 1) indicated that Task 3 was cognitively most complex for this group of participants. On the whole, it seems certain that there was an interaction between participants’ English language proficiency and the measurement of cognitive complexity for the four tasks.

6.3 Measured cognitive load for English L2 speakers: Open-ended responses

This section turns to a report of findings from participants’ own explanations for why the task at hand was easy or difficult to perform and why it required little or much mental effort. In response to the open-ended questions following each task performance, participants provided explanatory comments for their task difficulty and mental effort ratings. All participants provided comments in their first language, Japanese. They typically provided one or two sentence responses, although some participants gave more detailed explanations for their ratings. The length of the comments appeared to be more related to individual tendencies of the participants themselves, rather than any association with the tasks. All participants, without exception, answered the open-ended questions and provided some sort of explanation for why they perceived the task to be easy/difficult and require high/low mental effort. In what follows,
L2 participants’ explanatory comments are explored, first as a single group and then for each of the high and low proficiency groups.

6.3.1 Overall patterns

This section reports on the overall patterns found for each of the four picture-based narrative tasks, in terms of the three superordinate categories coded for (i.e., conceptual input, code complexity, performance factors) as well as the ‘Other’ category, for Study Part 1 and Part 2 separately. Each source of task complexity will be considered in turn, for all of the tasks in both study parts. Note that sometimes the same person provided comments that explained why a task at hand was difficult in some ways but easy in other ways; hence the percentage of participants who gave comments to justify why the task was easy (or required low perceived effort) and why it was difficult (or required high perceived effort) does not necessarily add up to 100%. As a reminder, Part 1 participants were asked to engage in each narrative task while simultaneously responding to the letter color change (i.e., the secondary color-detection task), while Part 2 participants exclusively engaged in the narrative tasks without any dual-task distractions.

6.3.1.1 Conceptual input

Conceptual input was identified to entail six subordinate categories: (a) storyline/picture quality, (b) simple input making task difficult, (c) logic, (d) number of elements, (e) background, and (f) familiarity/relatability. Patterns in participants’ responses for each task (on its conceptual input) will be reported in the following sections.

6.3.1.1.1 Task 1

As can be seen in Table 21 and Figure 27, almost 40% of the participants in Part 1 expressed that the conceptual input made Task 1 easy to handle and/or require low mental effort.
The majority attributed this easiness specifically to its simple and straightforward storyline.

Some example comments are as follows:

#18 絵の内容がすぐに理解できた [I could understand the storyline right away]

#25 比較的絵がシンプルで説明しやすかった [The pictures were relatively simple and easy to describe.]

#39 今回は話しが作りやすくって … 話しながら写真を読みとる必要がなかった [It was easy to come up with a story, so … there was no need to try to figure out the storyline while speaking.]

Among these 21 participants, eight of them expressed that having just one character to refer to made the task easy to handle. For example, participant 32 said that the task was easy because “there was only one character and his action was easy to describe (登場人物が一人で、行動も簡単なものであったから).” A few participants also attributed their low ratings to the fact that “the story took place in the same setting (同じ場面でストーリーが組み立てられていた)” (Participant 10) and to their familiarity with the story content (Participant 53: “日常でもよく起こることなので、ストーリーはたてやすかったです [It was easy to come up with a story because it happens quite often in my daily life]).

At the same time, almost 30% of the participants expressed that the conceptual input made Task 1 difficult to handle or to require high mental effort. It is interesting to note here that five of them attributed their high task difficulty ratings on Task 1 to the input being too simple:

#36 絵が淡々としたものだったため、自分で話を創らなければならない気がしたから [As the pictures were simple and plain, I felt that I had to create a story on my own.]
The story is just about a man goes to bed and wakes up, so I could not think of a story.

Additionally, some participants (four in number) were also concerned with the logic of the storyline and expressed that:

#25 男の人がなぜ窓の外を見に行ったのかよく分からず、その場面を説明するのは少し困難だった [I didn’t know why the man went by the window to check outside, and it was a bit hard to describe that part of the story.]

#35 カーテンを閉めた理由、それでも光が朝入って来た理由が分からず、どう話をせばいいのか分からなかった [I didn’t know why he closed the curtain or why the sunlight came in in the morning with the curtain shut, so I didn’t know what to say.]

Table 21
Summary Table With the Number of Comments for Each Category of Conceptual Input (Task 1, Part 1)

<table>
<thead>
<tr>
<th>Task 1 (Part 1)</th>
<th>Easy/ Low mental effort</th>
<th>Difficult/ High mental effort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>39.62% (21)</td>
<td>28.30% (15)</td>
</tr>
<tr>
<td>Storyline/picture quality</td>
<td>17</td>
<td>6</td>
</tr>
<tr>
<td>Simple input making task difficult</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Logic</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Number of elements</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Background</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Familiarity/relatability</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

*Note.* The numbers in parentheses indicate the sheer number of comments provided to justify why a given task was easy/low mental effort consuming or difficult/high mental effort consuming. The total number of participants in Part 1 (i.e., the denominator) is 53. The numbers next to each subordinate category indicate the number of comments.
Note. *Storyline special* = Simple storyline made the task difficult to do; *Familiarity* also includes relatability.

Figure 27. Input as a source of easiness and difficulty (Part 1, Task 1).
Study Part 2 participants’ comments were similar to the ones provided by Part 1 participants; however, proportionally more participants of Part 2 expressed that the conceptual input made Task 1 easy to do and/or require little mental effort (see Table 22 and Figure 28). In the second phase, more than 60% of the participants attributed their low rating on Task 1 to the conceptual input. Similarly to Part 1, the simple and straightforward storyline was the most popular reason why the task was easy and/or required low mental effort. Close to 40% of the participants attributed their low ratings to Task 1’s simple storyline. Also, 11 participants expressed that Task 1 was easy/required low mental effort because the story content was familiar or relatable to them:

#11 絵の中に人の感情が読み取れたので、自分のことのように考えることができ、喋りやすかった [I could easily understand the man’s feelings, so I could tell a story as if it was my own, which made the story-telling easy to do.]

#38 男が眠り、起床するも眠たく、アラームを切る、という流れは、自分もこれまでに何度も経験したことがあるから。[The sequence of going to bed, waking up sleepy, and turning the alarm off is something that I have experienced many times in the past.]

The next popular reason had to do with the number of elements. Similar to Part 1 participants, 10 participants in the second phase expressed that having just one character in the story made the task easy and/or require low mental effort. A few people also thought that the man’s actions were all reasonable and it was easy to understand why he took certain actions in a given picture frame and that no change in the background made the task easy to do.

Turning our attention to the comments explaining why Task 1 was difficult to do and/or required high mental effort, 32.84% of the participants’ comments were categorized in this
group, and very similar results were found to those in Study Part 1 as well. Among them, 15
participants expressed that Task 1 was difficult and/or required high mental effort because of its
somewhat ambiguous storyline. In addition to participants’ who expressed that the storyline was
overall hard to decipher, a handful of participants commented that it was hard to understand one
particular part of the storyline. A representative example comment is provided below:

#4 頭を使った所は、男の人は枕をアラームに投げつけているのではなく、単純に
机を倒してしまったのだろうと思ったので、途中でその違いに気付き、訂正した時
[When I used mental effort the most is when I realized that the man did not just turn the
table over, but he threw his pillow at the alarm clock while I was talking and I had to fix
it.]

About the logic of the storyline, some participants expressed that they did not know why the man
went to check outside (similarly to Part 1) or why he decided to go to bed again after once
waking up. Lastly, two of the participants in this phase commented that the simple input of a
man sleeping made it difficult for them to understand the storyline or decide what to focus on.
Table 22
Summary Table With the Number of Comments for Each Category of Conceptual Input (Task 1, Part 2)

<table>
<thead>
<tr>
<th>Task 1 (Part 2)</th>
<th>Easy/ Low mental effort</th>
<th>Difficult/ High mental effort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>61.19% (41)</td>
<td>32.84% (22)</td>
</tr>
<tr>
<td>Storyline</td>
<td>33</td>
<td>15</td>
</tr>
<tr>
<td>Simple input made task difficult</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Logic</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Number of elements</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Background</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Familiarity/relatability</td>
<td>11</td>
<td>0</td>
</tr>
</tbody>
</table>

*Note.* The numbers in parentheses indicate the sheer number of comments provided to justify why a given task was easy/low mental effort consuming or difficult/high mental effort consuming. The total number of participants in Part 2 (i.e., the denominator) is 67. The numbers next to each subordinate category indicate the number of comments.
Note. Storyline special = Simple storyline made the task difficult to do; Familiarity also includes relatability

Figure 28. Input as a source of easiness and difficulty (Part 2, Task 1).
6.3.1.1.2 Task 2

Task 2, overall, revealed radically different patterns from Task 1. More participants in both Part 1 and Part 2 expressed that the conceptual input made Task 2 difficult or require high mental effort (see Tables 23 & 24 and Figures 29 & 30). In Part 1, the majority of participants who rated high in task difficulty and mental effort (43% overall) for Task 2 attributed their high ratings to its unclear storyline, perhaps stemming from the need to pay attention to small details of the pictures in order to successfully understand its storyline:

#47 イラストの後半が何を表しているかよく分からない。物語の展開が読めなかった [I don’t know what the latter part of the pictures is depicting. I couldn’t figure out the storyline.]

#33 絵が細かくて、3コマ目で男の人が何を釣り上げているのか判別できなくて焦ったから [The picture was so fine that I couldn’t figure out what the man was catching in the third frame and I got panicked.]

The twist of this story is that a girl with a makeshift fishing rod manages to catch one fish after another, while a man with a professional outfit and fishing rod can only catch garbage and ends up catching his own boot. To successfully figure the storyline out, then, it is imperative for the speaker to understand that the man caught a can, not a fish, in the third picture frame, then caught his own boot instead of a fish, and that he fell into the water in the last two frames.

Additionally, five participants expressed that having more characters to refer to in the input made Task 2 hard to handle and/or require greater mental effort. Interestingly, among them, two mistakenly thought that the two characters were both men. This misunderstanding, in turn, made the task harder (perhaps than intended) for them because they were faced with unintentional
functional and consequently linguistic demands of distinguishing among similar characters. Here are their comments:

#2 登場人物が男2人で、聞き手に分かるように差別化するのが難しかった [There were two male characters, and it was difficult to distinguish between them in a way that the listener would know (who I was referring to).]

#8 あっちの男、こっちの男みたいに使い分けできないので、とりあえず知ってる単語をつなげて表現してみた [I cannot distinguish (between the two characters) by saying ‘this man’ and ‘that man,’ so I described them by using whatever the words I know.]

Curiously, some participants commented that it was not necessarily the two similar characters but the fact that their actions were intertwined that made Task 2 difficult and/or require high mental effort. For example:

#42 2人の行動を関連づけて説明しなければならず… [In describing the story, I had to relate the two characters’ actions to each other.]

Three participants in Study Part 1 also attributed their high task difficulty on Task 2 to its overly simple input:

#36 絵はシンプルだったが、それゆえに話が単調になると思い余計に考える事が多かった [The pictures were simple, but because of that, I thought the story might turn out monotonous and so I had a lot to think about.]

By contrast, in Part 1, approximately 30% of the participants expressed that Task 2 was easy/required little mental effort because of its simple conceptual input. The majority of comments to this end were related to its simple, straightforward storyline, but a few participants
also attributed their low ratings to having to deal with only a few characters and a fixed background.

Similarly, in Part 2, about the same proportion of participants (i.e., 25.37%) thought Task 2 was relatively easy and/or required less mental effort, due to its simple conceptual input, especially its apparent storyline:

#43 他のイラストに比べて、起きていることが明確だったように感じたから

[Compared to the other pictures, I felt that what was happening (in the story) was obvious.]

A handful of participants also expressed that having only two characters in the input, the fixed background, and relatability (e.g., #8 自分も物語の中にいると思ってストーリーを考えられたから [I could think of a story as if I were in it.]) made the task easy to handle and/or require less mental effort.

Differently from Part 1, in Part 2 (where participants engaged solely in the narrative tasks, without interference from the secondary color detection task), more than half of the participants expressed that Task 2 was difficult and/or required high mental effort due to its conceptual input. Regarding its storyline, in addition to the comments expressing that the storyline was hard to figure out – the kind of comments we saw in Part 1 – there were comments closely related to participants’ effort to tell a good, effective story (provided by five participants). Here are some of their comments:

#19 最後のところまでストーリー順序を保つのが頭を使った [I used much mental effort to keep the order of the story right until the end.]

#34 4 枚目のイラストと 5 枚目のイラストの違いを説明することが難しかったりしたため [It was difficult to describe the fourth and fifth pictures differently.]
It might be the case, then, that figuring out the storyline of Task 2 itself was not more difficult or more mental effort consuming for Study 2 participants than for Study 1 participants; however, attempts to come up with an effective story, which might have become possible due to a lack of the secondary task distraction, made the task more difficult and more attention consuming for some participants in Study Part 2.

Additionally, nine participants in Part 2 (out of 10 whose comments were about the number of elements) also commented that having to refer to two characters, especially because they are doing different things and yet their actions are related, made Task 2 difficult and/or effort consuming:

#9 男の人と女の子の2人がいるので、それぞれについて考えないといけなかったから [There were a man and a girl, so I had to think about each of them.]

#16 2人の行動が対照的だったので、その対照的さを説明するのに頭を使った [The two characters’ actions were contrasting, so I used much mental effort to describe the contrast.]

The other participant who provided a comment about the number of elements was more concerned with the number of objects, rather than the number of characters involved in the story:
#4 物が多く出てくる。つり針、ブーツ、枝、つりざお etc… [There were a lot of objects in the story, like a fishhook, boots, a twig, a fishing rod, etc…]

Seven participants also expressed that the simple input of Task 2 made it more difficult.

Lastly, seven participants pointed to the fact that the six picture frames contained little information and little change, which made the task harder to deal with and require higher mental effort:

#69 シーンの変化があまりなかったため、情報が少なかったため、話す内容を考えるのに時間がかかりました [The story had little change in its scenes and little information, so it took a while to think of what to talk about.]

Table 23
Summary Table With the Number of Comments for Each Category of Conceptual Input (Task 2, Part 1)

<table>
<thead>
<tr>
<th>Task 2 (Part 1)</th>
<th>Easy/ Low mental effort</th>
<th>Difficult/ High mental effort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>30.19% (16)</td>
<td>43.40% (23)</td>
</tr>
<tr>
<td>Storyline</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>Simple input made task difficult</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Logic</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Number of elements</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Background</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Familiarity/relatability</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Note. The numbers in parentheses indicate the sheer number of comments provided to justify why a given task was easy/low mental effort consuming or difficult/high mental effort consuming. The total number of participants in Part 1 (i.e., the denominator) is 53. The numbers next to each subordinate category indicate the number of comments.
Note. *Storyline special* = Simple storyline made the task difficult to do; *Familiarity* also includes relatability

*Figure 29.* Input as a source of easiness and difficulty (Part 1, Task 2).
Table 24
Summary Table With the Number of Comments for Each Category of Conceptual Input (Task 2, Part 2)

<table>
<thead>
<tr>
<th>Task 2 (Part 2)</th>
<th>Easy/ Low mental effort</th>
<th>Difficult/ High mental effort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25.37% (17)</td>
<td>52.24% (35)</td>
</tr>
<tr>
<td>Storyline</td>
<td>15</td>
<td>23</td>
</tr>
<tr>
<td>Simple input made task difficult</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Logic</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Number of elements</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>Background</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Familiarity/relatability</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Note. The numbers in parentheses indicate the sheer number of comments provided to justify why a given task was easy/low mental effort consuming or difficult/high mental effort consuming. The total number of participants in Part 2 (i.e., the denominator) is 67. The numbers next to each subordinate category indicate the number of comments.
Note. *Storyline special* = Simple storyline made the task difficult to do; *Familiarity* also includes relatability.

*Figure 30.* Input as a source of easiness and difficulty (Part 2, Task 2).
6.3.1.1.3 Task 3

Unlike Tasks 1 and 2, the overall patterns in responses for Task 3 in each phase turned out to be somewhat distinct. In Part 1, a bigger proportion of participants expressed that Task 3 was easy and/or did not require much mental effort due to its conceptually simple input, whereas in Part 2, the proportions of participants who thought Task 3 was easy (and/or not-so-mental effort consuming) and difficult (and/or mental effort consuming) were about the same. In Part 1, as seen in Table 25 and Figure 31, more than half of the participants expressed that it was easy to figure out the storyline. Two participants each also attributed their low task difficulty ratings to a fixed background and the number of characters:

#48 動くキャラクターがこどもだけだったから [The characters in action were only the children.]

#52 そのストーリーの中での登場人物が、1つの家庭の中にいることから、人物たちの関係性がつかみやすかったから [The characters were all in one home, so it was easy to imagine their relationship with each other.]

Interestingly, only a few participants expressed that it was difficult to figure out the storyline of Task 3 per se. There were more participants who expressed that some specific parts of Task 3 was conceptually complicated, and these were related to challenges: to (a) distinguish between the fourth and fifth picture frames, (b) realize that the mother was in fact sleeping, not being distracted by the boy in Frame 3, and lastly (c) describe the bowling ball with a face on it:

#2 ストーリーはとてもやさしかったが、4枚目と5枚目のイラストを違うように説明できなかった [The story itself was easy (to describe), but I could not distinguish between the fourth and fifth picture frames.]
#20 絵の細部（お母さんが寝ている）を注視しなければ、ストーリーの組み立てが難しいから [Creating a story requires paying attention to the details of the pictures (i.e., the mom is sleeping).]

#18 日本語でも何と言っていいのか分からないもの（顔が描かれたボール）が出てきたから [There was something that I wouldn’t even know how to describe in Japanese (i.e., the ball with a face on).]

Additionally, some participants (four in number) also attributed their high ratings to linguistic consequences of having to deal with multiple characters as well as logic of the storyline:

#35 lady/mother, boy/son/brother などをどのように使おうかで迷った。登場人物をどう指せばいいのかで悩んだ [I had a hard time deciding how to use terms, such as lady/mother and boy/son/brother. I wasn’t sure how to refer to each character.]

#37 ボーリングボールがなぜ出てきたのか、どこから出てきたのか分からなくて説明に悩んだ [I didn’t know where the bowling ball came from and why, so I wasn’t sure how to describe the pictures.]
Table 25
Summary Table With the Number of Comments for Each Category of Conceptual Input (Task 3, Part 1)

<table>
<thead>
<tr>
<th>Task 3 (Part 1)</th>
<th>Easy/ Low mental effort</th>
<th>Difficult/ High mental effort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>52.83% (28)</td>
<td>32.08% (17)</td>
</tr>
<tr>
<td>Storyline</td>
<td>28</td>
<td>13</td>
</tr>
<tr>
<td>Simple input made task difficult</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Logic</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Number of elements</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Background</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Familiarity/relatability</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Note. The numbers in parentheses indicate the sheer number of comments provided to justify why a given task was easy/low mental effort consuming or difficult/high mental effort consuming. The total number of participants in Part 1 (i.e., the denominator) is 53. The numbers next to each subordinate category indicate the number of comments.
Note. Storyline special = Simple storyline made the task difficult to do; Familiarity also includes relatability

Figure 31. Input as a source of easiness and difficulty (Part 1, Task 3).
Moving to Part 2, this time, more than half of the participants found Task 3 to be difficult and/or require high mental effort due to its conceptual input (see Table 26 and Figure 32). The majority of the comments had to do with its storyline. Similarly to Part 1, in addition to some participants who expressed that it was difficult to figure out the overall storyline, there was a greater number of participants who claimed that they had a hard time (a) trying to refer to the bowling ball, (b) distinguishing between similar picture frames, (c) realizing that the mother was taking a nap, (d) reading characters’ intentions and feelings, and (e) connecting each picture frame to create a coherent story. The last two themes are unique to this particular study part:

#67 1枚づつのスライドの説明ではストーリーになりずらかったため [It was hard to create a story by simply describing each slide.]

#52 登場人物…（特に子どもたち）が何を考えて、行動をとったのか思いつくのに少し時間がかかりました [It took a while to reason about the characters’ (especially children’s) intentions behind their actions.]

#53 母は驚いているのか、怒っているのか絵から読みとれなかったのも少し苦労した [I couldn’t figure out whether the mother was surprised or angry, so I had a hard time a bit.]

Similar to Task 2, then, the high proportion of participants expressing that Task 3 was difficult/effort consuming due to its storyline seems to be related to their efforts to tell a good, effective story. Furthermore, 11 participants expressed that having to deal with multiple characters made the task difficult and/or require high mental effort. Here, in addition to a challenge of referring to each character properly – a kind of comment we saw in Part 1 participants’ responses – other linguistic as well as conceptual demands stemming from having multiple characters came up in Part 2 participants’ comments:
The subject changes frequently, for example, from a boy to a girl to a woman and then to the boy and the girl, so I had to change the subject and use appropriate verb conjugations while I was telling the story.

There were a lot of characters, so I had to keep track of who did what.

Lastly, a few participants also expressed that the relatively simple input made the task difficult to deal with, especially to elaborate on the simple story and to translate a full-on story that they could easily come up with in Japanese.

On the other hand, more than 40% of the participants attributed easiness of and low effort required for Task 3 to its simple conceptual input. Among the subordinate categories, its straightforward storyline was the most common answer to justify their low ratings on difficulty and mental effort, similarly to Part 1. Three participants also gave comments that were related to the number of elements. Here, interestingly, in addition to a response that there were only a few characters involved in the story, one person each expressed that dissimilarity among the characters in Task 3 and a small number of main characters made the task easy to handle.

Task 2 had two male characters, whereas in this task I was able to distinguish among the characters by referring them as a baby, a mother (woman), guys (a boy and a girl).
#46 メインの woman と kids に注目すればストーリーは分かりやすく伝えることが出来ると思った [I thought the story can be easily conveyed if I paid attention to the woman and the kids as main characters.]

Table 26
Summary Table With the Number of Comments for Each Category of Conceptual Input (Task 3, Part 2)

<table>
<thead>
<tr>
<th>Task 3 (Part 2)</th>
<th>Easy/ Low mental effort</th>
<th>Difficult/ High mental effort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>43.28% (29)</td>
<td>50.75% (34)</td>
</tr>
<tr>
<td>Storyline</td>
<td>27</td>
<td>24</td>
</tr>
<tr>
<td>Simple input made task difficult</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Logic</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Number of elements</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>Background</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Familiarity/relatability</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Note. The numbers in parentheses indicate the sheer number of comments provided to justify why a given task was easy/low mental effort consuming or difficult/high mental effort consuming. The total number of participants in Part 2 (i.e., the denominator) is 67. The numbers next to each subordinate category indicate the number of comments.
Note. *Storyline special* = Simple storyline made the task difficult to do; *Familiarity* also includes relatability

*Figure 32.* Input as a source of easiness and difficulty (Part 2, Task 3).
6.3.1.1.4 Task 4

Finally, looking at Task 4, the patterns in participants’ responses seem quite similar to Task 2. Overall, in both phases of the study, far more participants expressed that Task 4 was difficult and/or required high mental effort rather than the opposite because of its conceptual input (see Tables 27 & 28 and Figures 33 & 34). More specifically, almost 60% of the participants in Part 1 attributed their high ratings in task difficulty and mental effort on Task 4 to its conceptual input, in particular its complicated storyline and a large number of characters involved in the story. Here’s a sample comment about the conceptually complex storyline:

#34 ストーリーが分からなかったので、英語を使いながらストーリーを理解しようとしてしまった [I couldn’t understand the storyline, so I was trying to figure it out while telling the story in English.]

The comments regarding the number of elements in this task are particularly illuminating and interesting. Having a large number of characters to deal with in Task 4 required participants not only to distinguish among similar characters (linguistic demands), but also to figure out the relationship among them, who should be described, who’s who in each picture frame, and so on (conceptual demands):

#46 はしごを登る人、カフェでお茶する人、ドライバー、犬の飼い主の因果関係を説明しようと頭を使いました [I was trying to explain the cause-and-effect relationship of the man on the ladder, people having tea at a café, the driver of the car, and the owner of the dog, and this required a lot of mental effort.]

#41 4人くらいの登場人物がいて誰を説明すべきなのかと最初迷った [There were four characters or so, and I wasn’t sure who I should describe at first.]
There were so many people involved in the story that I couldn’t tell who was carried away by an ambulance.

Additionally, perhaps related to having multiple characters in this case, some participants felt that they had a lot to describe in Task 4.

There was a lot to be described.

By contrast, there were seven participants in Part 1 who thought Task 4 presented a straightforward storyline and hence it was relatively easy to deal with and/or not very effort consuming.

Table 27
Summary Table With the Number of Comments for Each Category of Conceptual Input (Task 4, Part 1)

<table>
<thead>
<tr>
<th>Task 4 (Part 1)</th>
<th>Easy/ Low mental effort</th>
<th>Difficult/ High mental effort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>13.21% (7)</td>
<td>58.49% (31)</td>
</tr>
<tr>
<td>Storyline</td>
<td>7</td>
<td>17</td>
</tr>
<tr>
<td>Simple input made task difficult</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Logic</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Number of elements</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>Background</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Familiarity/relatability</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Note. The numbers in parentheses indicate the sheer number of comments provided to justify why a given task was easy/low mental effort consuming or difficult/high mental effort consuming. The total number of participants in Part 1 (i.e., the denominator) is 53. The numbers next to each subordinate category indicate the number of comments.
Figure 3.1. Input as a source of easiness and difficulty (Part 1, Task 4).
The patterns for Part 2 are very similar to those for Part 1. More than 60% of the participants in Part 2 gave comments related to its conceptual input to justify why Task 4 was difficult and/or required high mental effort. Among the subordinate categories, the most popular one was the number of elements. In Part 2 too, we see comments about linguistic and conceptual demands related to having to deal with multiple characters, as well as comments to express that there was a lot to describe:

#11 やはり登場人物が増えてくると、それをどうやって区別しようか考えるのに必死になってしまい、あまり他のことに気がまわらなくなる [When the number of characters grows in number, I work so hard to figure out how to distinguish among them that I cannot pay much attention to other things.]

#52 誰を主役にして話を組み立てるべきか、考えるのに時間がかかったから [It took some time to decide who the main character should be.]

#8 ストーリーの中のどの人が同一人物なのかがよく分からなくて、被害を被った人が誰なのかいまいち分からず、どう説明したらいいのか、分からなかったから [I wasn’t sure who’s who is the story, and I wasn’t too sure who was injured. So, I didn’t know what to describe the pictures.]

#55 一つの絵あたりの情報量が多く感じたので、あれもこれもと説明していたら、自分が何を言っているかわからなくなってしまった [Each picture frame had a lot of information and I was explaining everything. Then, I got confused as to what I was saying.]

20 participants also expressed that the storyline was hard to figure out. To give one interesting comment:
The story became difficult all of a sudden. Until now, each picture frame depicted similar time frames, but this time there was a wide gap between frames, so I had a hard time describing the pictures.

Although this comment was a minor one, it is interesting to see that the amount of imagination or the amount of gap to fill in required to tell the story depicted in Task 4 made the task conceptually difficult.

On the other hand, approximately 20% of the participants in Part 2 expressed that the conceptual input made Task 4 easy and/or low mental effort consuming. The majority of the comments had to do with its storyline, but a few people also attributed their relatively low ratings to the familiarity/relatability of the story content and having things to talk about (i.e., a large number of elements).

### Table 28
Summary Table With the Number of Comments for Each Category of Conceptual Input (Task 4, Part 2)

<table>
<thead>
<tr>
<th>Task 4 (Part 2)</th>
<th>Easy/ Low mental effort</th>
<th>Difficult/ High mental effort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20.90% (14)</td>
<td>62.69% (42)</td>
</tr>
<tr>
<td>Storyline</td>
<td>11</td>
<td>20</td>
</tr>
<tr>
<td>Simple input made task difficult</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Logic</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Number of elements</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>Background</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Familiarity/relatability</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

*Note. The numbers in parentheses indicate the sheer number of comments provided to justify why a given task was easy/low mental effort consuming or difficult/high mental effort consuming. The total number of participants in Part 2 (i.e., the denominator) is 67. The numbers next to each subordinate category indicate the number of comments.*
Note. *Storyline special* = Simple storyline made the task difficult to do; *Familiarity* also includes relatability

*Figure 34.* Input as a source of easiness and difficulty (Part 2, Task 4).
6.3.1.2 Code complexity

As can be seen in the series of Tables and Figures below, all four tasks posed high linguistic code demands on many of the participants in both Part 1 and Part 2 of the study. This finding is not so surprising because all these participants were living in Japan at the time of the experiment and were speakers of English as a foreign language. As described in the method section, their comments regarding code complexity were categorized into the following three subcategories: (a) task-induced, (b) effort for a better performance (language-related), and (c) strategy to deal with a given task (language-related). In what follows, I will report findings for each task.

6.3.1.2.1 Task 1

In both Part 1 and Part 2, approximately half of the participants expressed that the linguistic aspects of Task 1 made it difficult and/or required lots of mental effort (see Tables 29 & 30 and Figures 35 & 36). The most popular comments in this superordinate category had to do with task-induced code complexity. Within this subordinate category, the comments made by Part 1 participants included: (a) a perceived lack of general English proficiency, vocabulary, and grammar, (b) a lack of specific task-essential words and expressions, and although rarely (c) increased effort for linguistic aspects of task performance caused by the simple conceptual input.

#11 英語が苦手だから [I’m not good at English.]

#1 「まくら」が出てこない [I didn’t know how to say ‘pillow.’]

#31 今回は話が作りやすくって，それを描写するために頭を使った [This time, it was easy to come up with a story and I used much mental effort to actually describe it (in English).]
Task 1 elicited similar comments from Part 2 participants, with regards to the task-induced code complexity; however, there was one additional theme that the task elicited in Part 2: participants’ efforts to paraphrase words and phrases that they did not know. For example:

#47 自分の知っている簡単な単語に置き換えるのに頭を使った [I used much mental effort in paraphrasing with simple words that I do know.]

The next subordinate category, effort for a better performance, accounted for about 30% of the code-complexity-related comments given to explain why Task 1 was difficult and/or required much mental effort. These comments had to do with: (a) learners’ attention to accuracy and fluency, as well as (b) their efforts to follow the task instructions to tell a better story.

#28 文法がごちゃごちゃなことに途中で気づいたから（難しく感じた） [Because I realized that my grammar was messed up while I was talking, (I thought the task was difficult).]

#34 ことばにつまってしまい、止まりすぎたらだめだと思って焦ってことばを探した [I wasn’t fluent. I thought I shouldn’t stop too much and I was looking for words frantically.]

#40 面白いストーリーにしようと思ったが、途中でアイディアが切れてしまった。そうしたら、出てくるはずの表現も出てこなくなり、難しかったと感じた [I tried to tell an interesting story, but I ran out of ideas along the way. Then, I couldn’t even say expressions that I should have been familiar with and thought the task was difficult.]

In Part 2, participants provided largely similar responses, except that one person made a comment about lexical and syntactic complexity in this phase, as seen below:
By contrast, there were a handful of participants in both Part 1 and Part 2 who said that the language required to tell the story of Task 1 was easy and simple. A few people in each part of the study also confessed that they ended up devoting only little mental effort in an effort to deal with a cognitively demanding activity at hand:

#5 (Part 1) 分からない単語は考えても出てこないので、半分あきらめが入っている [I cannot think of words that I don’t know, so I was half giving up.]

#48 (Part 2) 最後を説明するとき単語が上手く出てこなくて簡単に言うことしかできなかった [When describing the last part, I couldn’t come up the right words, so I could describe it only briefly.]

It is interesting to observe that all comments (for this task) related to participants’ strategies to deal with a given task, in both parts of the study, were provided to justify why they could devote only little mental effort to Task 1, and these were not cited in reference to its difficulty.
Table 29
*Summary Table With the Number of Comments for Each Category of Code Complexity (Task 1, Part 1)*

<table>
<thead>
<tr>
<th>Task 1 (Part 1)</th>
<th>Easy/ Low mental effort</th>
<th>Difficult/ High mental effort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>22.64% (12)</td>
<td>47.17% (25)</td>
</tr>
<tr>
<td>Task-induced</td>
<td>10</td>
<td>22</td>
</tr>
<tr>
<td>Effort for a better performance (language-related)</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Strategy to deal with a given task (language-related)</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

*Note.* The numbers in parentheses indicate the sheer number of comments provided to justify why a given task was easy/low mental effort consuming or difficult/high mental effort consuming. The total number of participants in Part 1 (i.e., the denominator) is 53. The numbers next to each subordinate category indicate the number of comments.

Table 30
*Summary Table With the Number of Comments for Each Category of Code Complexity (Task 1, Part 2)*

<table>
<thead>
<tr>
<th>Task 1 (Part 2)</th>
<th>Easy/ Low mental effort</th>
<th>Difficult/ High mental effort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>26.87% (18)</td>
<td>53.73% (36)</td>
</tr>
<tr>
<td>Task-induced</td>
<td>15</td>
<td>31</td>
</tr>
<tr>
<td>Effort for a better performance (language-related)</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>Strategy to deal with a given task (language-related)</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

*Note.* The numbers in parentheses indicate the sheer number of comments provided to justify why a given task was easy/low mental effort consuming or difficult/high mental effort consuming. The total number of participants in Part 2 (i.e., the denominator) is 67. The numbers next to each subordinate category indicate the number of comments.
Figure 35. Code complexity as a source of easiness and difficulty (Part 1, Task 1).

- Easy/low mental effort
- Difficult/high mental effort

- Effort for a better performance (lg-related)
- Strategy to deal with a given task (lg-related)
- Task-induced
Figure 36. Code complexity as a source of easiness and difficulty (Part 2, Task 1).
6.3.1.2.2 Task 2

The patterns in participants’ responses for Task 2 are even clearer than those for Task 1. The majority of the participants (more than 60% in Part 1 and a little more than 70% in Part 2) commented that Task 2 posed high code complexity (see Tables 31 & 32 and Figures 37 & 38). In addition to their perceived lack of English proficiency, vocabulary, and grammar in general, many of them were unsure of words related to fishing as seen in the following comments:

#71 (Part 2) 鉤りについての話だったが、釣りに使う物（うさぎやうきなど）の英単語がわからなかったので [The story was about fishing, but I didn’t know words for fishing tools (e.g., fishing rod, bob).]

#41 (Part 1) 魚が鉤にかかったとき何と言えばいいかわからなかった。「釣り」のシチュエーションに関する語彙がなかったため [I didn’t know what to say when the girl got a fish on the hook. I was short of words related to the situation of ‘fishing.’]

#12 (Part 2) “釣る”という英語が思い浮かばなかったのが大きかった。終始caughtを使っていて自信がない [The biggest reason (why I felt the task to be difficult) was because I couldn’t think of how to say ‘to catch (a fish).’ I ended up using ‘caught’ but I’m not confident if that was correct.]

Additionally, in both Part 1 and Part 2, some participants expressed that they devoted additional mental effort to doing Task 2 in an effort to paraphrase words and phrases that they were not familiar with.

#71 (Part 2) 釣りに使う物の英単語がわからなかったので、それらを使わずにストーリーを伝えられるように考えた。[I didn’t know words for fishing tools, so I thought a lot to get the story across without using these words.]
Seven participants in Part 1 and 21 participants in Part 2 also touched upon the idea that their effort to tell a better story made Task 2 more difficult and/or require more mental effort. Their comments revolved around three main themes: (a) their attention to accuracy and fluency (although the latter was not so common), (b) their efforts to follow the task instructions, and interestingly (c) their attempt to vary their choice of lexis. Regarding the last theme, participants provided the following comments:

#32 (Part 1) 動詞の fish と名詞の fish の言い換えがないか話しながら探していたから [I was looking for synonyms for a verb ‘fish’ and a noun ‘fish’ while talking.]

#32 (Part 2) 初めは簡単に思えたが、同じ表現ばかり使うことになってしまった（ので、難しく感じた） [I thought it was easy at first, but I ended up using the same expressions over and over again (and so I thought the task was difficult.).]

#27 (Part 2) この絵も動きがあまりないので、同じような文の繰り返しをするのは避けたかったから、違う雰囲気の文章で伝えようと思った [This picture set contained not many motions. I wanted to avoid using the same sentences repeatedly, so I tried to tell a story with different sentences.]

Although less than 10% of the participants in both Part 1 and Part 2, a few participants expressed that Task 2 only required simple words and expressions. Lastly, participants’ comments about their strategies to deal with Task 2 were more closely related to their mental effort ratings than their difficulty ratings – a similar trend observed in relation with Task 1. Some of the comments revealed that the high cognitive complexity posed by Task 2 left some participants no choice but to avoid using words that they were not familiar with or repeat the same words and phrases to get the task done, which in turn prevented them from paying much mental effort as a result. The other comments were perhaps more idiosyncratic and included
themes such as coming up with a story directly in English, telling a bare minimum story, and not worrying about accuracy to just say whatever came to mind. Hence, high code complexity posed by Task 2 seems to have had distinct effects on learners’ mental effort. The consequence for some participants was that they devoted more attentional resources, for example, by finding ways to describe their story with the language that they did have, to deal with high cognitive demands. On the other hand, the same phenomenon had a negative effect on other participants, and they gave up or ended up not devoting much mental effort as a result.

Table 31
Summary Table With the Number of Comments for Each Category of Code Complexity (Task 2, Part 1)

<table>
<thead>
<tr>
<th>Task 2 (Part 1)</th>
<th>Easy/ Low mental effort</th>
<th>Difficult/ High mental effort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task-induced</td>
<td>3</td>
<td>28</td>
</tr>
<tr>
<td>Effort for a better performance (language-related)</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Strategy to deal with a given task (language-related)</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Note. The numbers in parentheses indicate the sheer number of comments provided to justify why a given task was easy/low mental effort consuming or difficult/high mental effort consuming. The total number of participants in Part 1 (i.e., the denominator) is 53. The numbers next to each subordinate category indicate the number of comments.
Table 32
Summary Table With the Number of Comments for Each Category of Code Complexity (Task 2, Part 2)

<table>
<thead>
<tr>
<th>Task 2 (Part 2)</th>
<th>Easy/ Low mental effort</th>
<th>Difficult/ High mental effort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>16.42% (11)</td>
<td>70.15% (47)</td>
</tr>
<tr>
<td>Task-induced</td>
<td>6</td>
<td>37</td>
</tr>
<tr>
<td>Effort for a better performance</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>(language-related)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strategy to deal with a given task</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>(language-related)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. The numbers in parentheses indicate the sheer number of comments provided to justify why a given task was easy/low mental effort consuming or difficult/high mental effort consuming. The total number of participants in Part 2 (i.e., the denominator) is 67. The numbers next to each subordinate category indicate the number of comments.
Figure 37. Code complexity as a source of easiness and difficulty (Part 1, Task 2).
Figure 38. Code complexity as a source of easiness and difficulty (Part 2, Task 2).
6.3.1.2.3 Task 3

Task 3 also posed relatively high cognitive demands in terms of its code or linguistic aspects (see Tables 33 & 34 and Figures 39 & 40). More than half of the participants in both parts of the study expressed that the language required to tell the story made Task 3 difficult and/or mental effort consuming. Once again, participants provided comments regarding their perceived lack of English proficiency and a lack of specific words and phrases that were essential or useful in telling the story of Task 3. These specific words and phrases included: to sleep, to hold, to exchange, to take something out of somewhere, to put a baby to sleep, to pick a baby up in one’s arm, to plot out, to play a prank, secretly, and a cradle. Similarly to the other tasks, some participants also attributed their high effort ratings to their attempts to paraphrase these unfamiliar expressions with the ones that they do know. Additionally, five people in Part 1 and 11 people in Part 2 expressed that their efforts to follow the task instructions, and attention to accuracy and sometimes to fluency, made the task difficult and/or require much mental effort. Curiously, in both parts of the study, we see participants making comments about their attention to verb tenses more often than in the other tasks (two participants and four participants, respectively, in each part). Here are some sample comments:

#52 (Part 1) 話す順番や時制に集中していたから [I was paying attention to the order of the story and the verb tenses.]

#12 (Part 2) やはりこれも時制をどうするか困った。前回までのを含えて過去形と心に決めていても、絵を見ると現在形にしたくなる。[I wasn’t sure which verb tense to use. Based on my own performances so far, I’ve decided to use the past tense, but when I see the pictures, I am inclined to use the present tense.]
It was difficult to tell the story, paying careful attention to the verb tenses in English.

Turning our attention to participants’ comments to justify why Task 3 was easy and/or required low mental effort, language-wise, a rather small proportion of participants expressed that Task 3 only required familiar words that could easily be accessed. In addition, two participants in Part 2 commented on their strategies to deal with Task 3. One of them said that the strategy to stick with the past tense instead of the present tense worked out well, which made the task easier to handle for her. The other one expressed that he prioritized task accomplishment more than his performance quality, and this strategy had a consequence of requiring only little mental effort.

Table 33
Summary Table With the Number of Comments for Each Category of Code Complexity (Task 3, Part 1)

<table>
<thead>
<tr>
<th>Task 3 (Part 1)</th>
<th>Easy/ Low mental effort</th>
<th>Difficult/ High mental effort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>13.21% (7)</td>
<td>58.49% (31)</td>
</tr>
<tr>
<td>Task-induced</td>
<td>7</td>
<td>27</td>
</tr>
<tr>
<td>Effort for a better performance</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>(language-related)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strategy to deal with a given task</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(language-related)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. The numbers in parentheses indicate the sheer number of comments provided to justify why a given task was easy/low mental effort consuming or difficult/high mental effort consuming. The total number of participants in Part 1 (i.e., the denominator) is 53. The numbers next to each subordinate category indicate the number of comments.
Table 34
Summary Table With the Number of Comments for Each Category of Code Complexity (Task 3, Part 2)

<table>
<thead>
<tr>
<th>Task 3 (Part 2)</th>
<th>Easy/ Low mental effort</th>
<th>Difficult/ High mental effort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>17.91% (12)</td>
<td>68.66% (46)</td>
</tr>
<tr>
<td>Task-induced</td>
<td>11</td>
<td>42</td>
</tr>
<tr>
<td>Effort for a better performance (language-related)</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>Strategy to deal with a given task (language-related)</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

*Note.* The numbers in parentheses indicate the sheer number of comments provided to justify why a given task was easy/low mental effort consuming or difficult/high mental effort consuming. The total number of participants in Part 2 (i.e., the denominator) is 67. The numbers next to each subordinate category indicate the number of comments.
Figure 39. Code complexity as a source of easiness and difficulty (Part 1, Task 3).
Figure 40. Code complexity as a source of easiness and difficulty (Part 2, Task 3).
6.3.1.2.4 Task 4

The patterns for Task 4, especially in Part 1, could not be any clearer. As can be seen in Tables 35-36 and Figures 41-42 below, Task 4 posed very high code complexity to the majority of the participants (i.e., 73.58% of the participants in Part 1 and 65.67% in Part 2). Similarly to the other three tasks, participants’ comments about code complexity for Task 4 included themes of: (a) their perceived lack of English proficiency in general, and (b) their unfamiliarity with some specific words, such as to go up a ladder, to run onto the sidewalk, to run into, to jump onto, to avoid the dog, to swerve, to hit someone, to let something go, to get injured, to arrest, a prison, an ambulance, a ladder, and a dog leash. A handful of participants also expressed that Task 4 was difficult and/or required high mental effort because they were paying particular attention to accuracy, fluency, and meeting the task instructions. Unlike the rest of the tasks, Task 4 elicited fewer comments about linguistic consequences of carefully following the task instructions than comments about accuracy and fluency, within this subordinate category of ‘effort for a better performance (language-related).’ Only a small portion of participants (especially small in Part 1) said that they could easily come up with the words needed to tell the story of Task 4.
### Table 35
**Summary Table With the Number of Comments for Each Category of Code Complexity (Task 4, Part 1)**

<table>
<thead>
<tr>
<th>Task 4 (Part 1)</th>
<th>Easy/ Low mental effort</th>
<th>Difficult/ High mental effort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task-induced</td>
<td>1.89% (1)</td>
<td>73.58% (39)</td>
</tr>
<tr>
<td>Effort for a better performance (language-related)</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>Strategy to deal with a given task (language-related)</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

*Note.* The numbers in parentheses indicate the sheer number of comments provided to justify why a given task was easy/low mental effort consuming or difficult/high mental effort consuming. The total number of participants in Part 1 (i.e., the denominator) is 53. The numbers next to each subordinate category indicate the number of comments.

### Table 36
**Summary Table With the Number of Comments for Each Category of Code Complexity (Task 4, Part 2)**

<table>
<thead>
<tr>
<th>Task 4 (Part 2)</th>
<th>Easy/ Low mental effort</th>
<th>Difficult/ High mental effort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task-induced</td>
<td>11.94% (8)</td>
<td>65.67% (44)</td>
</tr>
<tr>
<td>Effort for a better performance (language-related)</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Strategy to deal with a given task (language-related)</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

*Note.* The numbers in parentheses indicate the sheer number of comments provided to justify why a given task was easy/low mental effort consuming or difficult/high mental effort consuming. The total number of participants in Part 2 (i.e., the denominator) is 67. The numbers next to each subordinate category indicate the number of comments.
Figure 41. Code complexity as a source of easiness and difficulty (Part 1, Task 4).
Figure 42. Code complexity as a source of easiness and difficulty (Part 2, Task 4).
6.3.1.3 Performance factors

Performance factors included the following two themes related to participants’ perceived performance quality: (a) participants’ strategy to deal with a given task and (b) effort for a better performance. As seen in Tables 37 and 38 below, overall participants in both Part 1 and Part 2 commented on performance factors least often, in comparison with the other factors reported above. Regardless of the tasks, these comments were related to (a) participants’ efforts to follow the task instructions carefully, to pay attention to other performance factors (e.g., whether the story is easy to understand, describing details of the story, adding characters’ emotions to the story), and to perform well in general, as well as (b) their strategies to deal with high cognitive demands of each task (e.g., prioritizing task accomplishment than performance quality). See Table 39 below for some example comments for each task in each part of the study.

Table 37
Summary Table With the Number of Comments for Each Category of Performance Factors (Part 1)

<table>
<thead>
<tr>
<th>Part 1</th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Task 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy/ Low ME</td>
<td>Difficult/ High ME</td>
<td>Easy/ Low ME</td>
<td>Difficult/ High ME</td>
<td>Easy/ Low ME</td>
</tr>
<tr>
<td>11.32% (6)</td>
<td>11.32% (6)</td>
<td>11.32% (6)</td>
<td>15.09% (8)</td>
<td>1.89% (1)</td>
</tr>
<tr>
<td>Effort for a better performance</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>Strategy to deal with a given task</td>
<td>6</td>
<td>0</td>
<td>6</td>
<td>0</td>
</tr>
</tbody>
</table>

Note. ME = mental effort; The numbers in parentheses indicate the sheer number of comments provided to justify why a given task was easy/low mental effort consuming or difficult/high mental effort consuming. The total number of participants in Part 1 (i.e., the denominator) is 53. The numbers next to each subordinate category indicate the number of comments.
Table 38

Summary Table With the Number of Comments for Each Category of Performance Factors (Part 2)

<table>
<thead>
<tr>
<th>Part 2</th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Task 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easy/ Low ME</td>
<td>Difficult/ High ME</td>
<td>Easy/ Low ME</td>
<td>Difficult/ High ME</td>
</tr>
<tr>
<td></td>
<td>4.48% (3)</td>
<td>8.96% (6)</td>
<td>1.49% (1)</td>
<td>7.46% (5)</td>
</tr>
</tbody>
</table>

Effort for a better performance strategy to deal with a given task

|        | 0 | 6 | 0 | 5 | 0 | 8 | 0 | 6 |

|        | 3 | 0 | 1 | 0 | 2 | 0 | 1 | 0 |

**Note.** ME = mental effort; The numbers in parentheses indicate the sheer number of comments provided to justify why a given task was easy/low mental effort consuming or difficult/high mental effort consuming. The total number of participants in Part 2 (i.e., the denominator) is 67. The numbers next to each subordinate category indicate the number of comments.

It is interesting to observe that most of the comments about performance factors were justifications of why the task required high or low mental effort, rather than their difficulty ratings.
| Task 1 | #11 ふりしぼってやろうと思ったから [I thought I should try really hard.] | #4 思いつきでどんどんしゃべっていった [I spoke whatever came to my mind.] | #41 話を短くまとめてしまったため [I ended up summarizing the story briefly.] |
|        | #21 今回は順序にこだわってみました [This time, I decided to focus on the order.] |                                                          | |
| Task 2 | #7 全てのイラストをストーリーに含ませること、明確に表現するのが難しいので [It was too difficult to include all pictures in the story and to describe it clearly.] | #22 魚釣りというもののが伝わればいいと思った [I thought it would be enough to convey that it's fishing.] | #49 2人のシチュエーションのちがい、最後に何が起こったかを具体的に話すのが難しくて頭を使いました。[It was difficult to talk concretely about the difference between the two characters in their situations and what happened in the end, so I used much mental effort.] |

(continued)
### Part 1

<table>
<thead>
<tr>
<th>Effort for a better story (performance-related)</th>
<th>Strategy to deal with a given task (performance-related)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 3</td>
<td></td>
</tr>
<tr>
<td>#13 さっきよりストーリーを考えようとしたから [I tried to think of a story (rather than simply describing pictures) more than before.]</td>
<td>#27 途中で諦めて絵を飛ばしてしまったから [I gave up and skipped a picture in the middle.]</td>
</tr>
<tr>
<td>#46 surprise しようとしているので、その内容を相手にキチンと伝えるには、どこをどう話し、強調すべきか悩んだ [The children were trying to surprise the mother, so I wondered what I should say in what way and what to emphasize in order to convey the story content to the listener properly.]</td>
<td>#53 子どもたちはいたずらをしているのか、お母さんにかまってもらいたいので、など余計なことまで考えてしまった [I was thinking too much, like whether the children are playing a prank on the mother or they wanted to get her attention.]</td>
</tr>
<tr>
<td>Task 4</td>
<td></td>
</tr>
<tr>
<td>#14 創造力を使わなければならなため [I had to use creativity.]</td>
<td>#13 絵のストーリーを考える余裕がないから [I couldn’t afford thinking of a story.]</td>
</tr>
</tbody>
</table>
| #12 今回初めて創造性のことも考えて英語で説明しようとしたから。話したであろうセリフを入れてみたが、話をつけられ Sailor が少し減った気がする。例えば ... セリフを入れてみたことで少しやりやすくなった [For the first time, I paid attention to creativity when describing the pictures in English this time. I inserted lines that the characters would have said and referred to their feelings. But it was difficult.]

### Part 2

<table>
<thead>
<tr>
<th>Effort for a better story (performance-related)</th>
<th>Strategy to deal with a given task (performance-related)</th>
</tr>
</thead>
<tbody>
<tr>
<td>#21 絵を見るのに必死で、面白いいことが思いつかなかった [I was focusing so much on looking at the pictures that I could not think of anything creative.]</td>
<td>#12 今までのものを含めて、工夫してみたことで、難しいという気持ちが少し減った気がする。例えば ... セリフを入れてみたことで少しやりやすくなった [Reflecting on the tasks that I have done so far, I came up with and used a few strategies. I feel like they helped me perceive this task less difficult. For example, I inserted characters’ lines to the story, and it made the task easier to handle.]</td>
</tr>
</tbody>
</table>
6.3.1.4 Other

Finally, a handful of comments did not belong to any of the three superordinate categories. Among the comments that were coded into this group, the most popular ones in Part 1 (regardless of the tasks) had to do with the dual-task requirement where they had to react to the color change simultaneously while engaging in the narrative tasks. 21% (Task 1) to 32% (Task 4) of the participants commented on the need to perform dual tasks:

#19 (Task 1) 同時に A の文字の色の変化と絵を見なければいけなかった [I have to pay simultaneous attention to the color change of the letter A and the picture prompt.]

#36 (Task 2) だんだん A を見れなくなり、見ていないことに気づいてしゃべれなくなるという負のスパイラルに陥ったから [I got into this vicious cycle where I was unable to check A and got panicked realizing that I hadn’t been able to, and I couldn’t talk, either.]

#40 (Task 3) ストーリーを考えて説明することには慣れてきたが、A の色が変わるのに反応する余裕があまりないことから [I am getting used to coming up with a story and describing it, but I couldn’t afford reacting to the color change.]

#1 (Task 4) 絵が複雑で、プラス色の変化の同時進行が難しかったため [The pictures were complicated, and it was hard to react to the color change simultaneously.]

In Part 2, and also in Part 1 next to the task requirement of reacting to the color detection, 4% (Task 2) to 9% (Tasks 1 and 4) of the participants provided comments that they got used to using English and to the structure and style of the narrative task as they progressed through the experiment. These comments confirm the importance of counterbalancing the order of tasks (which indeed was done in the current study).
6.3.1.5 Summary of the overall patterns

Based on Part 1 and Part 2 participants’ explanatory comments for their ratings, a summary of the overall patterns for each task is provided here. Overall, Task 1 posed low conceptual demands for many of the participants. Its storyline was straightforward and easy to figure out, and the need to simply refer to a single character made the task easy and/or require low mental effort. It was interesting to observe, however, that this simplicity of input added extra difficulty for a few participants in both Part 1 and Part 2 because it required them to create a ‘story’ rather than simply narrating what happened in the picture prompt. Linguistically, although telling of the Task 1 story in English was not an easy task (and it never was for any of the four tasks overall), it seems to have required the simplest and most basic linguistic forms (e.g., structures, words, phrases) among the four tasks, as perceived by the participants.

Unlike Task 1, overall Task 2 posed high conceptual demands. Approximately half of the participants in both parts of the study attributed their high ratings on task difficulty and/or mental effort for Task 2 to its conceptual demands. In particular, many of them thought that its storyline was ambiguous, complicated, and difficult to understand. Successful understanding of the storyline also required attention to details of the pictures; hence some participants expressed that the visual input (i.e., the pictures) of Task 2 was so fine and detailed that it was difficult to decipher what was being depicted. Number of elements, in this task, was used not only as a justification for participants’ low ratings, but more commonly to explain why the task was difficult and/or effort consuming. It is curious to observe here that it does not seem to be the number of characters per se that made Task 2 difficult and effort consuming. In relation to the number of elements, participants expressed that the following particular aspects of the characters raised the level of difficulty and mental effort: (a) similarity between the two characters, and (b)
their related and yet distinct actions and behaviors. Additionally, one respondent commented that the need to refer to a lot of objects (rather than characters) in the story made the task difficult for her. Linguistically, Task 2 required words and expressions related to fishing for its successful completion and the strategy to vary the language to express the repetitive actions of the two characters (i.e., fishing) for a more sophisticated task performance, both of which made the task additionally difficult and mental effort consuming for many of the participants.

Task 3 was more similar to Task 1 than to Task 2 in terms of its conceptual input. Approximately half of the participants in both parts of the study expressed that the storyline of Task 3 was straightforward and easy to figure out. Yet, some specific aspects of the input made the task difficult and/or mental effort consuming. These aspects include: (a) the two similar picture frames (which in turn made it difficult/effort consuming to include both frames in the story in order to fulfill the scoring criterion); (b) difficulty in identifying one of the mother’s actions (i.e., sleeping instead of being distracted by the children); and (c) an object that is hard to describe (i.e., a bowling ball with a creepy face). For this task, the category, ‘number of elements’ was primarily used to explain why the task was difficult and/or required high mental effort. When it was used as a justification for their low difficulty/effort ratings, however, participants commented that the number of elements included in Task 3 is actually small. Although there are seemingly four characters in Task 3, a mother, her son, her daughter, and her new-born baby, the only characters in action for most of the story are the children, and they take actions together for the common purpose of surprising the mother. In addition, in thinking about linguistic consequences of having multiple characters, the task appears to be more difficult when there are similar characters than distinct ones, because the latter allows the speaker to use different terms (e.g., a boy and a girl), rather than noun modifiers to further describe each
character (e.g., a man who is wading in the river and a man on the river bank). These features of Task 3 clearly contrast with Task 2, where the two characters (that looked similar for some participants) had different intentions and were acting individually. These observations beg the question of what the ‘elements’ in the number of elements are supposed to mean, and emphasize the criticalness of considering similarities and differences among characters and how their actions are related. Linguistically, more than half of the participants in both parts of the study attributed their high ratings on difficulty and/or mental effort to the language required to tell the story of Task 3.

Lastly, Task 4 was perceived to be both conceptually and linguistically demanding. Approximately 60% of the participants in Study Part 1 and Part 2 expressed that the conceptual input made Task 4 difficult and mental effort consuming. The most popular answers within the category of conceptual input were the storyline and the number of elements. An important observation can be made about the number of elements here as well. Having to deal with a large number of characters in Task 4 posed not only linguistic demands, in referring to all characters and distinguishing among similar ones, but also conceptual demands, such as figuring out the relationships among them, who should be described, who’s who in each picture frame, and so on. Language-wise, Task 4 required the most unfamiliar language forms for its successful completion, as perceived by the learners.

As reported above, performance factors did influence participants’ perceptions of difficulty and mental effort to some extent, but less often than the other two superordinate categories (at least from the perspectives of the learners). Their influence also seemed to have been more relevant to participants’ understandings of the task instructions and what a ‘good’ performance
might be as well as their motivation and ability to put these instructions into practice, rather than a specific task design feature (i.e., which task it is that they are engaging in).

6.3.2 Explanatory comments by proficiency

In this section, participants’ explanatory comments are reported by proficiency groups. To maximize the possibility of difference in proficiency, the analysis here will focus mostly on the comparison between the low and high proficiency groups. However, the numbers of comments provided by the mid proficiency participants for each of the superordinate categories will be reported and the patterns in their comments are referred to occasionally when they are thought to add value to the proficiency analysis.

6.3.2.1 Task 1

As seen in Tables 40 and 41, in both Study Part 1 and Part 2, the conceptual input became a source of easiness and/or low mental effort proportionally more as L2 proficiency increased (i.e., comparing low with high proficiency groups, or even comparing all three proficiency groups). At the same time, however, in Part 1, we see that a greater number of the high proficiency participants, than in the low proficiency group, attributed their difficulty in narrating the Task 1 story (or high mental effort) to its conceptual input. Comparing the comments provided by participants in each proficiency group, it is apparent that all, but one, of the comments offered by the high proficiency participants had to do with their concern with the logic of the storyline and the need to elaborate on the simple input. Here are some example comments from them:

#20 なぜ再び眠りにつくのか、考えないといけないから [I have to consider why he went back to bed again.]
On the contrary, among the three low proficiency participants who expressed that the conceptual input made Task 1 difficult, only one expressed that the simple input made the task more difficult:

#16 夜の寝てるだけの写真など、どう説明していいか分からずとぼしてしまった [I wasn’t sure what to do with pictures, like the one where a man is simply sleeping, and I ended up skipping those pictures.]

Reasoning about the character’s intentions behind his actions, paying attention to creativity, and actually coming up with a story rather than simply describing the sequence of pictures are all closely related to performance factors. In other words, they could have simply said that the man goes to bed at night and wakes up in the morning, without any elaboration (much like the low proficiency participants did). Instead, the high proficiency participants made an effort to deliver a better performance, which in turn made the task more difficult and mental effort consuming. Hence, we can perhaps safely conclude that the high proficiency participants’ attempts to do a good job on Task 1, especially given such a simple conceptual input, unexpectedly increased their perceived difficulty and mental effort to a small degree. Recall, however, that in terms of

---

8 Yet, they were categorized as the conceptual input factors in this dissertation because participants’ concerns about the logic of the storyline or the need to elaborate on the simple input did not come about by chance, rather they are there largely because of the kind of input (i.e., task) with which they were confronted.
the degree of complexity, the low proficiency participants rated Task 1 to be more difficult and more mental effort consuming than did their high proficiency counterparts. In Part 2, although more high proficiency participants attributed their perceived task difficulty to its conceptual input (a similar trend to Part 1), the trend is less emphatic. It may be the case that the extra attentional resources made available by not having to respond to the secondary color detection task enabled the high proficiency participants to actually put their concerns about the logic of the storyline and the need to elaborate on the simple input into practice.

Another clear pattern that we see in participants’ responses to Task 1 is that in both parts of the study, dramatically lower proportions of the high proficiency participants, compared with the low proficiency participants, expressed that Task 1 posed high code complexity. This finding is not surprising given that high proficiency learners likely possessed a bigger pool of vocabulary, more sophisticated understandings of grammar, better control over their language use, and so on.

Table 40
Summary Table With the Number of Comments for Each Superordinate Category by Proficiency (Task 1, Part 1)

<table>
<thead>
<tr>
<th>Task 1 (Part 1)</th>
<th>Low proficiency</th>
<th>Mid proficiency</th>
<th>High Proficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easy/L Diff/H</td>
<td>Easy/L Diff/H</td>
<td>Easy/L Diff/H</td>
</tr>
<tr>
<td>Conceptual input</td>
<td>25.00% 15.00%</td>
<td>42.86% 28.57%</td>
<td>52.63% 42.11%</td>
</tr>
<tr>
<td>Code complexity</td>
<td>20.00% 70.00%</td>
<td>28.57% 50.00%</td>
<td>21.05% 21.05%</td>
</tr>
<tr>
<td>Performance factors</td>
<td>20.00% 20.00%</td>
<td>7.14% 7.14%</td>
<td>5.26% 5.26%</td>
</tr>
</tbody>
</table>

Note. L = Low mental effort; Diff = Difficult; H = High mental effort; The numbers in parentheses indicate the sheer number of comments provided to justify why a given task was easy/low mental effort consuming or difficult/high mental effort consuming. The total number of participants in Part 1 (i.e., the denominator) is 20 for the low proficiency group, 14 for the mid proficiency group, and 19 for the high proficiency group.
Table 41
*Summary Table With the Number of Comments for Each Superordinate Category by Proficiency (Task 1, Part 2)*

<table>
<thead>
<tr>
<th>Task 1 (Part 2)</th>
<th>Low proficiency</th>
<th>Mid proficiency</th>
<th>High Proficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easy/L</td>
<td>Diff/H</td>
<td>Easy/L</td>
</tr>
<tr>
<td>Conceptual input</td>
<td>50.00%</td>
<td>25.00%</td>
<td>59.46%</td>
</tr>
<tr>
<td></td>
<td>(8)</td>
<td>(4)</td>
<td>(22)</td>
</tr>
<tr>
<td>Code complexity</td>
<td>31.25%</td>
<td>68.75%</td>
<td>24.32%</td>
</tr>
<tr>
<td></td>
<td>(5)</td>
<td>(11)</td>
<td>(9)</td>
</tr>
<tr>
<td>Performance factors</td>
<td>6.25%</td>
<td>12.50%</td>
<td>2.70%</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(1)</td>
</tr>
</tbody>
</table>

*Note.* L = Low mental effort; Diff = Difficult; H = High mental effort; The numbers in parentheses indicate the sheer number of comments provided to justify why a given task was easy/low mental effort consuming or difficult/high mental effort consuming. The total number of participants in Part 2 (i.e., the denominator) is 16 for the low proficiency group, 37 for the mid proficiency group, and 13 for the high proficiency group.

6.3.2.2 Task 2

Participants’ open-ended comments on Task 2 revealed some potentially unexpected results when comparing proficiency groups. Recall that Task 2 was perceived to be second easiest and least effort consuming by the low proficiency learners in both parts of the study, whereas exactly the same task was rated to be second most or most difficult and effort consuming by the high proficiency participants in the two parts. As seen in Tables 42 and 43, twice as many high proficiency participants in Part 1 (i.e., 10 participants out of 19) and 10 times as many participants in Part 2 (i.e., 10 participants out of 13), as compared with their counterparts in the low proficiency group, expressed that the conceptual input made Task 2 *difficult* and required *high* mental effort. As seen in their example comments below, many of these high proficiency participants’ comments indicated that their perceived difficulty and high mental effort were closely associated with what we can interpret to be their concerns for a good task performance (in addition to simply not being able to understand what the input was depicting):
#22 (Part 1) 2人の行動を関連づけて説明しなければなかったから [I had to related the two characters’ actions in describing the pictures.]

#19 (Part 2) 順序を正しく説明しようとしたから [I was trying to get the order of the pictures right.]

#36 (Part 1) 絵はシンプルだったが、それゆえに話が単調になると思い余計に考え る事が多かった [The pictures were simple, but because of that, I thought the story might turn out monotonous and so I had a lot to think about.]

#42 (Part 2) 同様の動作が繰り返され、ストーリーにあまり展開が見られないから [The similar actions are repeated and there is not much change in the storyline.

By contrast, we see that 25% of the low proficiency participants in Part 1 (5 participants in number) confessed that they were focusing more on task completion than on the quality of their task performance:

#28 魚釣りというのが伝わればいいと思った [I thought it would be enough to express that it’s a story about fishing.]

#21 実際は何を言っているのかわからないだろうが、雰囲気は伝わると思う [The listener would probably not understand what I am saying actually, but I think the general idea was conveyed.]

Linguistically, note that there are three (out of 20) and four (out of 16) low proficiency participants, in Study 1 and Study 2 respectively, who expressed that the language required to tell the Task 2 story was easy or less mental effort consuming, whereas there was only a single high proficiency participant who expressed the same idea. Looking more closely into these low proficiency participants’ comments, four expressed that the task only required easy vocabulary, and critically three touched upon the fact that they were using the same expressions repeatedly:
The linguistic consequences of the monotonous nature of the Task 2 input, pointed out by some low proficiency participants, were also mentioned by some of the mid and high proficiency learners in Part 2 (four and two, respectively, to be more specific). However, they used the idea to explain why the task was difficult (or required high mental effort) than easy (or required low mental effort). Here are some of their representative comments:

#32 (mid proficiency) 同じことを説明するのに違った言い方を考えることができなかった（が、それを考えようとして頭を使った）[In describing the same thing, I could not think of different expressions (but I used much mental effort to this end).]

#7 (mid proficiency) 二人ともが釣りをしていたため、同じ単語ばかり使ってしまっていた。（そのため、課題が難しかったと感じた）[Both characters were fishing, so I ended up using the same words (so I thought the task was difficult)]

#50 (high proficiency) 繰り返しの表現をするのが難しかったです [It was difficult to find expressions for repetition.]
Taken together, the same features of the task input seemed to have had distinct effects on the English learners’ perceived difficulty and mental effort for Task 2, depending on their proficiency levels. The repetitive nature of the actions in Task 2 seemed to have made the task easy conceptually, at least for some of the low proficiency learners, because it allowed them to complete the task more or less successfully (in their terms) simply by conveying the gist of the story (i.e., fishing). The same task feature may also have made the task linguistically easy and less mental effort consuming, because they could use the same expressions repeatedly. On the contrary, the same phenomenon may have worked against the mid and high proficiency learners, perhaps because of their attempts to realize their more sophisticated understandings of what a good performance might be, as we have seen earlier.

Table 42
Summary Table With the Number of Comments for Each Superordinate Category by Proficiency (Task 2, Part 1)

<table>
<thead>
<tr>
<th>Task 2 (Part 1)</th>
<th>Low proficiency</th>
<th>Mid proficiency</th>
<th>High Proficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easy/L</td>
<td>Diff/H</td>
<td>Easy/L</td>
</tr>
<tr>
<td>Conceptual input</td>
<td>30.00%</td>
<td>30.00%</td>
<td>21.43%</td>
</tr>
<tr>
<td></td>
<td>(6)</td>
<td>(5)</td>
<td>(3)</td>
</tr>
<tr>
<td>Code complexity</td>
<td>15.00%</td>
<td>65.00%</td>
<td>14.29%</td>
</tr>
<tr>
<td></td>
<td>(3)</td>
<td>(13)</td>
<td>(2)</td>
</tr>
<tr>
<td>Performance factors</td>
<td>25.00%</td>
<td>15.00%</td>
<td>7.14%</td>
</tr>
<tr>
<td></td>
<td>(5)</td>
<td>(3)</td>
<td>(1)</td>
</tr>
</tbody>
</table>

Note. L = Low mental effort; Diff = Difficult; H = High mental effort; The numbers in parentheses indicate the sheer number of comments provided to justify why a given task was easy/low mental effort consuming or difficult/high mental effort consuming. The total number of participants in Part 1 (i.e., the denominator) is 20 for the low proficiency group, 14 for the mid proficiency group, and 19 for the high proficiency group.
Table 43
**Summary Table With the Number of Comments for Each Superordinate Category by Proficiency (Task 2, Part 2)**

<table>
<thead>
<tr>
<th>Task 2 (Part 2)</th>
<th>Low proficiency</th>
<th></th>
<th>Mid proficiency</th>
<th></th>
<th>High Proficiency</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easy/L</td>
<td>Diff/H</td>
<td>Easy/L</td>
<td>Diff/H</td>
<td>Easy/L</td>
<td>Diff/H</td>
</tr>
<tr>
<td>Conceptual input</td>
<td>50.00%</td>
<td>6.25%</td>
<td>18.92%</td>
<td>57.50%</td>
<td>23.08%</td>
<td>76.92%</td>
</tr>
<tr>
<td></td>
<td>(8)</td>
<td>(1)</td>
<td>(7)</td>
<td>(23)</td>
<td>(3)</td>
<td>(10)</td>
</tr>
<tr>
<td>Code complexity</td>
<td>25.00%</td>
<td>81.25%</td>
<td>16.22%</td>
<td>78.38%</td>
<td>7.69%</td>
<td>30.77%</td>
</tr>
<tr>
<td></td>
<td>(4)</td>
<td>(13)</td>
<td>(6)</td>
<td>(29)</td>
<td>(1)</td>
<td>(4)</td>
</tr>
<tr>
<td>Performance factors</td>
<td>0.00%</td>
<td>6.25%</td>
<td>2.70%</td>
<td>5.41%</td>
<td>0.00%</td>
<td>15.38%</td>
</tr>
<tr>
<td></td>
<td>(0)</td>
<td>(1)</td>
<td>(1)</td>
<td>(2)</td>
<td>(0)</td>
<td>(2)</td>
</tr>
</tbody>
</table>

*Note. L = Low mental effort; Diff = Difficult; H = High mental effort; The numbers in parentheses indicate the sheer number of comments provided to justify why a given task was easy/low mental effort consuming or difficult/high mental effort consuming. The total number of participants in Part 2 (i.e., the denominator) is 16 for the low proficiency group, 37 for the mid proficiency group, and 13 for the high proficiency group.*

6.3.2.3 Task 3

In terms of conceptual input, the percentages of the comments explaining why Task 3 was easy (or less mental effort consuming) and difficult (or more mental effort consuming) were relatively balanced in both of the proficiency groups (except for in Part 1 for the high proficiency group). In general, then, the conceptual input of Task 3 seemed to have posed relatively similar degrees of cognitive complexity on English learners regardless of their proficiency levels. Tables 44 and 45 summarize frequencies of comments for each superordinate category provided by participants in each proficiency group.

Comparing the responses in Part 1 with Part 2, it became apparent that there were proportionally more high proficiency participants who perceived the task to be conceptually simple in Part 1. Comments provided by the high proficiency participants in Part 2 suggested their considerable efforts to tell a good story, which in turn may have made the task less easy and more effort consuming. Six participants out of 13 in the high group made nine comments to express their perceived difficulty in: (a) reading characters’ feelings and intentions, (b)
describing events that happened simultaneously (e.g., while the mother was sleeping, the children played a prank on her; while the girl was holding the baby, the boy put a ball in her cradle), (c) describing details of the story, and (d) telling a coherent story that would be easy to understand for a listener. These comments were provided much less often by the low proficiency participants in both parts of the study and by participants in Part 1 overall. Recall that this pattern was what we observed with the overall trends for Part 1 and Part 2 reported above.

Similarly to Task 1, code complexity of Task 3 was less of a problem for the high proficiency learners in both Part 1 and Part 2, and by far the most salient challenge for low proficiency learners.

Table 44
Summary Table With the Number of Comments for Each Superordinate Category by Proficiency (Task 3, Part 1)

<table>
<thead>
<tr>
<th>Task 3 (Part 1)</th>
<th>Low proficiency</th>
<th>Mid proficiency</th>
<th>High Proficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easy/L</td>
<td>Diff/H</td>
<td>Easy/L</td>
</tr>
<tr>
<td>Conceptual input</td>
<td>40.00%</td>
<td>30.00%</td>
<td>35.71%</td>
</tr>
<tr>
<td></td>
<td>(8)</td>
<td>(6)</td>
<td>(5)</td>
</tr>
<tr>
<td>Code complexity</td>
<td>10.00%</td>
<td>70.00%</td>
<td>7.14%</td>
</tr>
<tr>
<td></td>
<td>(2)</td>
<td>(16)</td>
<td>(1)</td>
</tr>
<tr>
<td>Performance factors</td>
<td>5.00%</td>
<td>11.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(0)</td>
</tr>
</tbody>
</table>

Note. L = Low mental effort; Diff = Difficult; H = High mental effort; The numbers in parentheses indicate the sheer number of comments provided to justify why a given task was easy/low mental effort consuming or difficult/high mental effort consuming. The total number of participants in Part 1 (i.e., the denominator) is 20 for the low proficiency group, 14 for the mid proficiency group, and 19 for the high proficiency group.
### Table 45
*Summary Table With the Number of Comments for Each Superordinate Category by Proficiency (Task 3, Part 2)*

<table>
<thead>
<tr>
<th>Task 3 (Part 2)</th>
<th>Low proficiency</th>
<th>Mid proficiency</th>
<th>High Proficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easy/L</td>
<td>Diff/H</td>
<td>Easy/L</td>
</tr>
<tr>
<td>Conceptual input</td>
<td>50.00%</td>
<td>43.75%</td>
<td>45.95%</td>
</tr>
<tr>
<td>Code complexity</td>
<td>18.75%</td>
<td>75.00%</td>
<td>16.22%</td>
</tr>
<tr>
<td>Performance factors</td>
<td>0.00%</td>
<td>6.25%</td>
<td>5.41%</td>
</tr>
</tbody>
</table>

Note. L = Low mental effort; Diff = Difficult; H = High mental effort; The numbers in parentheses indicate the sheer number of comments provided to justify why a given task was easy/low mental effort consuming or difficult/high mental effort consuming. The total number of participants in Part 2 (i.e., the denominator) is 16 for the low proficiency group, 37 for the mid proficiency group, and 13 for the high proficiency group.

### 6.3.2.4 Task 4

As can be seen in Tables 46 and 47, on Task 4, the conceptual input became more and more frequently mentioned as the source of difficulty and high mental effort as the proficiency level increased (as a group). This pattern is similar to that found on Task 2. Inspecting their comments, it became apparent that the high proficiency learners were more affected by having to deal with a large number of elements than were the low proficiency participants, although the proportion of comments attributing conceptual input as a difficulty factor was higher for all groups on this task than on the other three tasks. Table 48 presents the number of comments given specifically on ‘number of elements’ as justifications for participants’ high ratings on difficulty or mental effort for Task 4, by each proficiency group, together with example comments. Proportionally more high proficiency participants felt that the large number of characters involved in the story made Task 4 more difficult and mental effort consuming. This is likely due again to the high proficiency participants’ attempts to perform well. In order to tell a detailed story that is easy to understand for hypothetical listeners, it is imperative for the speaker
to fully understand the storyline, figure out who should be described or included in the story and who should be portrayed as main characters, and clearly distinguish among similar characters and describe how they interact with each other. Many of the high proficiency participants’ comments seem to reflect their struggle to meet these demands.\footnote{\textit{\textsuperscript{9}}} Once again, code complexity became less of an issue as participants’ proficiency increased, though it was attributed as a source of difficulty by a larger proportion of participants in all proficiency groups on this task in comparison with the other three tasks.

\textit{Table 46}  
\textit{Summary Table With the Number of Comments for Each Superordinate Category by Proficiency (Task 4, Part 1)}

<table>
<thead>
<tr>
<th>Task 4 (Part 1)</th>
<th>Low proficiency</th>
<th>Mid proficiency</th>
<th>High Proficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easy/L</td>
<td>Diff/H</td>
<td>Easy/L</td>
</tr>
<tr>
<td>Conceptual input</td>
<td>0.00%</td>
<td>45.00%</td>
<td>21.43%</td>
</tr>
<tr>
<td>Code complexity</td>
<td>0.00%</td>
<td>80.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Performance factors</td>
<td>5.00%</td>
<td>10.00%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

\textit{Note.} L = Low mental effort; Diff = Difficult; H = High mental effort; The numbers in parentheses indicate the sheer number of comments provided to justify why a given task was easy/low mental effort consuming or difficult/high mental effort consuming. The total number of participants in Part 1 (i.e., the denominator) is 20 for the low proficiency group, 14 for the mid proficiency group, and 19 for the high proficiency group.

\footnote{\textit{\textsuperscript{9}}} It might also be the case, however, that the low proficiency participants’ perceived difficulty in the language required to tell the story of Task 4 had overwhelmed them and prevented them from judging the effects of conceptual input on their perceptions. Future studies, then, should ask participants to comment on both aspects of the task at hand, rather than making it open like I did in this exploratory study.
Table 47
**Summary Table With the Number of Comments for Each Superordinate Category by Proficiency (Task 4, Part 2)**

<table>
<thead>
<tr>
<th>Task 4 (Part 2)</th>
<th>Low proficiency</th>
<th>Mid proficiency</th>
<th>High Proficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easy/L</td>
<td>Diff/H</td>
<td>Easy/L</td>
</tr>
<tr>
<td>Conceptual input</td>
<td>12.50%</td>
<td>56.25%</td>
<td>24.32%</td>
</tr>
<tr>
<td></td>
<td>(2)</td>
<td>(9)</td>
<td>(9)</td>
</tr>
<tr>
<td>Code complexity</td>
<td>6.25%</td>
<td>75.00%</td>
<td>13.51%</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(12)</td>
<td>(5)</td>
</tr>
<tr>
<td>Performance factors</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
</tr>
</tbody>
</table>

*Note. L = Low mental effort; Diff = Difficult; H = High mental effort; The numbers in parentheses indicate the sheer number of comments provided to justify why a given task was easy/low mental effort consuming or difficult/high mental effort consuming. The total number of participants in Part 2 (i.e., the denominator) is 16 for the low proficiency group, 37 for the mid proficiency group, and 13 for the high proficiency group.*
Table 48
Number of Comments with Some Example Responses on ‘Number of Elements’ as Justifications for High Ratings on Difficulty or Mental Effort for Task 4 by Proficiency

<table>
<thead>
<tr>
<th>Task 4</th>
<th>Low proficiency</th>
<th>Mid proficiency</th>
<th>High Proficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of comments</td>
<td>16.67% (3)</td>
<td>30.77% (4)</td>
<td>25.00% (4)</td>
</tr>
</tbody>
</table>
| Example responses | #10 登場人物 | #55 一つの絵あたり | #16 男の人が何 | #52 誰を主役 | #25 はしごを登る人、カフェでお茶する人、ドライバー、犬の飼い主の因果関係を説明しようと頭を使いました。
つけ設備数が多く感じ、区別するのが難しいかった [There were many characters, and it was difficult to distinguish among them.] | #42 違が男性のため、違うのがわかるよう
されるのに説明するのが難しかったから |
| | が多く、区別するのが難し | の情報量が多く感じ、区別する | が多く、区別するのが難し | の情報量が多く感じ、区別する | の情報量が多く感じ、区別する |
| | するのが難し | ので、あれもこれも | ので、あれもこれも | ので、あれもこれも | ので、あれもこれも |
| | と説明していたら、 | と説明していたら、 | と説明していたら、 | と説明していたら、 | と説明していたら、 |
| | 自分が何を言ってい | う区別して表現 | う区別して表現 | う区別して表現 | う区別して表現 |
| | るかわからなくなっ | していいか全然 | していいか全然 | していいか全然 | していいか全然 |
| | てしまった [Each | かったから [It | かったから [It | かったから [It | かったから [It |
| | picture frame had a | took some time | took some time | took some time | took some time |
| | lot of information and | to decide who | to decide who | to decide who | to decide who |
| | I was explaining | the main character | the main character | the main character | the main character |
| | everything. Then, I | should be.] | should be.] | should be.] | should be.] |
| | got confused as to what |               |               |               |               |
| | I was saying.] |               |               |               |               |

Note. The numbers in parentheses indicate the sheer number of comments provided to justify why Task 4 difficult/high mental effort consuming. The total number of participants in Part 1 (i.e., the denominator) is 20 for the low proficiency group, 14 for the mid proficiency group, and 19 for the high proficiency group; The total number of participants in Part 2 (i.e., the denominator) is 16 for the low proficiency group, 37 for the mid proficiency group, and 13 for the high proficiency group.
In order to look more into the kind of comments provided by the low and high proficiency participants about code complexity, their comments were further categorized into general and specific code complexity. Responses are categorized into ‘specific’ code complexity when participants listed specific words/expressions/grammar points or particular parts of the story that they had trouble with. For example, a comment “I didn’t know how to say ‘trash’ and a ‘fishing rod’” would be grouped into specific code complexity. By contrast, a more generalized comment about their lack of English proficiency, such as “I am not good at English” would be grouped into general code complexity. When a comment included both general and specific aspects of code complexity, it was coded as specific. The number of comments for each category of code complexity and example responses for the low and high proficiency participants in Part 1 and Part 2 are provided separately in Tables 49 through 52. As can be seen in these tables, the high proficiency learners seemed to have been more aware of the specific issues that they encountered in doing the task at hand than were the low proficiency counterparts, generally speaking. The low proficiency participants often attributed their difficulty in narrating a story to a lack of their general English ability, whereas the participants in the high proficiency group tended to list specific words and phrases that they wanted to use but could not. Note, however, that the low proficiency participants showed their awareness of more particular issues with the language required to tell the stories of Task 2 and Task 3 in Part 2.
Table 49
Number of Comments with Some Example Responses on General versus Specific Task-Induced Code Complexity as Justifications for High Ratings on Difficulty or Mental Effort in Part 1 (Low Proficiency Group)

<table>
<thead>
<tr>
<th></th>
<th>Task 1</th>
<th></th>
<th>Task 2</th>
<th></th>
<th>Task 3</th>
<th></th>
<th>Task 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>General</td>
<td>Specific</td>
<td>General</td>
<td>Specific</td>
<td>General</td>
<td>Specific</td>
<td>General</td>
</tr>
<tr>
<td>Number of comments</td>
<td>84.62%</td>
<td>15.38%</td>
<td>83.33%</td>
<td>16.67%</td>
<td>86.67%</td>
<td>13.33%</td>
<td>87.50%</td>
</tr>
<tr>
<td></td>
<td>(11)</td>
<td>(3)</td>
<td>(10)</td>
<td>(2)</td>
<td>(13)</td>
<td>(2)</td>
<td>(14)</td>
</tr>
</tbody>
</table>

Example responses

- #11 英語が苦手だから [I am not good at English.]
- #5 「まくら」が出てこない [I couldn’t remember how to say ‘pillow.’]
- #14 文章を作る能力がない [I don’t have an ability to create sentences.]
- #3 「ごみ」と「釣竿」の単語がわからなかった [I didn’t know how to say ‘trash’ and a ‘fishing rod.’]
- #28 英語を取らない [English didn’t come out.]
- #4 「〜から〜を取りだす」など熟語を探すのに時間がかかった [It took a while to remember expressions like ‘take something out of somewhere.’]
- #27 思い浮かんだ日本語を英文にできなかった [I could not translate Japanese into English.]
- #10 犬に引っ張られ、うしろにおじさんと犬が歩いているなどを英語で説明するのが難しかった [It was difficult to describe that the old man was being dragged by the dog, the old man and the dog were walking in the background, etc.]

Note. The percentages represent the proportions of participants within each task who gave comments about general or specific task-induced cognitive complexity; hence the percentages within each task sum up to 100%. 
Table 50
Number of Comments with Some Example Responses on General versus Specific Task-Induced Code Complexity as Justifications for High Ratings on Difficulty or Mental Effort in Part 2 (Low Proficiency Group)

<table>
<thead>
<tr>
<th>Task</th>
<th>General</th>
<th>Specific</th>
<th>General</th>
<th>Specific</th>
<th>General</th>
<th>Specific</th>
<th>General</th>
<th>Specific</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>77.78%</td>
<td>22.22%</td>
<td>25.00%</td>
<td>75.00%</td>
<td>30.00%</td>
<td>70.00%</td>
<td>75.00%</td>
<td>25.00%</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Example responses

<table>
<thead>
<tr>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>#36 英語が苦手</td>
</tr>
<tr>
<td>#59 二度寝を何というのか、調べて確認しておきたい[I would like to check how to say ‘to go back to bed again.’]</td>
</tr>
<tr>
<td>#2 単語がなかった</td>
</tr>
<tr>
<td>#66 「つりざお」等</td>
</tr>
<tr>
<td>#58 適確な英語が見つけられません</td>
</tr>
</tbody>
</table>
| #65 時の経過を表す単語がつかれないといった絶対に言わないと伝わらないような単語が思い付かず…[I couldn’t remember words that I absolutely needed to say to get the meaning across, like words to express time passing and to hide...]
| #8 単語を思いすのに必死でした                                    |
| #65 ストーリーを説明する上に重要となる「逮捕される」という単語も他の言いまわしを思い付くことができず、それが思い出すためにかなりの時間を使ってしまった[I couldn’t paraphrase the word that is important in telling the story, ‘to be arrested,’ and I spend so much time on that.]

Note. The percentages represent the proportions of participants within each task who gave comments about general or specific task-induced cognitive complexity; hence the percentages within each task sum up to 100%.
Table 51
Number of Comments with Some Example Responses on General versus Specific Task-Induced Code Complexity as Justifications for High Ratings on Difficulty or Mental Effort in Part 1 (High Proficiency Group)

<table>
<thead>
<tr>
<th></th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Task 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>General</td>
<td>Specific</td>
<td>General</td>
<td>Specific</td>
</tr>
<tr>
<td>Number of comments</td>
<td>33.33%</td>
<td>6.67%</td>
<td>41.67%</td>
<td>58.33%</td>
</tr>
</tbody>
</table>
| Example responses | #51 説明することはやはり難しかった [I was difficulty to describe the pictures in English.]
#34 釣り竿という単語が出ていたか [I was looking for English expressions.]
#20 英語を考えてる ['fishing rod,' the desk fell over etc., but couldn’t think of them in English.]
#33 単語が出てこなかったから [I wanted to use a word ‘fishing rod,’ but couldn’t remember.]
#53 単語が出てこない [Words didn’t come out.]
#54 赤ちゃんの入っているかごの単語が思いつかない [I couldn’t think of a words for the basket in which they baby was.]
#36 単語が浮かばない [I couldn’t come up with words.]
#30 犬を中心話したが、受け身にするなど、英語的に難しかった [I told the story with the dog as the main character, and it was difficult language-wise to use passive forms etc.]

*Note. The percentages represent the proportions of participants within each task who gave comments about general or specific task-induced cognitive complexity; hence the percentages within each task sum up to 100%.*
Table 52

Number of Comments with Some Example Responses on General versus Specific Task-Induced Code Complexity as Justifications for High Ratings on Difficulty or Mental Effort in Part 2 (High Proficiency Group)

<table>
<thead>
<tr>
<th></th>
<th>Task 1</th>
<th></th>
<th>Task 2</th>
<th></th>
<th>Task 3</th>
<th></th>
<th>Task 4</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>General</td>
<td>Specific</td>
<td>General</td>
<td>Specific</td>
<td>General</td>
<td>Specific</td>
<td>General</td>
<td>Specific</td>
</tr>
<tr>
<td>Number of comments</td>
<td>0.00%</td>
<td>100%</td>
<td>50.00%</td>
<td>50.00%</td>
<td>33.33%</td>
<td>66.67%</td>
<td>25.00%</td>
<td>75.00%</td>
</tr>
<tr>
<td>Example responses</td>
<td>N/A</td>
<td>#68 「アラームが鳴る」</td>
<td>#53 英語の単語があまりすぐ思い出せなかった[I couldn't remember English words so quickly.]</td>
<td>#63 魚をつる、逃げられる、何かにあたりが全く分からなかった[I didn't know expressions such as 'to catch a fish,' 'to miss (a fish),' 'to hit something' and so on.]</td>
<td>#22 単語が思い浮かばなかった[I couldn't come up with words.]</td>
<td>#63 「悪だくみをする」「ボー ルを赤ちゃんととりかえる」等のうまい表現が思い浮かばず、苦労した[I couldn't come up with expressions like 'to play a prank,' 'to replace the baby with a ball,' etc., and I had a hard time.]</td>
<td>#42 使用する英単語のレベルが高かったから</td>
<td>#69 犬を結びしているも のを英語でどう表現す るかが難しかった[I was difficult to figure out how to describe the thing that is tied to the dog (i.e., a dog leash) in English.]</td>
</tr>
</tbody>
</table>

*Note.* The percentages represent the proportions of participants within each task who gave comments about general or specific task-induced cognitive complexity; hence the percentages within each task sum up to 100%.
6.3.2.5 Summary of open-ended response patterns by proficiency

Task 1 showed the most straightforward patterns in terms of the effects of task design features on learners’ perceived difficulty and mental effort for distinct proficiency groups. That is, the conceptual input of Task 1 made the task increasingly easier as learners’ L2 proficiency went up; similarly, Task 1 posed higher code complexity to the low proficiency participants than to their high proficiency counterparts. Note, however, that for a handful of high proficiency participants, their attempts to tell a good, elaborate story made the task conceptually difficult, especially given the simple input with a few somewhat ambiguous elements (e.g., why the man is checking outside, why he decided to go back to bed).

Task 2 yielded a very complicated picture of the role of proficiency. Unlike Task 1, Task 2 seemed to have had detrimental effects on the high proficiency participants more than their counterparts in terms of conceptual input and to some extent code complexity. More than half of the high proficiency participants in both study parts felt that Task 2 was conceptually difficult, and many of them were concerned about carefully following the task instructions and elaborating on the rather monotonous storyline, and thereby producing a good task performance. The low proficiency participants, on the contrary, seemed to have used the monotonous, repetitive input to their advantage and used the same words and phrases repeatedly or focused on task completion rather than worrying about the quality of their task performance, as a way of dealing with the given task and its associated cognitive complexity.

Similarly to Task 1, Task 3 showed a relatively straightforward relationship among task design, cognitive complexity, and L2 proficiency. Task 3 was perceived to be simple (or consume low mental effort) and difficult (or consume high mental effort) by similar proportions
of participants with distinct proficiency levels in both Part 1 and Part 2. Linguistically, the higher the proficiency level, the fewer issues they expressed to have had with the language required to tell a story in Task 3.

Lastly, Task 4 yielded a similar pattern to Task 2 in terms of its conceptual input. In other words, proportionally more high proficiency participants than low proficiency participants were affected by having to deal with multiple characters in the Task 4 story and felt that Task 4 was conceptually difficult and mental effort consuming. This rather unexpected trend was once again closely related to the high proficiency participants’ desire to do a good job on the task and their trust in their English abilities (i.e., they thought they should have been able to do what they could not do, which in turn made the task perceptually more difficult and effort consuming).

Linguistically, a lower percentage of high proficiency participants, as compared to their low proficiency counterparts, felt that Task 4 required high code complexity in both parts of the study.

Overall, regardless of the task, high proficiency participants’ more sophisticated understandings of what a good story or narrative performance might be seemed to make all of the tasks additionally difficult and mental effort consuming for these participants, as compared to the low proficiency learners. This finding gives support to the idea prevalent in cognitive psychology that intrinsic cognitive load of a task is largely dependent on learners’ characteristics (e.g., Kalyuga, 2007). That is, the relationship between task design and its putative effects on cognitive complexity varies depending on who the learners are. Moreover, the relationship among task design, cognitive task complexity, and learners’ L2 proficiency, as seen in this study, is quite complicated and far from straightforward. This finding underscores the need to validate our
assumptions about cognitive task complexity for a given task for a particular group of learners that we wish to work with. Furthermore, code complexity was less of a problem for the high proficiency participants than for the low (and mid) proficiency participants in general, and the high proficiency participants were more likely to be aware of (or at least list) specific issues with the language that they had in doing the task, rather than simply attributing difficulty to a general lack of English ability. This latter finding suggests an important difference in degree of awareness of the relationship between language form and task demands that was only, or primarily, available to the high proficiency participants.

6.4 Performance measures

Now that the degree of cognitive task complexity has been measured, we can finally investigate the relationship of measured, instead of presumed, cognitive complexity with task performance. Based on the cognitive load measures, we now know that the four tasks posed relatively similar degrees of cognitive load on the native speaker participants, though Task 1 and Task 3 likely have posed lower degrees of cognitive demands than the other two tasks. For the English-L2 participants, Task 1 was shown consistently to be least cognitively complex and Task 4 to be most complex, with the middle two tasks (Tasks 2 and 3) falling between the other two. When the English-L2 participants’ data were analyzed by proficiency, the order of complexity was shown to be Task 1, followed by Task 3, with Tasks 2 and 4 being the most complex for the high proficiency participants, while the order of complexity for the low proficiency participants was measured on the whole to be Task 1, Task 2, Task 3, and Task 4 sequentially (although the patterns observed on the four measures were less consistent between the various measures than for the native speaker or high proficiency counterparts – see Chapter 7 for a discussion on this
issue). In what follows, the relationship of measured cognitive complexity differences with task performance differences is considered in terms of the syntactic complexity, accuracy, lexical variety, and fluency indices. These findings will be reported firstly for the native speaker participants and then for the English L2 speakers. The English-L2 speakers’ data will also be analyzed for any patterns of difference between the high and low proficiency groups.

### 6.4.1 Native speakers of English

This section reports on the relationship between the measured cognitive load of each task and its effects on syntactic complexity, lexical variety, and fluency for the native speakers of English (recall that accuracy was presumed ‘native like’ for these native speaker participants). Tables 53 and 54 below summarize the means (or medians) and the standard deviations, as well as the effect sizes, for each performance measure in Part 1 and Part 2 separately. Note that the means and standard deviations for the $D$ measure are provided only for 28 participants (out of 30) who produced an adequate number of words (i.e., more than 50 words) for all four tasks in order for the $D$ scores to be calculated in Part 1. In Part 2, all 31 participants’ performances on all four tasks exceeded 50 words.
Table 53
Means/Medians (and Standard Deviations) and Effect Sizes (d) for Each Performance Measure on the Four Tasks in Part 1 (for Native Speakers of English, \(n = 30\))

<table>
<thead>
<tr>
<th></th>
<th>Syntactic Complexity</th>
<th>Lexical Variety</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLT</td>
<td>MLC</td>
<td>C/T</td>
</tr>
<tr>
<td>Task 1</td>
<td>12.50 (2.89)*</td>
<td>6.81 (0.92)*</td>
<td>1.81 (0.37)</td>
</tr>
<tr>
<td>Task 2</td>
<td>12.26 (2.90)</td>
<td>7.83 (1.13)</td>
<td>1.56 (0.30)</td>
</tr>
<tr>
<td>Task 3</td>
<td>14.79 (3.69)</td>
<td>7.72 (1.31)</td>
<td>1.87 (0.31)*</td>
</tr>
<tr>
<td>Task 4</td>
<td>14.49 (4.35)</td>
<td>8.44 (1.25)</td>
<td>1.58 (0.43)*</td>
</tr>
<tr>
<td>(d)</td>
<td>0.54</td>
<td>1.49</td>
<td>−0.57</td>
</tr>
</tbody>
</table>

Note. MLT = Mean length of T-unit; MLC = Mean length of clause; C/T = Clauses per T-unit; EFT = Error-free T-unit; W/S = Words per second; *The median is provided. **\(n = 28\)

Table 54
Means/Medians (and Standard Deviations) and Effect Sizes (d) for Each Performance Measure on the Four Tasks in Part 2 (for Native Speakers of English, \(n = 31\))

<table>
<thead>
<tr>
<th></th>
<th>Syntactic Complexity</th>
<th>Lexical Variety</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLT</td>
<td>MLC</td>
<td>C/T</td>
</tr>
<tr>
<td>Task 1</td>
<td>13.60 (3.66)</td>
<td>7.19 (1.05)</td>
<td>1.89 (0.47)</td>
</tr>
<tr>
<td>Task 2</td>
<td>13.21 (3.70)</td>
<td>7.55 (1.03)*</td>
<td>1.59 (0.45)*</td>
</tr>
<tr>
<td>Task 3</td>
<td>16.61 (3.98)</td>
<td>7.86 (1.18)</td>
<td>2.15 (0.60)</td>
</tr>
<tr>
<td>Task 4</td>
<td>14.97 (3.63)</td>
<td>8.47 (1.27)</td>
<td>1.71 (0.45)*</td>
</tr>
<tr>
<td>(d)</td>
<td>0.38</td>
<td>1.10</td>
<td>−0.39</td>
</tr>
</tbody>
</table>

Note. MLT = Mean length of T-unit; MLC = Mean length of clause; C/T = Clauses per T-unit; EFT = Error-free T-unit; W/S = Words per second; *The median is provided.
6.4.1.1 Syntactic complexity

Syntactic complexity was analyzed by the three separate measures: (a) mean length of T-unit (MLT), (b) mean length of clause (MLC), and (c) clauses per T-unit. Recall that MLT is a global measure of syntactic complexity, MLC indicates syntactic complexity at the phrasal level, and lastly clauses per T-unit is a subordination-based complexity measure. The results will be reported in this order of measures.

6.4.1.1.1 Mean length of T-unit

Inspection of outliers (once again the cut-off set at three standard deviations away from the mean) identified one outlier each in Study Part 1 and Part 2. In Part 1, the distribution for Task 1 was still shown to be substantially positively skewed and leptokurtic, which led me to use its median, instead of the mean, for representing central tendency. In Part 2, by contrast, once the outlier was removed, the distributions were all shown to be normal with no severe skewness or kurtosis. As can be seen in Figures 43 and 44 below as well as Tables 53 and 54 above, Task 1 and Task 2 elicited the shortest MLT (12.26-12.50 words in Part 1 and 13.21-13.60 words in Part 2 per T-unit on average) while Task 3 and Task 4 elicited longer MLT, with Task 3 eliciting quite a bit longer MLT (16.61 words per T-unit) than Task 4 on average (14.97 words) in Part 2. Note that the mean values of Tasks 3 and 4 in both study parts exceed the upper 95% confidence intervals for Tasks 1 and 2, which indicates a meaningful difference between these two sets of tasks on the measure of MLT. Cronbach’s alpha reliability estimate for MLT was found to be \( \alpha = .72 \) in Part 1 and \( \alpha = .56 \) in Part 2, suggesting that it was relatively consistent in distinguishing among participants.\(^{10}\) Cohen’s \( d \) was calculated between Task 1 and Task 4, the two tasks that

\(^{10}\) Note that somewhat low reliability estimates for the native speaker data are perhaps not unexpected, given the general lack of variability in their performances.
were measured to be maximally different in terms of their cognitive load, and it was found to be $d = .54$ in Part 1 and $d = .38$ in Part 2, indicating small to medium effect sizes (Cohen, 1988).

Also note that the difference was approximately two words between Task 1 and Task 4, for MLT, which is quite a large difference obtained simply by doing different tasks in the same basic genre of oral narration.

Figure 43. MLT for the four tasks with 95% confidence intervals in Part 1 (for native speakers of English).
6.4.1.1.2 Mean length of clauses

On this particular measure, one outlier was detected in Part 1, while no outliers were found in Part 2. Even after the outlier was removed, the distribution in Part 1 was shown to be substantially positively skewed and leptokurtic for Task 1. In Part 2, Task 2 was found to be substantially positively skewed. Otherwise, the distributions were all found to be normal in both study parts. Based on these observations, the decision was made to use the median when examining central tendency for the data whose distribution was shown to be substantially skewed. Figures 45 and 46 below show the graphic representation of the overall patterns. In terms of MLC, Task 1 elicited the shortest clauses (6.81-7.19 words per clause), while Task 4 elicited the longest clauses (8.44-8.47 words) in both study parts. The mean values of Task 4 clearly exceed the upper 95% confidence intervals for Task 1, which indicates a meaningful
difference between these two tasks on the measure of MLC. Tasks 2 and 3 did fall between the other two tasks in both Part 1 and Part 2, but in Part 1 the two tasks elicited similar MLC, whereas in Part 2 Task 3 elicited longer clauses than did Task 2. Cronbach’s alpha reliability estimate for MLC was calculated to be $\alpha = .71$ in Part 1 and $\alpha = .58$ in Part 2, suggesting that it was relatively consistent in distinguishing among participants. Cohen’s $d$ between Task 1 and Task 4 was found to be $d = 1.49$ in Part 1 and $d = 1.10$ in Part 2, indicating large effect sizes. Note that the difference between Task 1 and Task 4 on the measure of MLC was approximately 1.5 words.

![Figure 45. MLC for the four tasks with 95% confidence intervals in Part 1 (for native speakers of English).](image-url)
Figure 46. MLC for the four tasks with 95% confidence intervals in Part 2 (for native speakers of English).

6.4.1.1.3 Clauses per T-unit

The same two individuals who were found to be outliers on the MLT measure in each study part were again found to be outliers on the clauses per T-unit measure; hence their data on this particular measure was also eliminated from analysis. After the removal of the outliers, the data for Task 3 and Task 4 in Part 1 and for Task 2 and Task 4 in Part 2 were still shown to be substantially positively skewed, and Task 3 in Part 1 and Task 4 in Part 2 were also substantially leptokurtic. Based on this observation, the medians are used when reporting central tendencies for Tasks 3 and 4 in Part 1 and Tasks 2 and 4 in Part 2. As seen in Figures 47 and 48 (as well as Tables 53 & 54), Task 2 and Task 4 elicited the fewest clauses per T-unit (1.56-1.71 clauses per T-unit), while Task 1 and Task 3 elicited more clauses per T-unit (1.81-2.15 clauses) than the other two tasks in both study parts. Observe that the mean values of Tasks 1 and 3 exceed the
upper 95% confidence intervals for Task 2 in Part 1 and for both Task 2 and Task 4 in Part 2, which indicates a meaningful difference between these two sets of tasks on the measure of clauses per T-unit. (The upper 95% confidence intervals for Task 4 was slightly above the mean values of Tasks 1 and 3 in Part 1, largely due to a bigger standard deviation (and hence larger confidence intervals) of Task 4.) Note also that in Part 2 Task 3 elicited substantially greater clauses per T-unit than did Task 1 (0.26 clauses different). Cronbach’s alpha reliability estimate for the measure of clauses per T-unit was found to be $\alpha = .63$ in Part 1 and $\alpha = .59$ in Part 2, suggesting that it was relatively consistent in distinguishing among participants. Finally, Cohen’s $d$ between Task 1 and Task 4 was found to be $d = -.57$ in Part 1 and $d = -.39$ in Part 2, indicating small to medium effect sizes in favor of the simple task.
Figure 47. C/T for the four tasks with 95% confidence intervals in Part 1 (for native speakers of English).
Figure 48. C/T for the four tasks with 95% confidence intervals in Part 2 (for native speakers of English).

6.4.1.2 Lexical variety

Lexical variety was assessed by two measures: (a) $D$ and (b) TTR. Inspection for outliers revealed no outliers. The data for both measures were shown to be normally distributed with no severe skewness or kurtosis, except for Task 3 on TTR in Study Part 1 – it was substantially leptokurtic. Looking at the $D$ scores for the four tasks, it can be observed that all four tasks elicited approximately the same $D$ scores in both study parts, although in Part 2, Task 1 elicited lower $D$ score (44.49) than did the other three tasks, especially Task 2 (49.54) (see Figures 49 & 50). Observe that the confidence intervals are all overlapping with each other (except that the mean value of Task 2 in Part 2 is as high as the upper 95% confidence intervals for Task 1.) Cronbach’s alpha reliability estimate for this lexical variety measure $D$ was shown to be $\alpha = .91$
in Part 1 and $\alpha = .79$ in Part 2, suggesting that it was highly consistent in distinguishing among participants. Cohen’s $d$ for $D$ (between Task 1 and Task 4) was found to be $d = .01$ in Part 1 and $d = .26$ in Part 2, both indicating small effect sizes. TTR also revealed a similar trend in that the native speaker participants producing approximately the same TTR on average on all the four tasks (see Figures 51 & 52). Cronbach’s alpha reliability estimate for TTR was $\alpha = .84$ in Part 1 and $\alpha = .88$ in Part 2, suggesting that it was highly consistent in distinguishing among participants. Cohen’s $d$ between Task 1 and Task 4 for TTR was found to be $d = -.25$ in Part 1 and $d = -.33$ in Part 2, which indicate small negative effect sizes. Cognitive task complexity, therefore, did not seem to be much related with differences in lexical performance attributes of the native speakers.

![Figure 49. D score for the four tasks with 95% confidence intervals in Part 1 (for native speakers of English).](image)

261
Figure 50. D score for the four tasks with 95% confidence intervals in Part 2 (for native speakers of English).
Figure 51. TTR for the four tasks with 95% confidence intervals in Part 1 (for native speakers of English).
Figure 52. TTR for the four tasks with 95% confidence intervals in Part 2 (for native speakers of English).

6.4.1.3 Fluency

Fluency was measured by (a) the total number of words produced and (b) the number of words per second. The data for each task were inspected for outliers and for normality. For the measure of the total number of words, two outliers were identified in Part 2. After the removal of these outliers, the distributions for Task 2 in Part 1 (also leptokurtic) as well as for Tasks 3 and 4 in Part 2 were found to be positively skewed. Hence, the medians were used as the more accurate representation of their central tendencies for these tasks. As for the measure of words per seconds, no outlier was found and all data were shown to be normally distributed with no severe skewness or kurtosis. As can be seen in Figures 53-54 and Tables 53-54, the native speaker participants produced the longest stories on average when engaging in Task 4. Note, however,
that in Part 2, Task 2 elicited as many words as did Task 4. In fact, the mean values of Task 4 and Task 2 (only in Part 2) were found to be higher than the upper 95% confidence intervals for Tasks 1 and 3, indicating that these two sets of tasks elicited meaningfully different length of speech. Cronbach’s alpha reliability estimate for the measure of the total number of words was found to be $\alpha = .88$ in Part 1 and $\alpha = .94$ in Part 2, suggesting that it was highly consistent in distinguishing among participants. Finally, Cohen’s $d$ between Task 1 and Task 4 for the total number of words was found to be $d = .48$ in Part 1 and $d = .61$ in Part 2, both indicating medium effect sizes.

*Figure 53.* Total number of words for the four tasks with 95% confidence intervals in Part 1 (for native speakers of English).
Turning to the words per second measure, native speakers’ performances were more fluent for Task 1 than for the other tasks on average (see Figures 55 and 56). Note that the mean values of Task 1 exceeded the upper 95% confidence intervals for Tasks 2 and 4 in both Part 1 and Part 2. Cronbach’s alpha reliability estimate for the fluency measure of words per second was $\alpha = .87$ in Part 1 and $\alpha = .90$ in Part 2, suggesting that it was highly consistent in distinguishing among participants. Cohen’s $d$ for the measure of words per second was found to be $d = -.45$ in Part 1 and $d = -.48$ in Part 2, both indicating medium effect sizes in favor of the simple task.

![Figure 54. Total number of words for the four tasks with 95% confidence intervals in Part 2 (for native speakers of English).](image)
Figure 55. W/S for the four tasks with 95% confidence intervals in Part 1 (for native speakers of English).
Figure 56. W/S for the four tasks with 95% confidence intervals in Part 2 (for native speakers of English).

6.4.1.4 Summary of performance measures for native speakers

Overall, Part 1 with a color-detection task and Part 2 without it elicited very similar patterns in the performance measures for the native speaker participants. Interestingly, the four tasks that posed similar degrees of cognitive task complexity overall (with Tasks 1 and 3 and Tasks 2 and 4 posing somewhat distinct complexity) elicited performances that were different in quality. Task 3 (on all three measures) and Task 4 (on MLT and MLC) elicited syntactically more complex performances than did the other two tasks; Task 4 elicited the longest speech, whereas Task 1 elicited the most fluent speech of all the four tasks; and lastly, the tasks elicited similar lexical variety scores across tasks. Comparing Task 1 and Task 4 (the two tasks that were measured to be maximally different in cognitive terms), cognitive task complexity had small to
medium positive effects on almost all of the performance measures, except that it had large positive effects on MLC and small to medium negative effects on clauses per T-unit and on fluency in words per second.

6.4.2 English L2 speakers

Moving on to English L2 speakers’ task performances, this section reports on the relationship between the measured cognitive complexity of each task and its effects on syntactic complexity, accuracy, lexical variety, and fluency of participants’ task performances. As was the case with the cognitive load measures, the overall patterns will be reported first for each task, and then the results will be presented for each of the proficiency groups (again for each task), for Part 1 and Part 2 separately.

6.4.2.1 Overall trends

All the data were firstly inspected for outliers and checked for distributional characteristics. After the removal of outliers (i.e., with scores more than 3 standard deviations away from the mean), if the distribution was still shown to be skewed, as a central tendency indicator, the median, instead of the mean, was used. Tables 55 and 56 below summarize means (or medians) and standard deviations for each performance measure overall on the four tasks in Part 1 and Part 2. Note that the means and standard deviations for the $D$ measure are provided only for 25 participants in Part 1 and 37 participants in Part 2 who produced enough number of words (i.e., more than 50 words) for all four tasks for the $D$ scores to be calculated.
Table 55
Means/Medians (and Standard Deviations) and Effect Sizes (d) for Each Performance Measure on the Four Tasks in Part 1 (for English-L2 Speakers, n = 53)

<table>
<thead>
<tr>
<th></th>
<th>Syntactic Complexity</th>
<th>Accuracy</th>
<th>Lexical Variety</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLT</td>
<td>MLC</td>
<td>C/T*</td>
<td>EFT (%)</td>
</tr>
<tr>
<td>Task 1</td>
<td>7.43</td>
<td>5.78</td>
<td>1.24</td>
<td>41.53</td>
</tr>
<tr>
<td></td>
<td>(2.02)</td>
<td>(0.98)</td>
<td>(0.21)</td>
<td>(28.31)</td>
</tr>
<tr>
<td>Task 2</td>
<td>7.52</td>
<td>6.59</td>
<td>1.10</td>
<td>38.40</td>
</tr>
<tr>
<td></td>
<td>(1.73)</td>
<td>(1.06)</td>
<td>(0.16)</td>
<td>(28.06)</td>
</tr>
<tr>
<td>Task 3</td>
<td>9.29</td>
<td>6.73</td>
<td>1.31</td>
<td>42.42</td>
</tr>
<tr>
<td></td>
<td>(2.75)</td>
<td>(1.23)</td>
<td>(0.31)</td>
<td>(27.52)</td>
</tr>
<tr>
<td>Task 4</td>
<td>8.99</td>
<td>6.93</td>
<td>1.23</td>
<td>37.83</td>
</tr>
<tr>
<td></td>
<td>(2.24)</td>
<td>(1.29)*</td>
<td>(0.24)</td>
<td>(28.32)</td>
</tr>
<tr>
<td>d</td>
<td>0.73</td>
<td>1.00</td>
<td>−0.04</td>
<td>−0.13</td>
</tr>
</tbody>
</table>

*Note. MLT = Mean length of T-unit; MLC = Mean length of clause; C/T = Clauses per T-unit; EFT = Error-free T-unit; W/S = Words per second; *The medians is provided. **n = 25
Table 56
Means/Medians (and Standard Deviations) and Effect Sizes (d) for Each Performance Measure on the Four Tasks in Part 2 (for English-L2 Speakers, n = 67)

<table>
<thead>
<tr>
<th>Task</th>
<th>Syntactic Complexity</th>
<th>Accuracy</th>
<th>Lexical Variety</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLT</td>
<td>MLC</td>
<td>C/T</td>
<td>EFT (%)</td>
</tr>
<tr>
<td>Task 1</td>
<td>7.30</td>
<td>5.98</td>
<td>1.25</td>
<td>35.97</td>
</tr>
<tr>
<td></td>
<td>(2.26)*</td>
<td>(1.05)</td>
<td>(0.28)*</td>
<td>(25.20)</td>
</tr>
<tr>
<td>Task 2</td>
<td>7.67</td>
<td>6.62</td>
<td>1.17</td>
<td>34.46</td>
</tr>
<tr>
<td></td>
<td>(1.88)*</td>
<td>(1.09)</td>
<td>(0.22)*</td>
<td>(22.98)</td>
</tr>
<tr>
<td>Task 3</td>
<td>9.74</td>
<td>6.75</td>
<td>1.39</td>
<td>41.61</td>
</tr>
<tr>
<td></td>
<td>(2.80)</td>
<td>(1.18)</td>
<td>(0.35)*</td>
<td>(24.37)</td>
</tr>
<tr>
<td>Task 4</td>
<td>8.97</td>
<td>6.96</td>
<td>1.30</td>
<td>37.00</td>
</tr>
<tr>
<td></td>
<td>(1.89)</td>
<td>(1.19)</td>
<td>(0.21)</td>
<td>(22.71)</td>
</tr>
<tr>
<td>d</td>
<td>0.80</td>
<td>0.87</td>
<td>0.20</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Note. MLT = Mean length of T-unit; MLC = Mean length of clause; C/T = Clauses per T-unit; EFT = Error-free T-unit; W/S = Words per second; *The median is provided. **n = 37
6.4.2.1.1 Syntactic complexity

Syntactic complexity was analyzed by the three separate measures: (a) mean length of T-unit (MLT), (b) mean length of clause (MLC), and (c) clauses per T-unit. The results will be reported in this order of measures for each study part separately.

6.4.2.1.1 Mean length of T-unit

Inspections for outliers identified none in Part 1 and one outlier in Part 2. In Part 1, the data were all normally distributed, except that Task 1 was shown to be substantially leptokurtic. In Part 2, even after the removal of the outlier, the distributions for Task 1 and Task 2 were positively skewed and Task 2 was also leptokurtic. As can be seen in Figures 57 and 58 below, Task 3, the measured to be second simplest task, elicited the longest T-units on average (9.29 words per T-unit in Part 1, 9.74 words per T-unit in Part 2). In Part 1, Task 4 elicited a high average MLT score, similar in degree to Task 3, while the other two tasks elicited substantially shorter average MLT scores than Tasks 3 and 4. Notice that the mean values of Tasks 3 and 4 exceed the upper 95% confidence intervals for Tasks 1 and 2, which means that these two sets of tasks elicited meaningfully different average MLT scores. In Part 2, the pattern is very similar to Part 1, except that this time Task 4 elicited essentially identical MLT while Task 3 elicited still longer MLT values. We can observe in Figure 58 that in Part 2, the mean value of Task 3 exceeds the upper 95% confidence intervals for all the other tasks, indicating that the difference between these two sets of tasks were found to be meaningful, and that Task 4 elicited meaningfully longer T-units on average than did Tasks 1 and 2. Cronbach’s alpha reliability estimate for MLT was shown to be $\alpha = .85$ in Part 1 and $\alpha = .89$ in Part 2, suggesting that it was
highly consistent in distinguishing among participants. Cohen’s $d$ between Task 1 and Task 4 was found to be $d = .73$ in Part 1 and $d = .80$ in Part 2, indicating medium to large effect sizes.

Figure 57. MLT for the four tasks with 95% confidence intervals in Part 1 (for English L2 speakers).
Figure 58. MLT for the four tasks with 95% confidence intervals in Part 2 (for English L2 speakers).

6.4.2.1.1.2 Mean length of clause

There were no outliers found on the measure of MLC in either of the study parts. The data then were all shown to be normally distributed, except for Task 4 in Part 1 being positively skewed and leptokurtic. The data for Task 4 in Part 2 was also leptokurtic, but not skewed. For this reason, the decision was made to use the median, instead of the mean, in investigating central tendency for Task 4 solely in Part 1. Looking at Tables 55 and 56 above, it can be observed that MLC gradually increased consistently from Task 1 to Task 4. In particular, Task 1, the measured to be least complex task, elicited substantially shorter clauses (5.78 words per clause in Part 1 and 5.98 in Part 2 on average) than the other three tasks in both Study Part 1 and Part 2 (ranging from 6.59 to 6.96). The mean values of Tasks 2 through 4 were in fact found to
be higher than the upper 95% confidence intervals for Task 1 (see Figures 59 & 60 below).

Cronbach’s alpha reliability estimate for MLC was found to be $\alpha = .83$ in Part 1 and $\alpha = .76$ in Part 2, suggesting that it was highly consistent in distinguishing among participants. Cohen’s $d$ between Task 1 and Task 4 was found to be $d = 1.00$ in Part 1 and $d = .87$ in Part 2, indicating large effect sizes.

![Error Bars: 95% CI](image)

*Figure 59.* MLC for the four tasks with 95% confidence intervals in Part 1 (for English L2 speakers).
Figure 60. MLC for the four tasks with 95% confidence intervals in Part 2 (for English L2 speakers).

6.4.2.1.3 Clauses per T-unit

Inspections for outliers identified one outlier in Part 1 and none in Part 2. After removing the outlier, the data for this measure of clauses per T-unit were all positively skewed, with Task 3 and Task 4 also being quite leptokurtic in Part 1. Similarly, in Part 2, the data for all tasks, but Task 4, were positively skewed, and Task 2 was also shown to be leptokurtic. Comparing the four tasks, the salient pattern is that in both study parts, Task 2 elicited substantially fewer clauses per T-unit (1.10 clauses per T-unit in Part 1 and 1.17 in Part 2 on average), while the other tasks elicited essentially around the same, higher number of clauses per T-unit (see Tables 55 & 56). The mean values of Tasks 1, 3, and 4 were in fact found to be higher than the upper 95% confidence intervals for Task 2 in both Study Part 1 and Part 2, indicating a meaningful
difference between these two sets of tasks (see Figures 61 & 62). Note also that in Part 2, Task 3 elicited substantially higher clauses per T-unit values than did Tasks 1 and 4, in addition to Task 2. Cronbach’s alpha reliability estimate for the measure of clauses per T-unit was found to be $\alpha = .65$ in Part 1 and $\alpha = .77$ in Part 2, suggesting that it was generally consistent in distinguishing among participants. Cohen’s $d$ between Task 1 and Task 4 was found to be $d = -.04$ in Part 1 and $d = .20$ in Part 2, both indicating small effect sizes with the $d$ value for Part 1 going minimally in the negative direction (in favor of the simple task).

![Figure 61. C/T for the four tasks with 95% confidence intervals in Part 1 (for English L2 speakers).]
Figure 6. C/T for the four tasks with 95% confidence intervals in Part 2 (for English L2 speakers).

6.4.2.1.2 Accuracy

For the accuracy measure of error-free T-units, no outliers were found in either of the study parts and the distributions were found to be normal with no severe skewness or kurtosis for all four tasks. Figures 63 and 64 graphically represent the overall pattern for each of the four tasks. In these figures, we can observe that the four tasks elicited approximately the same rates of accuracy, ranging from 37.83% (Task 4) to 42.42% (Task 3) in Part 1. In Part 2, by contrast, Task 3 elicited somewhat greater degrees of accurate production (41.61%) than did the other three tasks (34.46%-37.00%). However, differences in accuracy rates were minimal between the four tasks, in general. Cronbach’s alpha reliability estimate for the accuracy measure was shown to be $\alpha = .84$ in Part 1 and $\alpha = .80$ in Part 2, suggesting that it was highly consistent in
distinguishing among participants. Finally, Cohen’s $d$ between Task 1 and Task 4 was found to be $d = -.13$ in Part 1 and $d = .04$ in Part 2, both indicating very small effect sizes.

*Figure 63.* Accuracy for the four tasks with 95% confidence intervals in Part 1 (for English L2 speakers).
Figure 64. Accuracy for the four tasks with 95% confidence intervals in Part 2 (for English L2 speakers).

6.4.2.1.3 Lexical variety

No outliers were found for the $D$ measure in either Part 1 or Part 1. The data distributions for this measure were shown to be normal except that Task 1 in Part 2 was positively skewed and leptokurtic. In both study parts, Task 3 elicited the biggest $D$ score among the four tasks (5.99 higher than Task 2, which elicited the lowest $D$ score), and substantially higher in Study Part 2 (7.79 higher than Task 2). As can be seen in Figures 65 and 66, the mean values of Task 3 were higher than the upper 95% confidence intervals for all the other tasks (except for Task 4 in Part 1), which means that Task 3 elicited meaningfully higher $D$ scores than did any of the other tasks. Cronbach’s alpha reliability estimate for the lexical measure of $D$ was found to be $\alpha = .61$ in Part 1 and $\alpha = .75$ in Part 2, suggesting that it was generally consistent in distinguishing among
participants. Cohen’s $d$ between Task 1 and Task 4 for $D$ was $d = .12$ in Part 1 and $d = .02$ in Part 2, both indicating quite small effect sizes.

Figure 65. $D$ score for the four tasks with 95% confidence intervals in Part 1 (for English L2 speakers).
Turning our attention to the other lexical variety measure, TTR, two outliers were found in Part 1 and none was found in Part 2. After eliminating these outliers from the analysis, the data distributions were shown to be normal except that Task 3 in Part 1 was substantially leptokurtic and Task 4 in Part 2 was positively skewed and leptokurtic. Unlike the $D$ score, the results of TTR indicated that Task 1, not Task 3, elicited the highest lexical variety among the four tasks in both Part 1 and Part 2 (0.60 in both study parts, see Tables 55-56 above and Figures 67-68). Caution must be exercised, however, in interpreting these results, because TTR is shown to be largely influenced by the number of words produced. As we will see in the next section, Task 1 elicited the shortest speech of all the four tasks. Hence, it may be the case that the largest TTR found for Task 1 was due to it having the smallest number of words produced. Cronbach’s alpha
reliability estimate for TTR was $\alpha = .73$ in Part 1 and $\alpha = .76$ in Part 2, suggesting that it was generally consistent in distinguishing among participants. Cohen’s $d$ between Task 1 and Task 4 for TTR was found to be $d = -.67$ in Part 1 and $d = -1.12$ in Part 2, indicating medium to large effect sizes in favor of the simple task, likely due to a confound with its text length.

*Figure 67.* TTR for the four tasks with 95% confidence intervals in Part 1 (for English L2 speakers).
Figure 68. TTR for the four tasks with 95% confidence intervals in Part 1 (for English L2 speakers).

6.4.2.1.4 Fluency

Two outliers were found in Part 1 and one in Part 2 for the measure of the total number of words. After removal of these outliers, the data for Tasks 2 and 3 were found to be positively skewed with Task 3 also being leptokurtic in Part 1; and the data for all four tasks were found to be positively skewed in Part 2. The results revealed that the English L2 participants produced increasingly greater number of words from Task 1 to Task 4 in both study parts (see Figures 69 & 70). Note that Task 4, the measured-to-be most complex task, elicited a meaningfully greater number of words (81.20 words in Part 1 and 80.00 words in Part 2 on average) than did Task 1, the measured-to-be simplest task, (53.63 words in Part 1 and 57.00 words in Part 2) in both Study Part 1 and Part 2 (judging based on the mean values of and the confidence intervals for
these two tasks). Cronbach’s alpha reliability estimate for the total number of words was found to be $\alpha = .89$ in Part 1 and $\alpha = .93$ in Part 2, suggesting that it was highly consistent in distinguishing among participants. Finally, Cohen’s $d$ between Task 1 and Task 4 for this measure was found to be $d = .94$ in Part 1 and $d = .68$ in Part 2, indicating medium to large effect sizes.

Figure 69. Total number of words for the four tasks with 95% confidence intervals in Part 1 (for English L2 speakers).
Moving on to the measure of words per second, inspections for outliers on this measure found no severe outliers in either of the two study parts. In terms of data distribution, all tasks, but Task 4 in Part 2, were positively skewed. In Part 2, Task 1 was also shown to be substantially leptokurtic. As seen in Tables 55 and 56 above, the average speech rates were similar across the four tasks, with Task 1, the measured-to-be simplest task, eliciting slightly more fluent speech than the other three tasks in both Part 1 and Part 2. 0.12 more words were produced on average for Task 1 than for Task 2, the task that elicited the least number of words per second (in both parts). Note the overlapping confidence intervals for the four tasks in both study parts (see Figures 71 and 72). Cronbach’s alpha reliability estimate for the fluency measure of words per second was found to be $\alpha = .98$ in Part 1 and $\alpha = .97$ in Part 2, suggesting that it was highly
consistent in distinguishing among participants. Cohen’s $d$ between Task 1 and Task 4 for this measure was found to be $d = -.18$ in Part 1 and $d = -.06$ in Part 2, both indicating negative, small effect sizes.

*Figure 71. W/S for the four tasks with 95% confidence intervals in Part 1 (for English L2 speakers).*
Figure 72. W/S for the four tasks with 95% confidence intervals in Part 2 (for English L2 speakers).

6.4.2.1.5 Summary of performance measures for English L2 speakers

Not surprisingly, comparing the task performances of native speakers’ and English L2 speakers’ in terms of the syntactic complexity, lexical variety, and fluency indices, the former produced better performances than did the latter according to all measures, except for TTR (which is likely too unstable for meaningful interpretation at the differing numbers of words produced per task). Focusing on the English L2 speakers then, clearly it was Task 3, the measured-to-be second simplest task, that elicited the best performance of all across all measures. Task 3 elicited the syntactically most complex, most accurate (in Part 2, although not to a meaningful extent), and lexically most diverse (on the $D$ measure) language production. In comparing the measured-to-be least and most complex tasks (Task 1 and Task 4), overall, the
most salient effect seems to have been in the area of syntactic complexity, with relatively large average differences between performances on the two tasks. Note also that Task 4 elicited by far the longest average performances (in number of words) and Task 1 the shortest. Comparing Task 1 and Task 4, cognitive complexity had medium to large positive effects on the performance measures of MLT, MLC (syntactic complexity), and total number of words. In other words, the English L2 participants benefited from increased cognitive complexity on these three dimensions of task performance, to quite a large extent. On the measures of clauses per T-unit (syntactic complexity), error-free T-units (accuracy), $D$ (lexical variety), and words per second (speech rate) cognitive task complexity had quite small, and sometimes negative, effects. Lastly, task complexity was found to have had medium to large, negative effects on TTR, which likely reflects the influence of text length on this unstable measure.

6.4.2.2 Performance measures by proficiency

This section reports on the results of the performance measures by participants’ English proficiency in Study Part 1 and Study Part 2. Once again, all low, mid, and high proficiency groups’ results will be presented; however, the comparison focuses on the low and high proficiency groups to allow maximal difference in their proficiency levels. Given (relatively) small sets of data for each proficiency group, outliers will not be eliminated from the analysis this time, but if the data are found to be skewed, the median is used, instead of the mean, as an indicator of central tendency.
6.4.2.2.1 Syntactic complexity

Tables 57 and 58 below summarize the relationship between the four tasks with differing degrees of cognitive complexity and the syntactic complexity of participants’ language production for each proficiency group. Recall that syntactic complexity was assessed by three measures: (a) mean length of T-unit, (b) mean length of clauses, and (c) clauses per T-unit. As can be seen in Tables 57 and 58, generally speaking, regardless of the tasks, the high proficiency participants produced dramatically more syntactically complex language than did their low proficiency counterparts on all measures of syntactic complexity. The results of each measure are reported in depth in the following sections.
Table 57
Means/Medians (and Standard Deviations) and Effect Sizes (d) for Syntactic Complexity Measures on the Four Tasks by Proficiency in Part 1

<table>
<thead>
<tr>
<th></th>
<th>Syntactic Complexity</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLT</td>
<td>MLT</td>
<td>MLC</td>
<td>C/T</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L (20)</td>
<td>M (14)</td>
<td>H (19)</td>
<td>L (20)</td>
<td>M (14)</td>
</tr>
<tr>
<td>Task 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5.94</td>
<td>7.91</td>
<td>8.57</td>
<td>5.19</td>
<td>5.95</td>
</tr>
<tr>
<td></td>
<td>(1.03)</td>
<td>(1.78)</td>
<td>(2.04)*</td>
<td>(0.71)</td>
<td>(1.16)</td>
</tr>
<tr>
<td>Task 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6.06</td>
<td>7.95</td>
<td>8.73</td>
<td>5.84</td>
<td>7.11</td>
</tr>
<tr>
<td></td>
<td>(1.03)</td>
<td>(1.02)</td>
<td>(1.63)</td>
<td>(0.94)</td>
<td>(0.98)</td>
</tr>
<tr>
<td>Task 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6.97</td>
<td>9.73</td>
<td>11.41</td>
<td>5.70</td>
<td>7.01</td>
</tr>
<tr>
<td></td>
<td>(1.25)</td>
<td>(1.86)</td>
<td>(2.62)</td>
<td>(0.97)*</td>
<td>(1.10)</td>
</tr>
<tr>
<td>Task 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7.31</td>
<td>9.46</td>
<td>10.40</td>
<td>5.86</td>
<td>7.91</td>
</tr>
<tr>
<td></td>
<td>(1.42)</td>
<td>(1.94)</td>
<td>(2.04)</td>
<td>(1.01)*</td>
<td>(1.31)*</td>
</tr>
<tr>
<td>d</td>
<td>1.10</td>
<td>0.83</td>
<td>0.90</td>
<td>0.77</td>
<td>1.58</td>
</tr>
</tbody>
</table>

Note. MLT = Mean length of T-unit; MLC = Mean length of clause; C/T = Clauses per T-unit; L = Low proficiency group (n-size); M = Mid proficiency group (n-size); H = High proficiency group (n-size); *The median is provided.
Table 58
Means/Medians (and Standard Deviations) and Effect Sizes \( (d) \) for Syntactic Complexity Measures on the Four Tasks by Proficiency in Part 2

<table>
<thead>
<tr>
<th>Task</th>
<th>Syntactic Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLT</td>
</tr>
<tr>
<td>Task 1</td>
<td>6.61 (1.70)</td>
</tr>
<tr>
<td>Task 2</td>
<td>6.94 (1.30)</td>
</tr>
<tr>
<td>Task 3</td>
<td>7.79 (2.23)</td>
</tr>
<tr>
<td>Task 4</td>
<td>7.57 (2.13)</td>
</tr>
</tbody>
</table>

| d    | 0.50 | 1.08 | 0.16 | 0.51 | 1.21 | 0.76 | 0.12 | 0.37 | −0.52 |

Note. MLT = Mean length of T-unit; MLC = Mean length of clause; C/T = Clauses per T-unit; L = Low proficiency group \((n\)-size\); M = Mid proficiency group \((n\)-size\); H = High proficiency group \((n\)-size\); *The median is provided.
6.4.2.2.1.1 Mean length of T-unit

The data were found to be normally distributed for all tasks and for all proficiency groups, except that the data for Task 1 for the high proficiency participants were found to be positively skewed and leptokurtic in Part 1. In Part 2, all data were found to be normally distributed, with no severe skewness or kurtosis. As can be seen in Tables 57-58 and Figures 73-74, Task 1 and Task 2 elicited by far the shortest T-units from both low and high proficiency groups among the four tasks in both study parts (ranging from 5.94 to 9.76 words per T-unit on average). The low proficiency participants, both in Part 1 and in Part 2, produced similar MLT on average for Task 3 and Task 4, while the high proficiency participants in both parts produced substantially longer T-units when engaging in Task 3 (1.01 words longer than Task 4 in Part 1 and 2.20 words longer than Task 4 in Part 2, both on average). Hence, there was a suggestive interaction effect between proficiency and cognitive task complexity for Task 3, on this measure of MLT. Cohen’s $d$ for the measure of MLT was calculated separately for each proficiency group between Task 1 and Task 4. For the measures of MLT, it was found to be $d = 1.10$ in Part 1 $d = .50$ in Part 2 for the low proficiency group (indicating medium to large effect sizes) and $d = .90$ in Part 1 and $d = .16$ in Part 2 for the high proficiency group (indicating a large effect size in Part 1 and a small effect size in Part 2, Cohen, 1988). Note that the mid proficiency group experienced substantially large effects of increased cognitive complexity on this measure of syntactic complexity in both study parts ($d = 0.83$ in Part 1 and $d = 1.08$ in Part 2).
Figure 73. MLT for the four tasks with 95% confidence intervals by proficiency in Part 1.
6.4.2.2.1.2 Mean length of clauses

Distributions of the data for all, but Task 3 and Task 4 for the low proficiency group and Task 4 for the mid proficiency group, were found to be normal in Part 1. The non-normally distributed data were all found to be substantially positively skewed and leptokurtic. In Part 2, the data for Task 1 for the mid proficiency participants were found to be leptokurtic; otherwise, the data were show to be normally distributed. Figures 75 and 76, as well as Tables 57 and 58 above, indicate that in terms of phrasal level complexity as measured by MLC, Task 1 elicited the shortest MLC (5.19 words per clause on average in Part 1 and 5.51 words in Part 2) from both low and high proficiency groups in both Part 1 and Part 2. The other three tasks elicited

Figure 74. MLT for the four tasks with 95% confidence intervals by proficiency in Part 2.
similar length of clauses from both proficiency groups in both study parts (when compared among the three tasks), except that once again Task 3 elicited substantially longer clauses on average than did Task 2 or Task 4, but solely among the high proficiency participants in Part 2. Cohen’s $d$ between Task 1 and Task 4 for the measure of MLC was found to be $d = .77$ in Part 1 $d = .51$ in Part 2 for the low proficiency group, and $d = 1.19$ in Part 1 and $d = .76$ in Part 2 for the high proficiency group, all indicating medium to large effect sizes (Cohen, 1988). Note that the mid proficiency participants benefited greatly from increased cognitive complexity on this measure of syntactic complexity as well ($d = 1.58$ in Part 1 and $d = 1.21$ in Part 2).

Figure 75. MLC for the four tasks with 95% confidence intervals by proficiency in Part 1.
Figure 76. MLC for the four tasks with 95% confidence intervals by proficiency in Part 2.

6.4.2.2.1.3 Clauses per T-unit

The data for this measure were found to be less normal than the other two syntactic complexity measures in Part 1. The data for Task 1 for the low and mid proficiency groups, Task 2 for the low proficiency group, and Task 4 for the high proficiency group were found to be substantially skewed. Additionally, the data for Task 1 for the mid proficiency group and Task 4 for the high proficiency group were also found to be substantially leptokurtic. In Part 2, the data for Task 1 and Task 3 for the mid proficiency group were shown to be substantially positively skewed. Otherwise, the data were found to be normally distributed. In terms of the last syntactic complexity measure, clauses per T-unit, clearly Task 2 elicited proportionally least subordination
(ranging from 1.00 to 1.28) and Task 3 elicited the most complex sentences (in terms of subordination) from both low and high proficiency participants (ranging from 1.24 to 1.63) (see Figures 77 and 78). In terms of Tasks 1 and 4, the low proficiency participants produced similar numbers of clauses per T-unit when engaging in these two tasks, compared to Task 3 in Part 1, while the low proficiency participants in Part 2 used as much subordination for Task 1 and Task 4 as for Task 2 (instead of Task 3). For the high proficiency participants, the two tasks fell between Task 2 and Task 3 in both parts of the study; however, in Part 1 they elicited substantially less complex language production than did Task 3, but substantially more complex production than did Task 2, while in Part 2, Task 4 and Task 1 elicited as complex sentences as did Task 2 and Task 3, respectively. On the whole, then, there found a clear proficiency effect, but not much in the way of a pattern for task effects, except that once again Task 3 elicited substantially more complex sentences (measured by clauses per T-unit) from the high proficiency group than did the other tasks. Cohen’s $d$ between Task 1 and Task 4 for the measure of clauses per T-unit was calculated to be $d = .42$ in Part 1 $d = .12$ in Part 2 for the low proficiency group (indicating small to medium effect sizes, Cohen, 1998) and $d = .32$ in Part 1 (indicating a small effect size) and $d = -.52$ in Part 2 for the high proficiency group (indicating a large, negative effect size, in favor of the simple task).
Figure 77. C/T for the four tasks with 95% confidence intervals by proficiency in Part 1.
Figure 78. C/T for the four tasks with 95% confidence intervals by proficiency in Part 2.

6.4.2.2.2 Accuracy

In Part 1, the data for Task 2 for the low proficiency group and Task 3 for the high proficiency group were found to be substantially positively skewed and leptokurtic; otherwise the data were all found to be normally distributed. In Part 2, all data were shown to be normally distributed with no severe skewness or kurtosis. As seen in the two tables below (i.e., Tables 59 and 60), generally speaking, the high proficiency participants produced substantially more accurate language production than did their low proficiency counterparts across tasks (see also Figure 79 and Figure 80) – a clear proficiency effect. Comparing the four tasks, overall, they elicited approximately the same accuracy rates, and this was the case for both low and high
proficiency groups. Note, however, that in Part 1, the low proficiency participants produced the least accurate sentences when engaging in Task 4 and that in Part 2, the difference between tasks that elicited the least (i.e., Task 2) and most (i.e., Task 3) accurate language was quite substantial (14.20% different). For the high proficiency participants, Task 3 and Task 1 in Part 1 and Task 3 and Task 4 in Part 2 elicited slightly higher accuracy rates than the other tasks. Cohen’s $d$ between Task 1 and Task 4 for the accuracy measure was calculated to be $d = .32$ in both parts of the study for the low proficiency group (indicating small to medium effect sizes) and $d = -.17$ in Part 1 and $d = .27$ in Part 2 for the high proficiency group (both indicating small effect sizes, Cohen, 1988).

Table 59

<table>
<thead>
<tr>
<th></th>
<th>Accuracy (%)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L (20)</td>
<td>M (14)</td>
<td>H (19)</td>
</tr>
<tr>
<td>Task 1</td>
<td>18.95 (18.49)</td>
<td>48.64 (18.78)</td>
<td>60.05 (27.06)</td>
</tr>
<tr>
<td>Task 2</td>
<td>17.00 (26.10)*</td>
<td>35.43 (18.90)</td>
<td>55.53 (27.51)</td>
</tr>
<tr>
<td>Task 3</td>
<td>18.15 (18.86)</td>
<td>52.29 (20.43)</td>
<td>63.00 (20.71)*</td>
</tr>
<tr>
<td>Task 4</td>
<td>13.50 (15.66)</td>
<td>48.43 (22.00)</td>
<td>55.63 (25.06)</td>
</tr>
<tr>
<td>$d$</td>
<td>$-0.32$</td>
<td>$-0.01$</td>
<td>$-0.17$</td>
</tr>
</tbody>
</table>

Note. L = Low proficiency group ($n$-size); M = Mid proficiency group ($n$-size); H = High proficiency group ($n$-size); *The median is provided.
Table 60
*Means (and Standard Deviations) and Effect Sizes (d) for the Accuracy Measure on the Four Tasks by Proficiency in Part 2*

<table>
<thead>
<tr>
<th></th>
<th>L (16)</th>
<th>M (37)</th>
<th>H (13)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>27.71 (22.24)</td>
<td>33.64 (21.75)</td>
<td>52.78 (31.46)</td>
</tr>
<tr>
<td>Task 2</td>
<td>19.53 (16.75)</td>
<td>35.10 (20.36)</td>
<td>51.04 (25.89)</td>
</tr>
<tr>
<td>Task 3</td>
<td>33.73 (23.74)</td>
<td>37.78 (22.05)</td>
<td>62.24 (21.52)</td>
</tr>
<tr>
<td>Task 4</td>
<td>21.60 (15.56)</td>
<td>35.48 (18.83)</td>
<td>60.28 (22.70)</td>
</tr>
<tr>
<td>d</td>
<td>−0.32</td>
<td>0.09</td>
<td>0.27</td>
</tr>
</tbody>
</table>

*Note. L = Low proficiency group (n-size); M = Mid proficiency group (n-size); H = High proficiency group (n-size)*

*Figure 79. Accuracy for the four tasks with 95% confidence intervals by proficiency in Part 1.*
Figure 80. Accuracy for the four tasks with 95% confidence intervals by proficiency in Part 2.

6.4.2.2.3 Lexical variety

Tables 61 and 62 summarize the results of the two lexical measures: (a) $D$ score and (b) Type-Token Ratio.
Table 61
Means/Medians (and Standard Deviations) for Lexical Variety Measures on the Four Tasks and Effect Sizes (d) by Proficiency in Part 1

<table>
<thead>
<tr>
<th></th>
<th>Lexical Variety</th>
<th>TTR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L (4) M (6) H (15) L (18) M (14) H (19)</td>
<td></td>
</tr>
<tr>
<td>Task 1</td>
<td>23.41 23.67 29.05 0.64 0.56 0.60</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(7.27) (4.06) (7.64) (0.11) (0.07) (0.07)</td>
<td></td>
</tr>
<tr>
<td>Task 2</td>
<td>18.71 24.55 25.91 0.56 0.51 0.52</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(9.16) (6.82) (10.35) (0.08) (0.11) (0.09)</td>
<td></td>
</tr>
<tr>
<td>Task 3</td>
<td>25.74 31.28 31.33 0.60 0.56 0.54</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(4.86) (11.87) (10.30) (0.07) (0.12) (0.09)</td>
<td></td>
</tr>
<tr>
<td>Task 4</td>
<td>21.99 23.79 29.53 0.59 0.51 0.50</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(5.09) (8.23) (6.25) (0.09) (0.10) (0.05)</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>−0.23 0.02 0.07 −0.50 −0.58 −1.64</td>
<td></td>
</tr>
</tbody>
</table>

Note. TTR = Type-Token Ratio; L = Low proficiency group (n-size); M = Mid proficiency group (n-size); H = High proficiency group (n-size); *The median is provided.

Table 62
Means/Medians (and Standard Deviations) for Lexical Variety Measures on the Four Tasks and Effect Sizes (d) by Proficiency in Part 2

<table>
<thead>
<tr>
<th></th>
<th>Lexical Variety</th>
<th>TTR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>Task 1</td>
<td>26.17 27.38 30.33 0.63 0.61 0.55</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(7.61) (8.59)* (8.85) (0.09) (0.08) (0.07)</td>
<td></td>
</tr>
<tr>
<td>Task 2</td>
<td>22.37 26.60 29.51 0.54 0.54 0.50</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(11.60) (7.87) (11.49) (0.11) (0.08) (0.07)</td>
<td></td>
</tr>
<tr>
<td>Task 3</td>
<td>35.18 32.44 37.32 0.63 0.56 0.53</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(17.15) (11.18) (9.74) (0.13) (0.07) (0.08)</td>
<td></td>
</tr>
<tr>
<td>Task 4</td>
<td>26.71 27.38 28.83 0.55 0.52 0.46</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(7.27) (8.45) (6.98) (0.10) (0.07) (0.06)</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>0.07 0 −0.19 −0.84 −1.20 −1.38</td>
<td></td>
</tr>
</tbody>
</table>

Note. TTR = Type-Token Ratio; L = Low proficiency group (n-size); M = Mid proficiency group (n-size); H = High proficiency group (n-size); * The median is provided.

6.4.2.2.3.1 D

The data for the D measure and TTR were both found to be normally distributed, except the data for Task 4 for the mid proficiency participants in Part 1. This set of data was shown to be substantially positively skewed and leptokurtic. In Part 2, the data for Task 1 for the mid proficiency group was found to be substantially positively skewed and leptokurtic.
proficiency group were substantially skewed; otherwise the data were all normally distributed. In both Part 1 and Part 2, Task 3 elicited the highest $D$ scores for both the low and high proficiency groups on average (see Figures 81 & 82). The difference between Task 3 and the other three tasks in Part 2 appears to be quite substantial for both proficiency groups. Cohen’s $d$ between Task 1 and Task 4 for $D$ was found to be small overall. For the low proficiency group, it was $d = -0.23$ in Part 1 and $d = 0.07$ in Part 2, both indicating small effect sizes although in distinct directions (Cohen, 1988). Similarly, for the high proficiency group, it was found to be $d = 0.07$ in Part 1 and $d = -0.19$ in Part 2, both indicating small effect sizes, again in different directions (Cohen, 1988).

![Graph](image)

*Figure 81.* $D$ score for the four tasks with 95% confidence intervals by proficiency in Part 1.
Moving on to the second lexical variety measure, Type-Token Ratio (TTR), the data for this particular measure were shown to be all normally distributed, with no severe skewness or kurtosis in Part 1 and Part 2. Inspecting the data presented in Tables 61-62 and Figures 83-84 below, Task 1 in Part 1 and Task 1 and Task 3 in Part 2 elicited substantially more varied lexical use (measured by TTR) than the other tasks for the low proficiency participants. For the high proficiency participants, in both study parts, Task 1 elicited the highest TTR, whereas Task 4 elicited the lowest TTR, with Tasks 2 and 3 falling between the two. As noted earlier, caution has to be exercised in interpreting these patterns because TTR has been shown to be very sensitive.
to text length. Indeed, participants across proficiency levels produced the shortest narrations when engaging in Task 1 (see Tables 63 & 64 below). Note also that the high proficiency participants produced substantially more words than did their low proficiency counterparts. This observation may also explain why TTR was generally lower for the high proficiency group on average than for the low proficiency group. For TTR, Cohen’s $d$ between Task 1 and Task 4 was calculated to be $d = -.50$ in Part 1 and $d = -.84$ in Part 2 for the low proficiency group (indicating medium to large effect sizes, in favor of the simple task) and $d = -1.64$ in Part 1 and $d = -1.38$ in Part 2 for the high proficiency group (both indicating quite large effect sizes, again in favor of the simple task, Cohen, 1988).

![Figure 83. TTR for the four tasks with 95% confidence intervals by proficiency in Part 1.](image)
Figure 8. TTR for the four tasks with 95% confidence intervals by proficiency in Part 2.

6.4.2.2.4 Fluency

The data distribution for the measure of the total number of words was found to be positively skewed for Tasks 1, 2, and 3 for the higher proficiency groups in Part 1. In Part 2, the data for Tasks 1, 2, and 4 for the mid proficiency group were shown to be positively skewed. As can be seen in the two tables below (i.e., Tables 63 & 64), the high proficiency group on average produced much longer stories than did the low proficiency group (see also Figures 85 & 86) – a clear proficiency effect. There also seems to have been an interaction effect between proficiency and tasks in that the low proficiency participants produced similar numbers of words across the four tasks (although in Part 2 Task 4 elicited a substantially larger number of words than did
Task 1), whereas the high proficiency participants produced increasingly more words from Task 1 to Task 4. In fact, for the high proficiency participants, the mean values of Task 4 clearly exceed the upper 95% confidence intervals for Task 1 in both study parts. For the total number of words, Cohen’s $d$ between Task 1 and Task 4 was calculated to be $d = .42$ in Part 1 and $d = .68$ in Part 2 for the low proficiency group (indicating medium effect sizes) and $d = 2.56$ in Part 1 and $d = 1.46$ in Part 2 for the high proficiency group (both indicating very large effect sizes, Cohen, 1988).

Table 63
*Means/Medians (and Standard Deviations) and Effect Sizes ($d$) for the Fluency Measures on the Four Tasks by Proficiency in Part 1*

<table>
<thead>
<tr>
<th>Task</th>
<th>Total Number of Words</th>
<th>Words per Seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L (20)</td>
<td>M (13)</td>
</tr>
<tr>
<td>Task 1</td>
<td>43.05</td>
<td>57.92</td>
</tr>
<tr>
<td></td>
<td>(16.63)</td>
<td>(22.78)</td>
</tr>
<tr>
<td>Task 2</td>
<td>41.95</td>
<td>64.31</td>
</tr>
<tr>
<td></td>
<td>(14.52)</td>
<td>(23.24)</td>
</tr>
<tr>
<td>Task 3</td>
<td>48.85</td>
<td>70.00</td>
</tr>
<tr>
<td></td>
<td>(17.45)</td>
<td>(23.66)</td>
</tr>
<tr>
<td>Task 4</td>
<td>50.75</td>
<td>83.23</td>
</tr>
<tr>
<td></td>
<td>(20.28)</td>
<td>(27.68)</td>
</tr>
<tr>
<td>$d$</td>
<td>0.42</td>
<td>1.00</td>
</tr>
</tbody>
</table>

*Note. L = Low proficiency group (n-size); M = Mid proficiency group (n-size); H = High proficiency group (n-size); * The median is provided.*
Table 64
*Means/Medians (and Standard Deviations) and Effect Sizes (d) for the Fluency Measures on the Four Tasks by Proficiency in Part 2*

<table>
<thead>
<tr>
<th></th>
<th>Total Number of Words</th>
<th></th>
<th>Words per Seconds</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>52.63</td>
<td>57.00</td>
<td>81.83</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>(20.00)</td>
<td>(24.15)*</td>
<td>(25.51)</td>
<td>(0.28)</td>
</tr>
<tr>
<td>Task 2</td>
<td>59.00</td>
<td>58.00</td>
<td>92.75</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>(30.50)</td>
<td>(32.45)*</td>
<td>(30.23)</td>
<td>(0.23)</td>
</tr>
<tr>
<td>Task 3</td>
<td>59.25</td>
<td>77.19</td>
<td>101.58</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>(26.14)</td>
<td>(31.75)</td>
<td>(33.83)</td>
<td>(0.24)</td>
</tr>
<tr>
<td>Task 4</td>
<td>69.13</td>
<td>74.00</td>
<td>126.08</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>(27.86)</td>
<td>(40.65)*</td>
<td>(34.45)</td>
<td>(0.26)</td>
</tr>
<tr>
<td>d</td>
<td>0.68</td>
<td>0.51</td>
<td>1.46</td>
<td>−0.19</td>
</tr>
</tbody>
</table>

*Note. L = Low proficiency group (n-size); M = Mid proficiency group (n-size); H = High proficiency group (n-size); * The median is provided.*

*Figure 85.* Total number of words for the four tasks with 95% confidence intervals by proficiency in Part 1.
Moving to the other fluency measure, words per seconds, the data distribution for this measure was found to be normal for all, but one of the tasks. The data for Task 2 for the low proficiency group were shown to be positively skewed and leptokurtic in Part 1. In Part 2, all data were shown to be normally distributed with no severe skewness or kurtosis. What stands out in the results of the fluency measure in Tables 63-64 and Figures 87-88 below is the fact that each proficiency group produced radically different speech rates, with the high proficiency group producing most words per second and the low proficiency group producing the least. Comparing the four tasks for each proficiency group, they elicited similar number of words per second.
regardless of the proficiency groups. Overall, then, there observed a clear categorical proficiency effect, but no task or interaction effects were found for this measure of fluency. For this measure of fluency, Cohen’s $d$ between Task 1 and Task 4 was calculated to be $d = -.47$ in Part 1 and $d = -.19$ in Part 2 for the low proficiency group and $d = -.42$ in Part 1 and $d = -.19$ in Part 2 for the high proficiency group, all indicating small to medium effect sizes, in favor of the simple task (Cohen, 1988).

Figure 87. W/S for the four tasks with 95% confidence intervals by proficiency in Part 1.
6.4.2.2.5 Summary of proficiency differences

All measures of task performance, except for the lexical variety measures, indicated that the high proficiency participants produced considerably better performances than did the low proficiency participants. Comparing the four tasks with differing degrees of cognitive complexity, once again Task 3, that was found to be second simplest for the high proficiency participants and the second most complex for the low proficiency participants, elicited the best performance of all the four tasks for both groups. More specifically, Task 3 elicited the syntactically most complex and lexically most diverse speech, with no apparent detrimental effects on accuracy or fluency (speech rate) for both proficiency groups (although to varying
degrees). Looking at the other three tasks, patterns were found to be similar across the two proficiency groups. Task 1 and Task 2 elicited syntactically and lexically least complex language production (in terms of MLT and MLC for Task 1 and MLT and clauses per T-unit for Task 2), while Task 4 elicited relatively (if not most) syntactically complex utterances (in terms of MLT and MLC). Importantly, however, there seemed to be (suggestive) interaction effects between proficiency and tasks for all syntactic complexity measures as well as the measure of the total number of words produced. In other words, Task 3 elicited syntactically the most complex speech (overall, at the phrasal level, and in terms of subordination) from the high proficiency participants, but not from the low proficiency counterparts; and the tasks elicited increasingly more words as the number of characters involved in the story went up, and with dramatic differences exhibited by the high proficiency participants. Accuracy and fluency (in terms of speed) seem to have been affected by cognitive task complexity to the smallest extent; any differences were observed to be more closely related to learners’ L2 proficiency levels than tasks. Note, however, that Task 3 tended to elicit slightly more accurate speech from both low and high proficiency groups in both study parts. It is interesting to observe that it was neither the measured-to-be most or least complex tasks that elicited the best performances from either of the high or low proficiency groups. Possible reasons behind this rather unexpected finding will be explored in the following chapter.

6.4.3 Influence of the dual-task methodology on task performance

Lastly, potential influences that the dual-task methodology may have had on participants’ task performance require some attention. Comparing Study Part 1 and Part 2, it can be observed that the two conditions had only a minimal perceptible influence on native speakers’ task
performances. In both the dual- or single-task conditions, they produced similarly complex (both syntactically and lexically) and fluent speech. Note, however, that the single-task condition elicited longer utterances from Task 2 and Task 4 (14 and 21 more words, respectively) than did the dual-task condition. Likewise, the secondary, color detection task in general did not seem to have influenced English L2 speakers’ task performance very substantially, either, except that the participants in the dual-task condition (regardless of tasks) produced shorter stories (about 10 words less on average) than did the participants in the single-task condition. When analyzed by proficiency, however, the dual- and single-task conditions elicited slightly different performances. The single-task condition elicited syntactically more complex (in terms of MLT and MLC), lexically more diverse, and more fluent utterances than did the dual-task condition from both low and high proficiency groups on average (regard less of tasks). Additionally, the low proficiency participants produced more accurate speech under the single-task condition than the dual-task, for all four tasks. Note that, given the small n-sizes of the low and high proficiency groups in both study parts, these differences may be more idiosyncratic than generalizable trends. Nevertheless, based on the overall trends with native and English-L2 speakers, it may be safer to conclude that the secondary task may have had minimal consequences for their task performances except for the length of speech.

6.5 Chapter summary

This chapter reported on the quantitative results of the four cognitive load measures for both native and English-L2 speaker participants, as well as the qualitative results of the two open-ended questions on the task difficulty and mental effort questionnaire gathered from the English L2 speakers. Additionally, it reported on the effects of measured cognitive complexity
on task performance indices for native and English-L2 participants. For the native speakers of English, although Task 2 and Task 4 were perceived to be more difficult and more mental effort consuming than the other two tasks, more objectively (measured by the dual-task methodology and time estimation), all the four tasks were measured to have posed relatively similar degrees of cognitive complexity to this particular group of native English speakers in the two study parts. These cognitively similar tasks, however, elicited task performances that were qualitatively different. Task 3 (on all three measures) and Task 4 (on MLT and MLC) elicited syntactically more complex performances than the other two tasks. Task 4 elicited the longest speech, whereas Task 1 elicited the most fluent speech of all the four tasks. Recall, however, the four tasks elicited only marginally different performances in terms of lexical variety regardless of their complexity levels.

Turning our attention to the results of the English L2 participants, overall, Task 1 was measured to be the simplest task and Task 4 to be the most complex task. The middle two tasks (in design) indeed fell between Task 1 and Task 4; however, Task 2 was shown to be more complex than Task 3, going against the prediction based on the task design feature of the number of elements. The participants’ open-ended comments revealed that Task 1 was both conceptually and linguistically simplest among the four tasks, whereas Task 4 was the complete opposite. Recall, however, that this simplicity of Task 1 made it seem more difficult and required higher mental effort perceptually for some participants, as they felt that they would have to exercise their creativity more rigorously to elaborate on the simple input and come up with a story. Task 2, then, was perceived to have posed high conceptual and linguistic demands like Task 4, although not quite as much. It was observed that the ambiguous storyline, the repetitive nature of
the character’s actions, and their related and yet distinct behaviors made Task 2 difficult and mental effort consuming. A lot of the participants were not familiar with vocabulary related to the topic (fishing), either, which increased perceived difficulty and mental effort even more. Lastly, regarding Task 3, the participants felt that in general it was easy to figure out its storyline (although some aspects of the conceptual input added extra perceived difficulty and mental effort) but that having to deal with many characters made the task conceptually and linguistically difficult and mental effort consuming. In turn, these four tasks with varying degrees of complexity had distinct effects on the English L2 participants’ task performance in terms of the CALF indices. It was Task 3 that elicited the best performance of all the four tasks. More specifically, it elicited the syntactically most complex, most accurate (in Part 2, although not to a meaningful extent), and lexically most diverse (on the $D$ measure) language production. Looking at the other three tasks, Task 1, the measured-to-be simplest task, elicited syntactically and lexically simplified language production, whereas Task 4, the measured-to-be most complex task elicited the most number of words and syntactically quite complex speech (although perhaps not quite as complex as Task 3). Lastly, Task 2, which posed conceptually and linguistically high demand, elicited overall the worst task performances in that it elicited syntactically and lexically simplified language production, and did not lead to any advantages on either accuracy or fluency.

Looking at the results of the English L2 speakers’ by their proficiency, in general the four tasks were found to be less demanding for the high proficiency participants and elicited better task performances from the high proficiency participants than their low proficiency counterparts. For the high proficiency participants, all measures, except for time estimation, revealed that the order of complexity was Task 1, followed by Task 3, with Tasks 2 and 4 being the most complex
ones (but they change places depending on the measures). For the low proficiency participants, on the other hand, the order of cognitive complexity was shown to be as predicted (based on the manipulated task design feature). The participants’ open-ended comments revealed that the high proficiency participants’ attempts to realize their more sophisticated understandings of a ‘good’ task performance required them to (attempt to) use various words and phrases to express essentially the same actions (e.g., fishing, catching a fish), focus on conveying the order of the pictures correctly, and elaborate on the monotonous, repetitive actions while engaging in Task 2. On the other hand, the same task was perceived to be easy and less mental effort consuming for at least some of the low proficiency participants because the repetitive nature of the actions allowed them to repeatedly use the same words and expressions and to complete the task more or less successfully (in their terms) simply by conveying the gist of the story. Among these tasks, once again Task 3 elicited the best performance (syntactically most complex and lexically most varied speech with no detrimental effects on accuracy or fluency in terms of speech rate) from both low and high proficiency participants, generally speaking. Here, it is curious why Task 3 – that was measured to pose lower cognitive load than Task 2 for the high proficiency participants but higher cognitive load than Task 2 for the low proficiency participants – elicited similar patterns in terms of its effects on learner task performance. Task 1 and Task 2 elicited syntactically and lexically least complex language production (in terms of MLT and MLC for Task 1 and MLT and clauses per T-unit for Task 2) – furthermore, neither accuracy nor fluency (speech rate) were positively affected. By contrast, Task 4 elicited relatively (if not the most) syntactically complex utterances (in terms of MLT and MLC) with no obvious disadvantage on accuracy or fluency (speech rate). These trends were observed for both proficiency groups. The
only clear interaction effect found in task performance was for the measure of the total number of words produced: The tasks elicited increasingly greater numbers of words as the number of elements or characters involved in the story went up, but only from the high proficiency group. Note, however, that we do see a suggestive interaction effect on syntactic complexity for Task 3. In other words, Task 3 elicited syntactically the most complex speech of the four tasks from, and only from, the high proficiency group.

Lastly, the influence of the secondary, color detection task on participants’ task performances were found to be minimal for both native and English-L2 speakers. The only exception was that the single-task condition elicited longer stories for all four tasks from the English L2 participants and for Task 2 and Task 4 from the native speakers of English.
CHAPTER VII: DISCUSSION

7.1 Introduction

This chapter discusses the results reported in Chapter 6, by emphasizing important patterns that respond to the original research questions, highlighting observations that shed new light on our understandings of task complexity, and referring to related theories and previous research in the fields of cognitive psychology, multimedia learning, and second language acquisition (with a focus on TBLT in particular). The discussion will be divided into two main sections, focusing in turn on: (a) validating the assumed relationship between task design and the level/source of cognitive complexity; and (b) interpreting the effects of measured cognitive complexity on task performance.

7.2 Validating the assumed relationship between task design and the level of cognitive complexity

This section firstly summarizes the overall results regarding measurement of cognitive load of the four oral narrative tasks, and it then addresses four important issues that have emerged from the overall patterns.

7.2.1 Summary of overall patterns

To reiterate, the first three research questions asked:

1. Can we detect differences in cognitive complexity of the four oral narrative tasks using (a) dual-task methodology, (b) time estimation, (c) task difficulty self-assessment, and (d) mental effort self-assessment for university-level Japanese-L1 learners of English as well as for native speakers of English, and does the measured cognitive complexity reflect the designed differences for both groups of participants?
2. What seems to make a task more or less complex for the English L2 participants?

3. Does L2 proficiency matter in terms of the level and types of cognitive complexity as well as measures of cognitive load for the English L2 participants?

As summarized in Chapter 6, the basic answers to the first three research questions were as follows. The four tasks were found to have posed similar degrees of cognitive complexity to the native speaker participants, according to the objective measures of cognitive load, although Task 2 (the designed-to-be second simplest task) and Task 4 (the designed-to-be most complex task) were perceived to be somewhat more difficult and more mental effort consuming than the other two tasks. By contrast, for the English L2 speaker participants, Task 1, the designed-to-be cognitively simplest task among the four, was indeed shown to be the simplest task by all measures of cognitive load. Similarly, Task 4, the cognitively most complex task by design, was consistently shown to be the most complex task. The two medium-difficulty tasks by design (i.e., Task 2 and Task 3) fell between Task 1 and Task 4, in terms of their cognitive demands, on all measures. However, the differences in the degree of cognitive complexity found between these two tasks tended to be small, with overlapping confidence intervals, and against the intended task design, with all measures (except for time estimation in Part 2) suggesting Task 2 to be slightly more cognitively demanding than Task 3. Note, however, that in Part 2, perceived task difficulty and mental effort ratings were substantially higher for Task 2 than for Task 3.

Turning to the second research question, conceptual input, code complexity, and performance factors were identified to be the main sources of easiness/difficulty and low/high mental effort. In general, the number of elements seems to have influenced the level of cognitive complexity as predicted: It was mentioned as a source of easiness for Task 1 and progressively...
more as a source of difficulty as task complexity increased by design. This particular finding seems to give support to Robinson’s (2001a) proposal about the relationship between the number of elements and cognitive task complexity. Looking across the four tasks, Task 1 posed conceptually as well as linguistically the lowest demands of all four tasks, whereas Task 4 was found to pose the highest demands in these dimensions. However, despite the fact that it only contained two characters, Task 2 was measured to be the second most complex task due to its conceptual (for reasons other than number of elements) and linguistic demands. To approximately half of the participants, its storyline was ambiguous and the input was in some places hard to decipher. On top of these conceptual difficulties, some of the task-useful, if not essential, words were related to fishing, and many of the participants expressed that they were unfamiliar with these words, which in turn increased their perceived difficulty and effort exerted. Lastly, by contrast, the storyline of Task 3 was easy to figure out in general, but the language required to tell the story, for many participants, was difficult. Performance factors, related to participants’ perceived performance quality, were also mentioned as a source of easiness/low mental effort and of difficulty/high mental effort; however, they were secondary to the other two superordinate categories and were less related to task design per se.

As for the last research question of the first section (i.e., proficiency effects), the four tasks were measured to have posed distinct levels of cognitive complexity to the low and high proficiency groups. Overall, the tasks posed higher cognitive demands to the low proficiency learners than to their high proficiency counterparts, except for Task 2. The order of the four tasks for the low proficiency group (based on trends in the combined cognitive load measures) was as predicted by task design: Task 1, Task 2, Task 3, and Task 4 (going from simple to more
complex), whereas the order for the high proficiency group was: Task 1 (simplest task) followed by Task 3, and then by Task 2 and Task 4 (most complex tasks). In an attempt to shed light on this difference in the order of complexity for the two proficiency groups, participants’ open-ended comments were consulted. They revealed that the repetitive, monotonous input of Task 2 (i.e., the main action of the two characters was fishing in all of the picture frames) may have worked to the advantage of at least some of the low proficiency participants, whereas the same input seemed to have worked against many of the high proficiency participants. That is, the repetitive nature of the actions in Task 2 made the task easy linguistically for the low proficiency learners, because they could use the same expressions repeatedly, and the same task feature also allowed them to complete the task more or less successfully (in their terms) simply by conveying the gist of the story (i.e., fishing). However, the same phenomenon appeared to have worked against the high proficiency learners because of their attempts to realize their more sophisticated understandings of what a good narrative performance might be (e.g., varying expressions to convey the same event, distinguishing among similar picture frames to include all picture frames in the story). In fact, Task 2 elicited the most number of comments about participants’ efforts at a linguistically better performance as a source of difficulty and mental effort, most of which came from the high proficiency participants. Overall, regardless of the task, high proficiency participants’ more sophisticated understandings of what a good story or narrative language performance might be seemed to make all of the tasks additionally difficult and mental effort consuming for these participants, as compared to the low proficiency learners. Additionally, across tasks, the high proficiency participants were more likely to be aware of the specific linguistic issues that they were confronted with while engaging in each task rather than simply
attributing difficulty/high mental effort to a lack of their general English proficiency (which we saw more often with the low proficiency group). Lastly, it is worth noting that the dual-task methodology and time estimation produced more robust results that seemed more in alignment with the self-assessment measures when used with the high proficiency group than with the low proficiency group. (Possible reasons for this interaction of proficiency with cognitive load measures will be explored further below.)

7.2.2 Further discussion

Four important issues emerge from the overall patterns reported in the previous section and merit further attention: (a) the relationship between learner factors and the level of cognitive load, (b) the relationship between the number of elements and the level of cognitive load, (c) types of cognitive load, and (d) differences between the four cognitive load measures themselves.

7.2.2.1 Relationship between learner factors and the level of cognitive load

One aspect of cognitive complexity that has received minimal attention in the field of TBLT, but has been the focus of investigation in cognitive psychology, is the relationship between learner factors and the level of cognitive load. As reviewed in Chapter 1, in cognitive psychology, learner variables (e.g., learners’ knowledge level of the task content) are presumed to have a critical effect on the degree of intrinsic cognitive load; in other words, the level of intrinsic cognitive load cannot be determined without taking into account who the target learners are (e.g., Sweller, 1994, Kalyuga, 2007). Findings from the current research have indicated that this may also be the case with the types of L2 communication tasks used in TBLT research and teaching. Clearly, participants at differing proficiency levels devoted distinct amounts of
attentional resources to doing the exact same tasks, and they also perceived them to be
differentially difficult and effortful – on the whole, the tasks were less complex for higher
proficiency learners, though with a few exceptions discussed below. While proficiency effects
have been attested in a handful of TBLT studies, they have always been in relation with
outcomes measures, not in relation with the cognitive complexity of the task designs per se.

Additionally, participants’ open-ended comments revealed that learners’ motivation to tell
a good, effective story, which also seemed to be closely associated with their proficiency
differences, clearly affected the level of cognitive complexity experienced by the participants,
though in perhaps unanticipated ways. Recall that the simple input of Task 1 (and Task 2 to some
extent) unexpectedly posed high task difficulty to some of the high proficiency participants.
They expressed that Task 1 was difficult because the input was so simple and minimal that they
could not come up with a satisfactory story to tell. However, this concern, that a dearth of
information made the task difficult, rather than easy, was observed only negligibly in low
proficiency participants’ comments. It seems to be the case, then, that the high proficiency
participants’ understandings of what a good narrative story might be and their high expectations
to realize these understandings in their performances made the task relatively more difficult for
them than it might otherwise have been. Previously, Tavakoli (2009), in her TBLT interview
study, found similarly that “a paucity of information in a picture story could affect TD [task
difficulty]” (p. 11). It is important to keep in mind, therefore, that simplicity of task input does
not always make a task easy for all learners.

Similarly, participants’ open-ended responses, especially with regards to the performance
factors, also revealed that learners sometimes chose to devote more or less mental effort to a task
at hand, regardless of the task input. There were several participants who expressed that a given task required high mental effort because they tried so hard to follow the task instructions carefully, to pay attention to other performance factors (e.g., whether the story is easy to understand, describing details of the story, adding characters’ emotions to the story), and basically to perform well. These comments were found across tasks, regardless of their intrinsic cognitive demands. Alternatively, several participants, again for all four tasks, expressed that they intentionally reduced the amount of mental effort needed for a given task so that they could deal with the high cognitive demands that it posed, for example by prioritizing task accomplishment over performance quality. At the individual level, this learner choice factor clearly influenced the level of cognitive load of each narrative task in the present study.

Learners’ choice over how much effort to devote has also been attested in multimedia learning (e.g., van Gog & Paas, 2008). In the domain of TBLT, this learner choice and its fundamental roles (although in relation with the consequences of their task performance, rather than the level of cognitive demands) have been pointed out by Skehan (2009) in explaining why simultaneous improvement of linguistic complexity and accuracy was made possible when a post-task activity of transcribing one’s utterances was given (see Chapter 1 for more details of his observation). These findings of the current study illustrate the complicated nature of the relationship between task design and its putative cognitive complexity, and they underscore the critical requirement to measure the level of cognitive task complexity, instead of assuming it based on task design features, for a particular group of learners with whom we are working.
7.2.2.2 Relationship between the number of elements and the level of cognitive load

One of the most interesting results, in relation to the original and perhaps simplistic design feature adopted to distinguish among the four tasks, was that regardless of other differences in task features, the number of elements (characters in the stories) was perceived firstly as a source of easiness and then progressively more often as a source of difficulty as the number of characters increased from Task 1 to Task 4. For Task 1, the measured-to-be simplest task, approximately 15% of participants in both Study Part 1 and Part 2 expressed that having to deal with only one character in the story made Task 1 easy and less mental-effort consuming, whereas nobody stated the opposite opinion. For the middle tasks 2 and 3, then, this factor was mentioned by some as a source of easiness (or low mental effort) and by others as a source of difficulty (or high mental effort), with proportionally greater emphasis on difficulty for Task 3 (which featured four characters, versus two in Task 2). Lastly for Task 4, the measured-to-be most complex task, the number of elements was referred to almost exclusively as a source of difficulty or high mental effort (by 26% of participants in Part 1 and close to 50% of the participants in Part 2), except for one person (in the high proficiency group) who expressed that having a lot to talk about made the task easy.

Probing their comments more in depth, it became apparent that having a large number of characters to deal with required participants not only to distinguish among similar characters (linguistic demands), but also to figure out the relationship among them, who should be described, who’s who in each picture frame, and so on (conceptual demands). The former type of complexity encouraged learners to push the limits of their interlanguage resources to meet the increased linguistics demands and focus on improved performance, hence it seems to have had
facilitative effects on learners’ task performances. This facilitative role of the number of elements is exactly why Robinson (2001b) has argued that a more complex task would elicit (syntactically and lexically) more complex and more accurate language production. What has been neglected until today, perhaps, is the possibility that there may also be deleterious effects from the conceptual demands associated with having to deal with a large number of elements, in terms of learners’ attentional allocation. While they had a lot to do linguistically, which may have lead to enhanced performance for some, the participants in this study also needed to figure out the relationship among many characters, who should be described, and who’s who in each picture frame (especially in Task 4), which likely required many of the English L2 participants to devote a large portion of attentional resources to that end and distracted them from paying attention to performance aspects per se. This distinction between facilitative and extraneous cognitive load will be discussed further in the next section, and the relationship between these two types of cognitive load and learner task performance will be explored in the second section of this discussion.

Lastly, as already discussed in Chapter 6, some of the participants’ comments regarding the number of elements beg the question of what should be considered as ‘elements.’ Although only a few, there were participants who expressed that Task 3 was easy because the characters in action were only the children or because they could refer to the characters by using different terms (e.g., a mother, a boy, a girl, a baby), rather than noun modifiers to further describe each character (e.g., a man who is wading in the river and a man on the river bank). These features of Task 3 clearly contrast with Task 2, where the two characters (that looked similar for some participants) were acting individually. This difference in the role played by the number of
characters may also add some clarification as to why overall Task 3 was measured to be less cognitively complex than was Task 2. These observations emphasize the criticalness of considering similarities and differences among characters and how their actions are related, in addition to simply counting the mere number of characters involved in each story.

7.2.2.3 Types of cognitive load: Facilitative versus extraneous cognitive load

Participants’ explanatory comments for their task difficulty and mental effort ratings overall revealed that complexity seemed to have been caused by two different types of cognitive load: (a) extraneous cognitive load that was perceived to be deleterious, wasteful, and irrelevant to task performance; and (b) facilitative cognitive load that encouraged learners to push the limits of their interlanguage resources and focus on improved performance. Interestingly, all of the tasks were found to pose a mix of both extraneous and facilitative load, though to clearly varying degrees.

Generally, extraneous cognitive load in this study context was caused by ambiguous (as seen in Task 2 and partially in Task 3), monotonous (as seen in Task 2), or complicated storylines (as seen in Task 4), and in particular by the need to pay attention to details of the pictures (as seen in Task 2, Task 4 and partially in Task 1 and Task 3) as well as excessively high code complexity (as seen in Task 2). Task 1 likely posed the least extraneous cognitive load due to the simplicity of the input. Even for Task 1, though, some participants found the second picture frame ambiguous, in terms of why the main character checked outside, or the fourth picture frame, where it was unclear why the light came in when the curtains were shut. For a few English L2 speakers, these ambiguities were demanding in an extraneous way, in that they had to devote attentional resources not to linguistic aspects of their performances but to figuring out the
minute details of the input. Task 2, by contrast, seems to have posed high extraneous cognitive load, conceptually, to approximately one third of the English L2 participants. For these participants, its storyline was quite ambiguous, its input was in some places difficult to decipher, and the monotonous action added extra cognitive demands in terms of coming up with things to say. Similarly, to this end, particularly in Study Part 2, participants commented that their attempts to tell a good, effective story based on such input – for example, by distinguishing between similar picture frames to make sure that all pictures are included in the story, or actually coming up with a story (based on the monotonous input) rather than simply describing what was happening in the sequence of pictures – made the task difficult to handle and/or perceptually increased mental effort exerted. In addition, Task 2 posed high extraneous cognitive load linguistically as well. As many as 28 participants in total, summing up Part 1 and Part 2 participants, expressed that their effort to tell an effective story required them to pay additional attention to linguistic forms, which could have been an example of facilitative, not extraneous, cognitive load. What was prevalent among these participants’ comments for Task 2 was their (especially high proficiency participants’) effort to vary expressions to refer to essentially the same actions of fishing, which ideally would increase their lexical variety for this particular task (as seen in the native speaker performance – Task 2, although marginally, elicited the largest $D$ values in both study parts). Unexpectedly, however, this increased attention to lexical variety was not reflected in either of the $D$ or TTR measures for the English L2 speakers. Curiously, among the 10 explanatory comments (provided by the English L2 participants) that explicitly referred to participants’ attention to varied uses of expressions for Task 2 (in total, summing up Part 1 and Part 2), four of them found Task 2 difficult because they wanted to use different
words to describe similar actions, but they could not, and ended up using the same expressions, and three other comments more explicitly expressed that they used a lot of mental effort trying to come up with different expressions but could not in the end. These participants then were spinning their wheels, ‘wasting’ their attention and mental effort by devoting them to the linguistic aspect for which they did not have enough linguistic resources. Turning to Task 3, yet another pattern emerged in that, despite the more complicated storyline and higher number of characters, it seemed to pose quite low extraneous load, similar to Task 1. Thus, the majority of the participants expressed that figuring out of the storyline per se was easy and did not require much mental effort, though a few participants were unsure of how to refer to the black ball with a face on it and whether the mother was sleeping or being distracted by one of the children. Lastly, as discussed in the previous section, Task 4 posed relatively high extraneous cognitive load for many of the participants, in that having to deal with a number of characters in the story, participants were required to use their attentional resources to first understand the input per se rather than to focus on their performances.

Facilitative cognitive load, on the other hand, was caused by learners’ effort to paraphrase unfamiliar expressions (as seen in all tasks), to deliver an effective narrative story (as seen in all tasks), and to deal with a large number of characters (as seen in all tasks, but Task 1). The four tasks elicited approximately the same number of comments about participants’ effort to paraphrase unfamiliar expressions with their existing linguistic resources. Summing across Study Part 1 and Part 2, six, eight, seven, and nine participants made such comments for Task 1, Task 2, Task 3, and Task 4, respectively. Regarding their effort to tell an effective story, again summing across both study parts, 15 participants for each of Task 1 and Task 3, 18 participants
for Task 2 (excluding the ones who made comments about their attention to varied uses of expressions) and nine participants for Task 4 expressed that their effort to tell a good, effective story required them to pay attention to accuracy and fluency of their task performances and to look for appropriate language forms as a tool to make the story creative, detailed, and coherent, which in turn increased their perceived mental effort and difficulty. Moving on to the last factor of facilitative cognitive load, as argued by Robinson (2001b), the degree of facilitative cognitive load increased as the number of elements involved in each task increased. None of this type of cognitive load was detected in Task 1, whereas approximately 15% of the participants in Part 1 (8 participants in number) and 43% in Part 2 (29 participants) explicitly expressed that the need to deal with a large number of characters in Task 4 posed linguistic demands of distinguishing among similar characters and dealing with increased amount of information. Task 2 and Task 3 fell between the other two tasks: A handful of participants expressed that they had to devote extra attention to how to refer to each character and the appropriate conjugation between the subject and the verb. To summarize, Task 1 seems to have posed low cognitive load in terms of both extraneous and facilitative types, whereas Task 4 seems to have posed high extraneous and high facilitative cognitive load. Task 2 and Task 3 likely posed relatively medium facilitative cognitive load, with the former also posing very high extraneous cognitive load and the latter low extraneous load.

These findings highlight an interesting relationship with Robinson’s (e.g., 2001a) notion of resource-directing and resource-dispersing dimensions of task design. Firstly, the relationship between increased attention triggered by resource-directing factors and learner task performance seems to be quite complicated. As seen in Task 2, even though learners’ attention is directed to
some specific linguistic aspects of their task performance in response to the demands posed by a given task, that does not necessarily guarantee that it has facilitative effects on their task performance. Hence, heightened attention may not always lead to improved performance. According to learners’ open-ended responses, Task 2 encouraged at least some participants to pay specific attention to the varied uses of expressions for the action of fishing, and yet these participants ended up using their attentional resources and mental effort somewhat in vain, trying to respond to task demands that were above the level of their linguistic competency. Future studies should investigate more rigorously this relationship between increased attention and task performance in the context of task complexity research. Secondly, the task design features which belong to what Robinson calls a resource-directing dimension (e.g., number of elements) may well simultaneously induce resource-dispersing (or what I have been calling extraneous) type of cognitive complexity. In general, in Task 4, having a large number of elements (or characters in this case) to refer to in the story posed not only high linguistic demands of distinguishing among similar characters and describing how their actions are related to each other, but also high conceptual demands of understanding the task input (e.g., who’s who in each picture frame, among many characters who should be the focus of the story and who should be included in it). The former type of demands seemed to have drawn participants’ attention to linguistic aspects of their task performance (as claimed by Robinson), but the latter demands, that were induced by the same task feature of [-few elements], most likely distracted them from focusing on improved performance. Thus, while Robinson’s Triadic Componential Framework and Cognition Hypothesis are useful in thinking about what factors might affect the level of cognitive task complexity and in turn how cognitive complexity may affect learner task performance, it is
nonetheless crucial to verify the assumed relationship between task design features and the actual level of cognitive complexity by independently measuring it.

7.2.2.4 Differences among the cognitive load measures

This section addresses issues related to measures of cognitive load themselves, including (a) their effect sizes, (b) interaction with proficiency, and (c) possibilities of using the cognitive load measures in the context of TBLT.

7.2.2.4.1 Effect sizes

This study revealed that, on the whole, the four measures of cognitive load yielded quite systematic findings for the English L2 participants – all measures identified Task 1 to be cognitively least complex and Task 4 to be cognitively most complex. With respect to these two tasks, the study indicated that the smallest effect size (mean difference) between Tasks 1 and 4 across the measures was $d = 0.34$ on reaction times to the dual-task measure and $d = -0.32$ (in Part 1) and $d = -0.39$ (in Part 2) on time estimation. These effect size values might represent, at least in this study and possibly as a baseline for future studies, a minimum amount of detectable difference in cognitive complexity between two tasks in order for them to be experienced distinctly. The patterns in effect sizes found among the four types of measures also deserve consideration. To summarize the overall trend, contrary to the dual-task measure and time estimation approach, which yielded the smallest $d$ values (indicating only small to medium effect sizes), both self-assessment measures (task difficulty and mental effort self-ratings) elicited substantial $d$ values ($d = 0.98$ in Part 1 and $d = 0.95$ in Part 2 for task difficulty, $d = 1.00$ in Part 1 and $d = 0.89$ for mental effort), indicating quite large effect sizes, approaching one standard deviation of difference between the simplest and most complex tasks. Note also that the exact
same relationship between the objective and the subjective measures was also observed for the native speaker participants (i.e., small differences on the first two measures, large differences on the self-assessment measures). While providing a degree of validation for the use of all of these measures for detecting complexity differences, these results also suggest that the effects of cognitive task complexity manipulation may tend to be exaggerated when assessed by subjective measures; that is, it may be that perceptions are influenced more easily by observable task manipulations than the actual amount of attention that is measured objectively to be devoted to the doing of tasks (measured here by the dual-task methodology and time estimation). Alternatively, it may also be the case that the measures of dual-task methodology and time estimation underestimate the effects of cognitive task complexity. Either way, it is highly advisable to continue to use multiple measures of cognitive load for triangulation purposes. As studies utilizing a variety of cognitive load measures accumulate, we should be able to compare them meta-analytically in an attempt to deepen our understandings of what they can reveal about the cognitive demands of tasks, and ultimately to validate our interpretations based on these (and possibly other) measures themselves.

7.2.2.4.2 Interaction with proficiency

The next issue that merits consideration is a potential interaction of the cognitive load measures with learner proficiency levels. Overall, clearer patterns were observed with the high proficiency group on various measures of cognitive load than with the low proficiency group. For the high proficiency group, Task 1 was unanimously (except for the time estimation measure used in Part 1) shown to have posed the lowest cognitive demands, and Task 2 and Task 4 to have posed the highest cognitive demands of all four tasks. Task 3, then, fell between these two
ends of the complexity spectrum. The order of complexity for the high proficiency group, therefore, was shown to be Task 1, followed by Task 3, and then by Tasks 2 and 4, with Task 1 being the least complex task.

The results with the low proficiency learners were more obscure, and confused possibly because of a three-way interaction among proficiency, tasks, and measures of cognitive load. The two self-rating measures revealed that cognitive complexity increased as designed: Task 1 was rated least difficult and least effort consuming, followed by Task 2 and Task 3 sequentially, and Task 4 was rated as most difficult and most effort consuming. Reaction time of the dual-task methodology and time estimation in Part 1, on the other hand, revealed a different picture in terms of the order of cognitive task complexity. Namely, Task 4 was shown to be unexpectedly less cognitively demanding than anticipated. As expected, high proficiency participants reacted to the color change most quickly when they were working on Task 1. When engaging in Task 4, high proficiency participants’ reaction time to the color change was very slow (1.37 seconds on average), and slowest among the four tasks. Given this trend with high proficiency participants, we would expect that low proficiency participants would have reacted to the color change (when working on Task 4) extremely slowly. Surprisingly, however, low proficiency participants reacted to the color change quite quickly when working on Task 4, in fact the quickest among the four tasks. A similar trend was observed with the time estimation measure (used in Study Part 1): the ratio of estimated time to actual time on task for low proficiency participants decreased sequentially from Task 1 up to Task 3; however, Task 4 was found to have elicited a higher ratio than Task 3.
It could be speculated that Task 4 was cognitively overwhelming for low proficiency participants (especially in Study Part 1, with the secondary task of reacting to the color change simultaneously while engaging in the narrative task), and that they may have given up on or tuned out from doing the primary, story-telling task. In fact, low proficiency participants’ explanatory comments for their task difficulty and mental effort ratings revealed that some were paying more attention to the secondary, color detection task or having a hard time concentrating on the primary, narrative task while engaging in Task 4:

#4 (Part 1) “A”の文字に気をとられていた [I was preoccupied by the letter A (the color change).]

#5 (Part 1) 考えていてもその間ずっと色の課題が続いているので、どちらにも集中しなければならず、集中して考えるのが困難であった [While I was thinking of a story, the color detection task was going on. So, I had to focus on both of them, and I had a hard time concentrating on thinking.]

By comparison, high proficiency participants seem to have been heavily engaged in Task 4, and their engagement in the task prevented them from paying attention to the color change:

#19 (Part 1) Aの文字に注目するのがよりできなくなってきた [It has become more difficult to pay attention to the letter A (the color change).]

# 34 (Part 1) 今回はストーリーに必死で A の色は見れていなかったかもしれない [This time, I tried so hard on the story that I may not have been able to see the color of A.]

# 47 (Part 1) 絵ばかり気をとられて、A のことを気にしなかったので、あまり space ボタンは押せなかった [I was paying most of my attention to the pictures, and I didn’t care much about A (the color change), so I couldn’t hit the space key so much.]
Similarly, overall, high proficiency participants perceived the narrative tasks to be less cognitively complex than did low proficiency participants, and yet they likely devoted as much or even more attentional capacity to doing the four tasks, as compared with their counterparts. The dual-task methodology indicated that, while high proficiency participants perceived the narrative tasks to be easier to handle or cognitively less demanding (demonstrated by their lower task difficulty and mental effort ratings), in reality they were engaging more heavily in narrating the stories and devoting more attentional resources to doing the tasks, especially on Task 2 and Task 4 (seen in their slower response rates to the color change), than their counterparts.

These findings suggest that although the dual-task methodology, time estimation, and task difficulty ratings have so far all been used as measures of what has been portrayed to be a singular construct of ‘cognitive task complexity’ in the field of TBLT, they are likely measuring distinct aspects of cognitive demands (if not distinct phenomena). When using measures of cognitive load, therefore, researchers should keep in mind the kinds of inferences that are warranted for each measure and choose appropriate measures of cognitive load (and interpret them with care). For the same reason, it is important to choose measures that tap into distinct aspects of cognitive demands when triangulating interpretations about cognitive task complexity.

The results of the present study also suggest that, when measuring cognitive complexity of a task that may be beyond the capacity of the learners, the dual-task methodology and time estimation may not produce accurate results. Both of these measures assume participants’ full engagement in the primary task at hand, and the violation of this assumption may result in inaccurate interpretations. Therefore, researchers should keep in mind that the dual-task
methodology and time estimation measures are likely to function best when learners’
engagement in the primary L2 task is guaranteed.

Last, comparing the $d$ values between the two proficiency groups for each of the four
measures, the dual-task methodology (reaction time) elicited bigger effect sizes for the high
proficiency group than for the low proficiency group (0.66 $SD$s bigger), whereas the other
measures elicited effect sizes with smaller differences between the proficiency levels. The low
effect size between T1 and T4 for the low proficiency group in the dual-task measure can
certainly be explained by the participants’ possible disengagement in the primary, story-telling
task (on Task 4) and their unexpectedly fast reaction times as a result.

The take-away point from these findings is that language proficiency not only interacts
with tasks but also with measures of cognitive load. Proficiency does seem to mediate cognitive
complexity: Participants at differing proficiency levels devoted distinct amounts of attentional
resources to doing the exact same tasks, and they perceived them differentially. More
interestingly, proficiency showed a clear interaction with two of the measures of cognitive
complexity. The dual-task methodology and time estimation functioned better – in the sense of
detecting designed differences among tasks, especially between Tasks 1 and 4 – with the high
proficiency participants than with the low proficiency counterparts, because the latter likely
failed to meet the prerequisites for these two measures, that is, full engagement in the primary
task. These techniques have so far been used with L1 speakers in the fields of cognitive
psychology and multimedia learning and, given their new introduction to the field of SLA, the
potential interaction with proficiency has not yet been highlighted or investigated. The findings
of the current study suggest that the choice of cognitive complexity measures should be made carefully, taking into account the learners’ proficiency levels.

7.2.2.4.3 Possibilities of the cognitive load measures: Do they function as intended in the context of TBLT research?

Before closing the discussion for the first section, the measures themselves bear closer consideration, especially regarding their potential contribution in research on TBLT.

7.2.2.4.3.1 Dual-task methodology

This study has provided some evidence supporting the usefulness of the dual-task methodology in measuring cognitive load of oral tasks typically used in the field of TBLT. This measure proved to be particularly useful in understanding the potential difference between the actual cognitive demands devoted by participants versus perceived mental effort or difficulty. This was the case in Study Part 1, where the dual-task methodology (and time estimation) exhibited an interaction with learners’ L2 proficiency and revealed that, while the low proficiency participants perceived Task 4 to be particularly difficult and mental effort consuming, they devoted fewer attentional resources to doing the task than did the high proficiency learners (who perceived the tasks to be less difficult/effort consuming). As an objective, direct measure of cognitive load, therefore, the dual-task methodology appears to have successfully revealed a differential amount of attentional resources actually devoted to doing the story-telling tasks by participants with differing proficiency levels, as opposed to perceived cognitive load (measured by the self-assessment measures). Note, however, that this interaction among the dual-task methodology, tasks, and learner L2 proficiency has likely been caused or at least exacerbated by the inevitable challenge of this cognitive load measure, namely the
secondary, color detection task. Task 4, the measured-to-be most complex task, together with the color detection task, perhaps overwhelmed the low proficiency participants and distracted their attention from the primary, story-telling task. The fact that Task 4 indeed elicited the longest, negative difference on the time estimation measure used with the low proficiency participants in Part 2 (without the color detection task) gives partial support to this speculation. Hence, the dual-task measure of cognitive load may be less effective, and therefore should only be employed and interpreted with caution, for participants with low L2 proficiency levels whose full engagement with the primary task is harder to guarantee.

Furthermore, the dual-task methodology requires some careful planning to make it work as intended. First, inserting a practice session prior to the main narrative tasks, where participants could experience what it is like to engage in the color detection task while telling a narrative story in English, turned out to be crucial. Participants’ open-ended responses revealed that many were nervous in the practice session because they had not experienced doing such narrative tasks in English and reacting to the color change simultaneously. However, by experiencing the whole process during the practice session, they felt that they could engage in the experiment more calmly and devote their attentional resources more effectively. These comments suggest the value of the practice task. Additionally, given the complex nature of the dual-task methodology, the instructions on how to engage in primary and secondary tasks must be presented in a way that facilitates and ensures participants’ understandings. Ideas (implemented in this experiment) include (a) recruiting participants who shared a single L1 so that task instructions can be given in their L1 to facilitate their understandings of task procedures, and (b) monitoring participants throughout the experiment.
Although the dual-task measure yielded one of the smallest effect sizes among the four measures (i.e., 0.34), it is important to point out that this effect size is larger than the effect size found for the same measure used in Révész et al. (2014), where Cohen’s $d$ for the dual-task measure was found to be 0.15 for the English as a second language learners (and the difference was found to be statistically significant). In fact, in the current study, if we only take into account the high proficiency participants (i.e., the population who most likely met the prerequisites for this measure, namely full engagement in the primary task), the effect size is as large as 0.58. In addition, comparing the mean reaction times, learners in Révész et al.’s study responded to the color change very fast on both simple and complex tasks (i.e., 0.354 seconds for the simple task, 0.348 seconds for the complex task), whereas the participants in the current study responded to the color change a lot slower in general (i.e., 1.18 seconds for the simple task, 1.33 for the complex task). These differences likely stem from our distinct choices of a secondary task. The background color detection task used in Révész et al.’s study was likely too salient for the participants, and therefore not sensitive enough to detect the fine difference between the two tasks, leading to a very small effect size and participants’ extremely fast responses to the color change (cf. Schoor et al., 2012). These findings, therefore, emphasize the importance of the choice of a secondary task in adopting the dual-task methodology.

All in all, the dual-task methodology may have good potential as an objective measure of cognitive load in the field of TBLT, especially if it can be operationalized carefully, if the secondary task can be chosen appropriately, and if learners’ engagement in the primary L2 task can be guaranteed.
7.2.2.4.3.2 Time estimation

Overall, the measure of time estimation in the current study yielded patterns in the degree of cognitive complexity similar to the dual-task methodology. It is also noteworthy that the dual-task methodology and time estimation yielded very similar $d$-values ($d = .34$ for the dual-task methodology and $d = −0.32$ in Part 1 and $d = −0.39$ in Part 2 for time estimation). This consistency in findings between the dual-task methodology and time estimation may shed new light on the relationship between them. As suggested by Block et al. (2010), the prospective time estimation approach may function very similarly to the dual-task methodology, measuring the actual, as opposed to the perceived or subjective, degree of cognitive demands of a task at hand. The use of the dual-task methodology and time estimation as measures of cognitive load has just begun in the field of TBLT, and it is difficult, if not impossible, to reach any definitive conclusions about the relationships between these measures at this point. To advance our understandings of such relationships, then, more studies are needed that utilize both the dual-task methodology and the prospective time estimation approach, so that we can compare results and reach a firmer conclusion about these measures of cognitive load.

When we analyze time estimation by proficiency, however, the measure elicited ambiguous results in terms of the level of cognitive complexity of the four narrative tasks for both low and high proficiency groups. For the high proficiency participants, the measure showed in Part 1 that the order of complexity was Task 3, followed by Task 1, Task 2, and Task 4 sequentially; in Part 2, then, it showed the order of complexity to be Task 1, Task 3, Task 2, and Task 4, with Task 1 being the simplest task. For the low proficiency participants, it elicited increasingly longer negative differences between their estimated time and the actual time spent,
up to Task 3 in Part 1. In Part 2, Task 3 was shown to be the least complex task, followed by Task 2, Task 1, and Task 4 sequentially. Some of these results (i.e., in Part 1 for the high proficiency group, and in Part 2 for the low proficiency group) are radically different from what the other measures of cognitive load have found and are hard to explain. At this point, I would argue that more studies need to investigate the relationship between the time estimation measure and learners’ L2 proficiency to confirm that it indeed works as intended with learners’ of distinct proficiency levels. If time estimation is to be used as a valid measure of cognitive load for the diverse types of participants that regularly participate in the context of TBLT research, it is imperative to operationalize the measure carefully and choose the appropriate paradigm that predicts the direction of estimation correctly (i.e., whether it should decrease or increase as cognitive demands of a task increase), especially given that the relationship between cognitive complexity and the estimated time duration changes depending on whether or not participants are aware of the time estimation requirement.

7.2.2.4.3.3 Task difficulty and mental effort self-assessment

The results of the two self-rating measures revealed similar trends across the four tasks. However, the additional open-ended questions about task difficulty and mental effort elicited sometimes distinct responses. Participants’ effort to paraphrase unfamiliar expressions, strategies to deal with a given task linguistically, and performance factors were by far more closely related to perceived mental effort exerted than to task difficulty, whereas the simple task input unexpectedly made the task difficult for some participants but not necessarily effort-consuming as a result. Based on these observations, I would argue that the two measures played different roles in the current study: Mental effort ratings seemed to have been more sensitive to the effort
or attention actually devoted to each task, whereas task difficulty ratings revealed the perceptual difficulty of each task, regardless of how much mental effort it required participants to actually invest. These arguments have also been made by Brünken et al. (2003) and van Gog and Paas (2008) in the context of multimedia learning. Hence, these two measures should be used for different purposes: The task difficulty ratings to explore learners’ perceived difficulty of a task, and the mental effort ratings to gauge the perceived amount of attention or effort the individual has put into doing the task. Lastly, it has to be emphasized that, as we have seen here already, the addition of the open-ended questions to the simple ratings of task difficulty and mental effort provided a tremendous amount of new information and insights into the relationship between task design and cognitive load from learners’ perspectives. Without their responses to these open-ended questions, it would have been impossible to understand why some tasks were measured to be cognitively more complex than others, why the level of cognitive complexity is different for participants’ with distinct proficiency levels, and what makes a task cognitively more or less complex in general (especially in relation with facilitative and extraneous cognitive load).

7.2.2.4.3.4 Data triangulation

Finally, the value of data triangulation, especially in this type of study, should be underscored. Because the independent measures of cognitive load (i.e., dual-task methodology, time estimation, self-assessment ratings) have not been validated in the context of TBLT research, and the meaning of any single measure may therefore not yet be entirely clear, this study benefited from the use of multiple measures. The measures used here supplemented each other in that they allowed a fuller and more accurate understanding of the phenomena at play in
the oral narrative tasks. Additionally, Cohen’s $d$ (effect sizes) differed largely across the four measures, ranging from $d = -0.32$ (time estimation in Study Part 1) to $d = 1.00$ (mental effort ratings in Part 1), indicating distinct sensitivity of measures to task complexity manipulation. Given these observations, I highly recommend the use of multiple measures of cognitive load for future studies on cognitive task complexity, including a direct measure (e.g., dual-task methodology, possibly time estimation) and a subjective measure of cognitive load (e.g., self-assessment), at least until we know more about the nature of these cognitive load measures and how to interpret the results they elicit. As argued earlier, the addition of the open-ended questions are imperative in understanding the nature of cognitive task complexity and the relationship between task design features and their putative effects on cognitive complexity.

7.3 Effects of measured cognitive complexity on task performance

Following the format of the previous section, in what follows, I will firstly summarize the overall results regarding the effects of measured cognitive complexity on the native speakers’ as well as English L2 speakers’ task performances. We will then move on to discussing three important issues that have emerged from these overall patterns.

7.3.1 Summary of overall patterns

To reiterate, the two relevant research questions asked:

1. What are the effects of increased cognitive task complexity on task performance in terms of syntactic complexity, accuracy, lexical variety, and fluency indices, for both the native and L2 speakers of English?
2. Does L2 proficiency matter in terms of the effects of cognitive complexity on English-L2 participants’ task performances?
As summarized in Chapter 6, the basic answers to these research questions are as follows. For the native speakers of English, the four tasks, which were measured to have posed relatively similar cognitive demands (for this particular population), elicited task performances that were qualitatively different. Task 3 and Task 4 elicited performances that were clearly more syntactically complex than the other two tasks. Task 4 elicited the greatest amount of speech, whereas Task 1 elicited the least amount but the most fluent speech of all four tasks. Recall, however, that the four tasks elicited only marginally different performances in terms of lexical variety regardless of their complexity levels. By contrast, for the English L2 participants, Task 3 – the measured to be second simplest task overall – elicited the best performance of all four tasks. That is, it elicited the syntactically most complex, most accurate, and lexically most diverse language production. Looking at the other three tasks, Task 1, the measured-to-be simplest task, elicited syntactically and lexically simplified language production, whereas Task 4, the measured-to-be most complex task elicited the greatest number of words and syntactically quite complex speech (although perhaps not quite as complex as Task 3). Lastly, Task 2 – the measured-to-be second most complex task – elicited overall the worst task performances in that it elicited syntactically and lexically simplified language production, and it did not lead to any advantages on either accuracy or fluency.

Analyzing English L2 participants’ task performances by proficiency groups, although Task 3 was measured to have posed distinct degrees of cognitive complexity to the low and high proficiency participants, especially in relation with Task 2, it elicited the best performances (syntactically most complex and lexically most varied speech with no detrimental effects on accuracy or fluency in terms of speech rate) from both proficiency groups in general. For the
other tasks as well, similar trends were observed across proficiency groups. Task 1 and Task 2 elicited syntactically and lexically least complex language production, and neither accuracy nor fluency (speech rate) were positively affected. By contrast, Task 4 elicited relatively (if not the most) syntactically complex utterances with no obvious disadvantage on accuracy or fluency (speech rate). A clear interaction effect was found for the measure of the total number of words produced: The tasks elicited increasingly greater numbers of words as the number of elements or characters involved in the story went up, but only from the high proficiency group. Additionally, there was a suggestive interaction effect on syntactic complexity for Task 3. That is, Task 3 elicited syntactically the most complex speech of all four tasks from the high proficiency group, but that was not the case with the low proficiency group. Hence, it seems that only the high proficiency group could well respond to and take advantage of the demands posed by this particular task, in terms of syntactic complexity. Note also that Task 2 and Task 4, that were measured to have posed similarly high cognitive complexity to the high proficiency participants, elicited radically different task performances. That is, Task 2 elicited syntactically and lexically simplified language production and the second shortest amount of speech with no advantage on either accuracy or fluency, whereas Task 4 elicited syntactically complex, and the longest, language production with no disadvantage on accuracy or fluency.

7.3.2 Further discussion

Three important issues emerge from the overall patterns reported above and merit further attention: (a) why Task 3 elicited the best performance, (b) proficiency effects, and (c) consideration of the contribution of each of the performance measures per se.
7.3.2.1 Why did Task 3 elicit the best performance of the four tasks?

As predicted by Robinson’s (2001a) Cognition Hypothesis (CH), Task 4 with the most number of characters (and hence the highest functional and linguistic demands of distinguishing among similar characters) was measured to have posed the highest cognitive demands overall. As for the effects on L2 task performance, it would be predicted that Task 4 would also elicit the best performance among the four narrative tasks used in the current study, at least in terms of linguistic complexity and accuracy outcomes. Although Task 4 did result in apparent effects on syntactic complexity, it was not the case that it resulted in the best performance among the four tasks. Rather, it was Task 3, the measured-to-be second simplest task, that elicited on the whole the best performance of all four tasks. In an attempt to probe this potentially unexpected result, the notion of facilitative versus extraneous cognitive load will be visited in relation with observed effects on task performance here. As summarized earlier, Task 1 and Task 3 seem to have posed low extraneous cognitive demands in that their storylines were easy to figure out, whereas Task 2 and Task 4 posed high extraneous cognitive demands. In terms of facilitative cognitive load, Task 1 posed the lowest demands and Task 4 the highest, with Tasks 2 and 3 falling between the two tasks. Task 1, then, posed lowest extraneous as well as facilitative cognitive load, and as a result, it was overall cognitively simplest – exactly the result obtained by all measures of cognitive load. Low extraneous load posed by Task 1 would have freed up some attentional resources to be devoted to task performance; however, with a lack of facilitative cognitive load, that is, a type of demand that plays an important role in encouraging learners to push the limits of their interlanguage and improve performance, Task 1 only ended up eliciting the shortest speech overall, but with no other effects. Task 2, then, posed very high extraneous
cognitive load and relatively medium facilitative cognitive load. Ideally, the somewhat higher facilitative cognitive load would have encouraged learners to pay additional attention to linguistic forms and focus on improved task performance; however, it might have been the case that the high extraneous cognitive load had already taken up most of the attentional resources and there was very little left to be devoted to linguistic aspects of task performance. Hence, it elicited the worst performances of all four tasks. Moving on to Task 4, it posed high cognitive demands, yet both in terms of extraneous and facilitative dimensions, and as a result, the latter likely created the need to produce longer utterances (perhaps to refer to all different characters and what they are doing) and syntactically complex sentences (perhaps to distinguish among similar characters and relate each others’ actions). Nonetheless, high extraneous cognitive load may have prevented the English L2 participants from paying as much attention to linguistic elements as they might have wanted to, and certainly less so than they did in Task 3. Lastly, then, Task 3 likely enjoyed the best combination of quite low extraneous load and relatively medium facilitative load. This combination allowed the English L2 speakers to devote the majority of their attentional resources to linguistically improved task performance, which, perhaps in turn, led to the syntactically most complex, most accurate, and lexically most diverse language production. This relationship between extraneous and facilitative cognitive load and task performance may also explain why Task 3 elicited the best performance from both low and high proficiency groups.

Generally, in L2 task complexity research, the role of extraneous cognitive load has been largely ignored, not only in relation with the levels of cognitive load per se, but with its putative effects on task performance. The idea of extraneous cognitive load is similar to Robinson’s
(2001a) resource-dispersing dimensions of task complexity in that both touch upon the issue of where learners’ attention is being allocated, whether to linguistic aspects of task performance or not. However, as discussed earlier, the possibility that a task design feature that is categorized as a resource-directing variable (or what I have been calling a facilitative type of cognitive load) can also yield a resource-dispersing type of complexity had been left out from the picture of the design-complexity relationship, and how extraneous and facilitative types of cognitive load influence task performance in intertwined, complicated ways had not yet been explored. Note, however, that the speculations made above about the relationship between extraneous and facilitative cognitive load and L2 task performance are in line with an observation made by Skehan (2009). Thus, simultaneous attention to linguistic accuracy and complexity can be made possible if we reduce the extraneous cognitive load, for example of figuring out a storyline (by making the task structure clear), and yet pose a high facilitative demand of integrating multiple pieces of information (such as foreground and background information, or such as the actions of multiple characters). Additionally, this idea of increasing facilitative demands, which can perhaps be related to germane cognitive load in Sweller’s (1994) model, while reducing the extraneous cognitive load, seems to be compatible with Cognitive Load Theory’s ultimate goal of optimizing cognitive load of a given task for better performance (and learning). Taken together, the results of the current study suggest that when choosing tasks for investigation, we need to be aware of not only the overall level of cognitive load imposed, but also the types of cognitive demand and their likely differential effects on task performance.
7.3.2.2 Proficiency effects

Across all performance measures, except the two lexical variety measures, clear proficiency effects were observed. That is, the high proficiency participants produced better performance on all aspects of CALF, but lexical variety, than did the low proficiency participants. Looking at the relationship between L2 proficiency and task performance, the four tasks with differing levels and types of cognitive complexity had very similar relative effects on task performances of both low and high proficiency groups. Interaction effects between learners’ L2 proficiency and tasks were found for the measure of the total number of words produced and for all measures of syntactic complexity. Thus, the high proficiency participants likely had enough language skills to take advantage of high facilitative cognitive load posed by Tasks 3 and 4 (in the form of pushing learners to refer to all characters and their actions, and to talk about what all is happening in each story), whereas the low proficiency participants did not. This finding underscores the argument that L2 proficiency cannot be neglected in understanding the relationship between cognitive task complexity and L2 task performance, as previous studies (e.g., Ishikawa, 2006) have illustrated.

7.3.2.3 Differences among performance measures

Finally, the performance measures used in the present study deserve further attention, in terms of what they were able to provide. Looking at the data for the English L2 participants, some performance measures elicited larger $d$-values than did other measures. In both study parts, MLT, MLC, and total number of words elicited large, positive effect sizes ranging from $d = .68$ (total number of words in Study Part 2) to $d = 1.00$ (MLC in Part 1). The other performance measures (except for TTR, which will be discussed later) yielded much smaller effect sizes (all
below $d = .20$). These findings suggest that syntactic complexity and fluency (amount rather than speed) of task performance benefitted the most by increased cognitive complexity overall.

Turning to the first CALF dimension of linguistic complexity, the syntactic complexity measure that is most commonly used in L2 task complexity research, clauses per T-unit (a measure of syntactic subordination), elicited radically different results from the other syntactic complexity measures used in this study: (a) MLT (global) and (b) MLC (phrasal-level). For the native speaker participants, MLT and MLC showed that Task 1 elicited the syntactically least complex language production, while Task 4 elicited the syntactically most complex language production. By contrast, the measure of clauses per T-unit revealed that Task 2 and Task 4 elicited the fewest clauses per T-unit, while Task 1 and Task 3 elicited more clauses per T-unit than the other two tasks. For the English-L2 speakers, although the picture is less clear, the same basic pattern is also observable. MLT and MLC measures revealed that Task 1 and Task 4 elicited substantially different degrees of syntactic complexity, with Task 4 eliciting more complex sentences, while clauses per T-unit indicated that Task 1 and Task 4 elicited very similar amounts of subordination. These findings underscore the danger of relying solely on the subordination-based syntactic complexity measure (i.e., clauses per T-unit) when investigating the effects of increased cognitive complexity on learner or native speaker performance.

In terms of accuracy, both study parts found minimal differences in the ratio of error-free T-units among the four tasks. It may be the case that the measure itself was an issue. As reviewed in Chapter 1, Robinson’s (2001b) CH predicts that a cognitively complex task – for example, the narrative task with nine characters designed for this study – has greater potential to draw learners’ attention to task-essential, communicatively non-redundant language forms,
namely noun modifiers, rather than to a variety of possible language forms. The use of a specific accuracy measure, therefore, may help us capture this potential local accuracy difference across tasks in future investigations.

Moving on to the lexical measures, the two measures elicited largely distinct outcomes. The $D$ measure detected only negligible differences in lexical variety between Task 1, the measured-to-be simplest task, and Task 4, the measured-to-be most complex task, leading to small effect sizes of $d = .12$ in Study Part 1 and $d = .02$ in Part 2. However, Task 3 was shown to have elicited a substantially larger $D$ score than the other four tasks. To this end, the effect sizes between Task 3 and Task 2 (that elicited the smallest $D$ scores) were found to be $d = .62$ and $d = .71$, indicating medium to large effect sizes. On the other hand, TTR yielded large, negative effect sizes between Task 1 and Task 4, $d = −0.67$ in Part 1 and $d = −1.12$ in Part 2, hence in favor of the *simple* task – a radically different result from the $D$ measure. As pointed out in previous literature, the TTR scores in the current study were likely affected by substantially different length of speech elicited by each task (with differing cognitive complexity). Therefore, caution must be taken in interpreting the TTR scores if data involve varying lengths of speech, and for such data $D$ may be a better measure of lexical variety to use.

Lastly, the two fluency measures also produced largely different results. The total number of words (amount) elicited large, positive effect sizes between Task 1 and Task 4 ($d = .94$ in Part 1, $d = .68$ in Part 2), whereas words per second (speed) detected only small differences among the four tasks (leading to small effect sizes between Task 1 and Task 4, $d = −.18$ in Part 1 and $d = −.06$ in Part 2). These results indicate the importance of using fluency measures that tap into different aspects of the construct.
CHAPTER VIII: CONCLUSIONS

8.1 Introduction

This last chapter of the dissertation explores limitations, provides concluding remarks by offering answers to the ‘so what?’ question and by offering theoretical and practical implications, and finally discusses ideas for future studies.

8.2 Limitations

Limitations of the present dissertation at the same time open up suggestions for future research. Here, I will offer three main limitations to the current dissertation: (a) alternatives for the measurement of performance outcomes in task complexity research, (b) sample size, and (c) measures of cognitive load. Firstly, in addition to the CALF measures used in the current study (and also widely used in other task complexity studies), there are several other performance measures that may give us new insights into the relationship between cognitive task complexity and L2 learners’ task performance. First and foremost, assessment of task performance quality (in terms of content, not language) merits attention. In the field of SLA, although uncommon, several studies have assessed L2 learners’ task performance based on their degrees of success in completing a given task, rather than solely on their linguistic performance. Brown et al. (2002), for example, developed task-based performance assessment instruments for English as a second language learners and rating rubrics for these instruments. These rubrics involved ratings based on real-world criteria where test-takers were evaluated according to elements that were thought to be essential for successful accomplishment of a given task (e.g., accuracy of performance content, appropriate pragmatics and style), as well as holistic performance ratings on other aspects of performance (e.g., code command, cognitive operations, communicative adaptation).
As illustrated in this example, when creating rating rubrics, it is important to consider and carefully choose performance aspects for assessment that are relevant and indeed essential for successful completion of the actual tasks we are asking learners to do. In the context of this dissertation, it would be interesting to assess: (a) to what extent the six picture frames were included in the story, (b) to what extent the order of the pictures was clear to the listener, and (c) to what extent the story was creative – the three performance areas that the participants were explicitly told would be assessed.

The next measures that might give us new insights are form-specific complexity and accuracy measures. Robinson’s CH predicts that a complex task along the resource-directing dimensions, for example number of elements, will elicit linguistically more complex and more accurate language production (as compared to a simple task) because such a task likely draws learners attention to task-relevant, communicatively non-redundant language (e.g., noun modifiers in the case of number of elements) (2001b), and importantly not to any possible language forms. Given that, it would make sense to use form-specific complexity and accuracy measures (although it begs a question of comparability – whether we can make a fair comparison between a complex task that in theory elicits more opportunities for the use of task-relevant language forms and a simple task that may not elicit any opportunity at all, and whether we can compare across studies that used tailored form-specific measures of complexity and accuracy).

Secondly, although overall this dissertation involved an n-size of 61 native speakers of English and 119 English L2 speakers, when it came down to the proficiency analyses, they involved relatively small sample sizes (ranging from n = 13 to n = 20). Thus, the proficiency analyses were only exploratory and their findings should be interpreted with some caution. These
rather small sample sizes stemmed partially from an attempt to make sure that the low and high
proficiency groups were maximally different in their proficiency levels (hence participants in the
middle needed to be left out) and to make the comparison across the two study parts possible
(hence the cut-off scores needed to be kept identical in the two study parts). These
operationalizations were important methodologically, but they did have some consequences.
Also, assuming a normal distribution of Japanese-L1, English learner’s proficiency levels (which
indeed was seen in the distributions of the proficiency test scores in the current study), it was
difficulty to find participants who would fall in the lower and higher ends. Having said that,
however, in order to advance our understandings of how the relationship between task design,
cognitive complexity, and task performance interacts with L2 proficiency, studies with more
substantial n-sizes of distinct proficiency groups are essential.

Lastly, given the new introduction of many of the cognitive load measures used in the
current study to TBLT, interpretation of data elicited by these measures was not easy and
straightforward. In particular, the data elicited by the time estimation measure, especially in
relation with proficiency, were in some places uninterpretable. Many of the cognitive load
measures available have been developed and utilized in other fields with native speakers in mind.
Thus, caution has to be taken when using such measures with a distinct population of L2
speakers with L2 communication tasks that are less likely to have been used in the original
fields. I would hope that more studies will use these new measures of cognitive load and
investigate their appropriacy for issues in the field of TBLT.
8.3. Concluding remarks

8.3.1 Importance of the current study: So what?

This dissertation sought to: (a) validate the assumed relationship between task design and cognitive complexity of four oral narrative tasks, by using multiple independent measures of cognitive load; (b) explore what factors made the four tasks cognitively more or less complex; (c) investigate the effects of measured cognitive complexity on participants’ task performance; and (d) explore the role of proficiency. One of the most important findings of this study is that (for the English L2 participants) the designed-to-be simplest task (Task 1) and the designed-to-be most complex task (Task 4) were shown to be substantially different in terms of cognitive complexity unanimously by all measures of cognitive load, while the difference in cognitive load between Task 2 and Task 3, that were designed to pose distinct cognitive demands, was found to be generally small. In addition, Task 2 was shown to be cognitively more complex than Task 3 by all measures of cognitive load (except for time estimation in Part 2). These findings suggest that, when testing the effects of cognitive task complexity, the difference in design features (e.g., number of elements) should be kept as large as possible to make the actual cognitive demands posed by these tasks meaningfully distinct. They also underscore the importance of validating the assumptions about cognitive task complexity before investigating the effects of increased cognitive complexity on participants’ task performance and language learning.

Another important finding comes from participants’ open-ended comments. These comments revealed a number of new insights into the relationship between task design features and cognitive complexity. Among them, one of the most noteworthy is that (a) conceptual input (including the manipulated task design feature of number of elements), (b) code complexity, and
(c) performance factors affected English-L2 participants’ perceived difficulty and mental effort exerted. The distinction between extraneous and facilitative types of cognitive load is another important observation to highlight. Task 1, the measured-to-be the simplest task, was interpreted to have posed low extraneous and facilitative cognitive load, whereas Task 4, the measured-to-be the most complex task, seemed to have posed high extraneous and facilitative load. The middle two tasks (i.e., Tasks 2 and 3) both posed relatively medium facilitative load, but Task 2 posed much higher extraneous load than did Task 3. Hence, it is essential to take into account factors other than a manipulated design feature and be aware of the two types of cognitive load—extraneous and facilitative—when designing experimental tasks. These findings that point to the complicated relationship between task design and its cognitive complexity also give additional support to the claim that the level and type of complexity need to be verified rather than assumed. Critically, these findings would not have been possible to obtain if the study had not adapted the questionnaire with the inclusion of open-ended questions. Eliciting learners’ reactions to a task at hand is crucial in understanding in depth the relationship between task design and cognitive demands imposed.

Similarly, methodologically, it is important to emphasize the value of using multiple measures of cognitive load in TBLT research. Understandings of cognitive complexity in the present study, including its relationship with L2 proficiency, would not have been possible with use of just one measure, and different measures of cognitive load were shown to possibly tap into distinct aspects of cognitive demands. Hence, by demonstrating that the use of multiple measures of cognitive load is, in fact, feasible, the study also proposes that such an approach become
central for TBLT research in order to enable the empirical validation of assumptions about the
cognitive complexity of L2 tasks.

Turning to the effects of cognitive complexity on task performance, it was Task 3, the
measured-to-be second simplest task, that elicited the best performance overall. It was speculated
that a combination of low extraneous cognitive load and high facilitative load that Task 3 posed
could account for this rather unexpected result. This relationship between extraneous and
facilitative types of cognitive load and task performance is particularly important in
understanding what exact aspects of task design or cognitive complexity draw or distract
attention from linguistic aspects and hence improve or deteriorate task performance.

Lastly, L2 proficiency seems to have interacted not only with tasks but also with measures
of cognitive load. Overall, high proficiency participants perceived the narrative tasks to be less
cognitively complex than did low proficiency participants, and yet they likely devoted as much
or more attentional capacity in doing the four tasks, as compared with their counterparts.
Additionally, Task 1 was shown to be cognitively simplest and Task 4 to be cognitively most
complex for both proficiency groups, but to varying degrees. Lastly, Task 2 was consistently
shown to pose higher cognitive demands than Task 3 (despite their cognitive complexity
manipulation), but only for high proficiency participants. These findings emphasize the
importance of language proficiency in interpreting cognitive task complexity. At a minimum,
without considering learners’ language proficiency, interpretations will be superficial and
applicable only to certain learner groups.

Turning to the relationship between L2 proficiency and task performance, generally
speaking, the high proficiency participants produced syntactically more complex, more accurate,
and more fluent speech than did the low proficiency counterparts (with lexical variety affected minimally by proficiency). Importantly, however, the measure of the total number of word showed an interaction effect with L2 proficiency. That is, the high proficiency participants had enough language ability to take advantage of the increased (facilitative) cognitive load and produced progressively longer stories, whereas the low proficiency participants produced similar number of words for each task. In addition, there was a suggestive interaction effect with proficiency on the syntactic complexity measures for Task 3; thus the task elicited the most complex utterances from the high proficiency learners, but not from the low proficiency counterparts. Hence, L2 proficiency cannot be neglected in understanding the relationship between cognitive task complexity and task performance, either.

Lastly and most critically, different cognitive load measures reveal different dimensions of task-learner relationships: The dual-task methodology and time estimation did not function as intended with the low proficiency learners, given their potentially low engagement with the primary, story-telling task. When choosing appropriate measures of cognitive load, therefore, we need to take into account learner proficiency and consider whether the given learners will be able to fully engage in the primary tasks at hand.

8.3.2 Implications

In this section, several key implications are drawn for future research into the effects of cognitive task complexity and in relation to second language teaching.

8.3.2.1 Implications for researching the effects of cognitive task complexity

This study offers a handful of implications for future studies that wish to investigate the effects of cognitive task complexity on task performance, interaction, and learning. First, in
designing experimental tasks, the difference in task design features should be kept large if we wish to ensure that the degree of cognitive complexity among tasks is meaningfully distinct. In the current dissertation, one of the most obvious trends was that across measures of cognitive load, for any proficiency group, Task 1 with one character and Task 4 with nine characters were unanimously shown to have posed distinct degrees of cognitive complexity for the English-L2 participants, regardless of other design factors. Note here that this observation was possible to make because this study used four tasks that gradually increased in complexity by design, rather than the typical dichotomous design of simple versus complex tasks. I would add that this advice should be particularly useful for researchers who are interested in investigating the effects of measured cognitive complexity, regardless of what task features exactly bring about the difference in complexity. If a researcher is interested in investigating the relationship between a particular task design feature and the degree of cognitive complexity such a task feature poses (i.e., to empirically verify the Triadic Componential Framework), s/he should be careful of changing solely the feature under investigation (and also careful to measure the level of complexity by multiple independent measures of cognitive load).

Second, in designing experimental tasks, it is imperative to keep in mind that various, sometimes unexpected, factors likely influence the level of cognitive complexity of tasks at hand. Regardless of what theory one intends to test, therefore, it will be wise to take into consideration the possibility of factors other than those manipulated. The current dissertation demonstrated that three overarching categories of factors influenced the levels of cognitive load of the four oral narrative tasks: (a) conceptual input (including but not exclusive to the design feature, number of elements), (b) code complexity, and (c) performance factors. Additionally, it
revealed that complexity differences were caused by two distinct types of cognitive load: (a) extraneous cognitive load that was perceived to be deleterious, wasteful, and irrelevant to task performance; and (b) facilitative cognitive load that encouraged learners to push the limits of their interlanguage resources and focus on improved performance. These findings underscore the importance of taking into account various factors when designing tasks that are maximally different in their levels of cognitive complexity (or intended to be similar to create equivalent tasks). Furthermore, the current study has also suggested that the two types of cognitive load had likely differential effects on L2 learners’ task performances and that a combination of low extraneous load (e.g., in the form of straightforward input) and high facilitative load (e.g., realized by the need to refer to many characters) seems to elicit the best task performance. Given that, if we can find a way to reduce extraneous load (e.g., in a narrative task, assisting learners with figuring out who’s who in the story, deciding which characters to describe or focus on; making sure that a task does not pose excessively high code complexity), the purer effects of what Robinson (2001a) calls a resource-directing factor (e.g., of having a large number of elements) may be more precisely testable.

Third, making sure that the difference in design is large enough, and taking into account factors other than the manipulated features, will still not be enough to guarantee an actual, meaningful difference in cognitive complexity of experimental tasks. It is critical to assess the level of cognitive task complexity by multiple measures of cognitive load to ensure that the tasks are indeed meaningfully different in terms of their levels of cognitive complexity (i.e., if that is to be considered the primary causal variable in this domain of research). To this end, it is important to choose measures of cognitive load wisely. Available cognitive load measures have
distinct advantages and disadvantages for the type of research conducted in TBLT, and they likely measure distinct aspects of cognitive demands (if not distinct phenomena). In choosing measures, their relationship with participants’ proficiency level and the primary task must be taken into account as well, because some measures (e.g., dual-task methodology, time estimation) require participants’ full engagement in the primary task, and some primary tasks (e.g., online written chat) do not necessarily allow the use of a visual-based secondary task (e.g., letter color detection task) for the dual-task methodology. Finally, the use of multiple measures of cognitive load is integral to understanding what separate pieces of information gathered by different measures are telling us. This policy becomes important especially if one is trying out new measures in the field; otherwise, it will be extremely difficult to interpret the date gathered by a single, newly introduced measure.

Lastly, when choosing syntactic complexity measures, researchers should not solely rely on subordination-based measures. As suggested by Norris and Ortega (2009), different types of syntactic complexity measures (e.g., global syntactic measure, such as MLT; phrasal-level complexity measure, such as MLC; subordination-based measure, such as clauses per T-unit) seem to have tapped into distinct aspects of syntactic complexity in this study, and the measure of clauses per T-unit elicited largely differing results from the other two measures. Hence, it would be wise to use multiple measure of syntactic complexity to make sure that we do not miss existing patterns that cannot be captured by certain measures. This argument applies not only to complexity measures but other performance measures, such as accuracy, lexical variety, and fluency measures. It has to be noted, however, that it is not always the more the better. Using multiple dependent measures in vain is not only redundant but also could lead to fishing for
results that would work in advantage of the researcher (or often times fishing for statistical significance). The choice of performance measures, then, must be firmly backed up by theoretical rationales that indicate a priori what the measure should be telling us about the relationship with task complexity differences.

8.3.2.2 Implications for teaching

What implications for L2 teaching can the present study provide? Although more studies are needed to offer any conclusive advice, of course, I do believe that the study makes unique contributions to L2 pedagogy. First of all, it demonstrates the use of cognitive load measures, some of which could be used in classroom settings as well as in experiments. I would argue that subjective ratings might be most appropriate for use in L2 classrooms, given their less intrusive and efficient nature. Additionally, participants’ open-ended comments also prove to be useful in tapping into unique insights about their mental processing and perceptions on cognitive task complexity in relation to pedagogic tasks used in the classroom. For this reason, I would argue that eliciting open-ended responses from students about their perceptions of a given task enables language teachers to gain more in-depth understandings of the phenomena and the extent to which they may or may not be conducive to language performance and learning.

Information about cognitive complexity of L2 pedagogic tasks gathered by these techniques would also, potentially, help language teachers in various phases of teaching, but in particular, task design/sequencing and assessment phases. It allows language teachers to design tasks with varying degrees of cognitive complexity for a given L2 learner group, in an effort to encourage learners to pay attention to distinct aspects of language use (e.g., performance quality, syntactic complexity, lexical variety, accuracy, and fluency). It also allows us to sequence tasks
according to their complexity levels (from simple to complex, for example) when designing or revising a syllabus for a particular course. Furthermore, cognitive complexity measures, particularly the ones focused on learners’ perceptions of L2 tasks, can serve the purpose of formative assessment. In other words, they help language teachers understand who might require extra support (given their high ratings on task difficulty, for example) and what makes a task easy or difficult for certain learners (based on their open-ended comments), as well as judge what needs to be reviewed and/or taught in the subsequent lessons (in response to learners’ perceived needs expressed through their open-ended comments).

Additionally, the current study suggests that although Robinson’s (2001a) CH and in particular Triadic Componential Framework give us a good starting point for task design, the relationship between task design and its cognitive complexity is more complicated than researchers previously seem to have assumed. The cognitive complexity of a task is influenced by various factors, including, for example, storyline complexity, visual quality of task prompts, and code complexity, as well as learner factors, in addition to the factors typically investigated in studies of the Triadic Componential Framework (Robinson, 2001a) (i.e., cognitive factors). Language teachers are advised, therefore, to pay close attention to all of these factors when designing more or less complex tasks that fulfill learning needs of their students. One clear finding from the study is that presumed differences in task design may not translate into actual, cognitive or perceptual differences as learners engage with tasks. Language teachers (or test designers) should consider the salience of any differences designed into a task, erring on the side of clear, obvious differences between the task factors presumed to induce greater or lesser
demand on learners, and pay close attention to the learners themselves as they engage in the tasks.

Lastly, the current study has shown that tasks with distinct task design and cognitive complexity may offer differing learning opportunities for learners. For instance, Task 2 seemed to have drawn participants’ (high proficiency participants’ in particular) attention to various ways to express the action of fishing. In this study, this focused or increased attention did not translate into learners’ task performance, largely because they did not have enough resources to this end. In a classroom setting, however, such increased attention provides us an optimal environment to give scaffolding and teach new expressions in a meaningful context where learners are confronted with the true need to mean. When choosing tasks to be used in a classroom, therefore, it will be important to pay attention to what linguistic forms or other aspects of task performance a task is likely to draw learners’ attention to, choose tasks accordingly, and plan how one would take advantage of learners’ increased attention to various linguistic and other performance aspects.

8.3.3 Future research

To conclude this dissertation, I would like to offer some ideas for future research, with the strong hope to advance our knowledge of the domain further. The present study was concerned with university-level, Japanese-L1 speakers of English with varying proficiency levels (as well as native speakers of English), and it is an empirical question whether the findings can be replicated in other languages and with other learner populations. Especially, it would be interesting to investigate whether a measured-to-be simplest task and a measured-to-be most complex task for a particular population in a particular L2 would also be measured to be simplest
and most complex for another population in another language. So far, the relationship between task design and cognitive complexity has been portrayed as if it were universal in most research. It is an empirical question, however, whether that is actually the case. The same thing can be said about the relationship between measured cognitive complexity and its effects on L2 task performance – does it apply equally to any target language?

Another area of replication needed is related to task type, mode, and complexity manipulations. This study used an oral narrative task and manipulated task complexity by the number of elements (or characters to be more specific) involved in each task. It found that various factors, including conceptual input, code complexity, and performance factors, contributed to the level of cognitive task complexity and that complexity was caused by both facilitative and extraneous cognitive load. It would be worth investigating whether the three superordinate categories of factors would affect cognitive complexity in similar ways for tasks of distinct modes, with different types, and with its complexity manipulated differentially. Also, it would be interesting and important to examine what causes facilitative and extraneous cognitive load and how they affect L2 learners’ task performance, using tasks of various types, modes, and complexity manipulations.

Lastly, future studies need to investigate more rigorously the relationship among task design, cognitive demands, and L2 learning, rather than one-time task performance. Especially, the relationship between learners’ increased attention to linguistic and other aspects of task performance induced by certain task design features and their L2 acquisition will be of particular importance not only to the advancement of the field of TBLT per se, but also to the improvement
of second language education, the heart of TBLT. By engaging in studies along these lines, I truly believe and hope that we can keep the field moving forward.
## Appendix A

A Time Estimation, Task Difficulty, and Mental Effort Questionnaire (English Version)

<table>
<thead>
<tr>
<th>ID # ( )</th>
<th>Task # ( )</th>
</tr>
</thead>
</table>

### 1. Was the story-telling task easy? (1=Very easy, 9=Very difficult)

| Very easy | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Very difficult |

### 2. What made the story-telling task easy/difficult? Please provide details.

### 3. How long do you think your story was? (Approximately ( ) minutes and ( ) seconds)

| Very, very low mental effort | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Very, very high mental effort |

### 4. How much mental effort did you put in to complete this task? (1=Very low mental effort, 9=Very high mental effort)

| Very, very low mental effort | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Very, very high mental effort |

### 5. What made the task require low/high mental effort? Please provide details.
実験用ID番号 ( )
課題番号 ( )

<table>
<thead>
<tr>
<th>とても簡単だった</th>
<th>とても難しかった</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

1. 6枚のイラストを見てストーリーを考え、それを英語で説明するのは簡単でしたか？

2. それはなぜですか？なるべく詳しく書いてください。

| 3. あなたが話したストーリーの長さはどれくらいだったと思いますか？ |
|----------------|----------------|
| 約 ( ) 分 ( ) 秒 |

<table>
<thead>
<tr>
<th>全く使わなかった</th>
<th>カなり使った</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

4. この課題をこなすのに、どのくらいの頭を使いましたか？

5. それはなぜですか？なるべく詳しく書いてください。
The boy walked up the street. He stepped on a piece of ice. He fell (1) down but he didn’t hurt himself.

MAN AND HIS PROGRESS

Man is the only living creature that can make and use tools. He is the most teachable of living beings, earning the name of Homo sapiens. (1) ________ ever restless brain has used the (2) ________ and the wisdom of his ancestors (3) ________ improve his way of life. Since (4) ________ is able to walk and run (5) ________ his feet, his hands have always (6) ________ free to carry and to use (7) _________. Man’s hands have served him well (8) ________ his life on earth. His development, (9) ________ can be divided into three major (10) ________ life. His development, (9) ________ can be divided into three major (10) ________ life.

Up to 10,000 years ago, (12) ________ human beings lived by hunting and (13) _________. They also picked berries and fruits, (14) ________ dug for various edible roots. Most (15) ________, the men were the hunters, and (16) ________ women acted as food gatherers. Since (17) ________ women were busy with the
children, (18) ________ men handled the tools. In a (19) ________ hand, a dead branch became a (20) ________ to knock down fruit or (21) ________ for tasty roots. Sometimes, an animal (22) ________ served as a club, and a (23) ________ piece of stone, fitting comfortably into (24) ________ hand, could be used to break (25) ________ or to throw at an animal. (26) ________ stone was chipped against another until (27) ________ had a sharp edge. The primitive (28) ________ who first thought of putting a (29) ________ stone at the end of a (30) ________ made a brilliant discovery: he (31) ________ joined two things to make a (32) ________ useful tool, the spear. Flint, found (33) ________ many rocks, became a common cutting (34) ________ in the Paleolithic period of man’s (35) ________. Since no wood or bone tools (36) ________ survived, we know of this man (37) ________ his stone implements, with which he (38) ________ kill animals, cut up the meat, (39) ________ scrape the skins, as well as (40) ________ pictures on the walls of the (41) ________ where he lived during the winter.

(42) ________ the warmer seasons, man wandered on (43) ________ steppes of Europe without a fixed (44) ________, always foraging for food. Perhaps the (45) ________ carried nuts and berries in shells (46) ________ skins or even in light, woven (47) ________. Wherever they camped, the primitive people (48) ________ fires by striking flint for sparks (49) ________ using dried seeds, moss, and rotten (50) ________ for tinder. With fires that he kindled himself, man could keep wild animals away and could cook those that he killed, as well as provide warmth and light for himself.
Appendix D
Elicited Imitation Test: List of Prompts

1. I have to get a haircut.
2. The red book is on the table.
3. The streets in the city are wide.
4. He takes a shower every morning.
5. What did you say you were doing today?
6. I doubt that he knows how to drive that well.
7. After dinner I had a long peaceful nap.
8. It is possible that it will rain tomorrow.
9. The houses are very nice but too expensive.
10. The little boy whose kitten died yesterday is sad.
11. That restaurant is supposed to have very good food.
12. I want a nice big house in which my animals can live.
13. You really enjoy listening to country music, don’t you?
14. She just finished painting the inside of her apartment.
15. Cross the street at the light and then just continue straight ahead.
16. The person I’m dating has a wonderful sense of humor.
17. She only orders meat dishes and never eats vegetables.
18. I wish the price of town houses would become affordable.
19. I hope it will get warmer sooner this year than it did last year.
20. A good friend of mine always takes care of my neighbor’s three children.
21. The black cat that you fed yesterday was the one chased by the dog.
22. Before he can go outside he has to finish cleaning his room.
23. The most fun I’ve ever had was when we went to the opera.
24. The terrible thief whom the police caught was very tall and thin.
25. Would you be so kind as to hand me the book which is on the table?
26. The number of people who smoke cigars is increasing every year.
27. I don’t know if the 11:30 train has left the station yet.
28. The exam wasn’t nearly as difficult as you told me it would be.
29. There are a lot of people who don’t eat anything at all in the morning.
Appendix E
Elicited Imitation Test: Written Instructions in Japanese

繰り返し課題の説明

英語で聞こえた通りにそのまま繰り返す課題です。

・英語の文をいくつか聞いてもらいます。
・各文の後、短い休止があって、トーン（電子音）が続きます。
・聞きえた通りにそのまま英語を繰り返してください。
・トーン（電子音）の後、繰り返すのに十分な時間が与えられます。
・できるかぎり繰り返してください。
・トーン（電子音）が流れるまでは、繰り返し始めないでください。

はじめに練習があります。この練習では、日本語の文がいくつか流れていますので、
聞きえた通りにそのまま日本語で繰り返してください。
Appendix F
Elicited Imitation Test: Written Instructions in Japanese (Translation)

Repetition Task Instructions

In this task, you will be asked to repeat what you hear in **Japanese**.

- You will hear a number of sentences in **Japanese**.
- After each sentence, there will be a short pause, followed by a tone sound.
- Your task is to try to **repeat exactly what you hear in Japanese**.
- You will be given sufficient time after the tone to repeat the sentence.
- **Repeat as much as you can.**
- **Do NOT start repeating the sentence until you hear the tone sound.**

- There will be a **practice** session at the beginning. Here, you will hear several sentences in **English**, so repeat exactly what you hear in **English**.
Appendix G
Background Questionnaire for Native Speakers of English

1. ID#

2. Gender: Male / Female

3. Age: ____________

4. Educational Status
   If you are an undergraduate student, please specify your School, major(s), minor(s), and certificate(s) (if applicable).

   School: ________________________________

   Major: ________________________________   Minor: ________________________________

   Certificate: ____________________________

   If you are a graduate student, please specify your degree program and major.

   Degree: Mater’s / PhD / Other   Major: ________________________________

5. What is your first language?
Appendix H
Background Questionnaire for English-L2 Speakers

1. 实験用 ID 番号：______________________

2. 性別： 男性 / 女性

3. 年齢：______________________

4. 学年：大学 一年生 / 二年生 / 三年生 / 四年生 / 五年生以上

   大学院 修士課程 / 博士課程 その他（  ）

5. 専攻：

   学部 ________________________________

   学科 ________________________________

6. 母語は何ですか？

7. 最近受けた英語のテスト（TOEFL, IELTS, TOEIC, または英検）の点数（もしくは級）、受験年月を教えてください。複数ある場合は、すべて書いてください。

<table>
<thead>
<tr>
<th>テストの種類 (最近受けたテストに丸をしてください)</th>
<th>点数（もしくは級）</th>
<th>受験年月（例：2014年5月）</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOEFL iBT CBT PBT ITP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IELTS TOEIC 英検</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOEFL iBT CBT PBT ITP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IELTS TOEIC 英検</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOEFL iBT CBT PBT ITP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IELTS TOEIC 英検</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Questions 1 through 3 are equivalents of the same numbered questions in the background questionnaire for native speakers of English in Appendix G. Questions 4 and 5 are equivalents of Question 4, and Question 6 is an equivalent of Question 5 in the native speaker version. Lastly, Question 7 asks about learners’ self-reported scores on standardized tests, such as TOEFL, IELTS, TOEIC, and Eiken.

378
Appendix I
Consent Form for Study 1 (Native Speakers of English)
IRB #2014-0197
Title: Effects of cognitive task complexity on second language learners’ task performance
Georgetown University
Consent to Participate in Research Study

**STUDY TITLE:** Effects of cognitive task complexity on second language learners’ task performance

**PRINCIPAL INVESTIGATOR:** Shoko Sasayama **TELEPHONE:** 202-687-6213

**ADVISOR:** Dr. John M Norris

**INTRODUCTION**
You are invited to consider participating in this research study. Please take as much time as you need to make your decision. Feel free to discuss your decision with whomever you want, but remember that the decision to participate, or not to participate, is yours. If you decide that you want to participate, please sign and date where indicated at the end of this form. If you have any questions, please ask the principal investigator (Shoko Sasayama).

**PURPOSE**
The purpose of this study is to investigate the effects of task complexity on second language learners’ task performance.

**STUDY PLAN**
You are being asked to take part in this study because your first language is English. You will be asked to tell a story in English based on five sets of pictures (your story will be audio-recorded) and fill out a questionnaire on your perceptions of each task. Half of you will also be asked to interact with a computer. Subsequently, you will be asked to fill out a background questionnaire. You will then be invited to participate in a post-task interview. The session should take you approximately 45 minutes to complete.

**RISKS**
There are no known risks associated with participating in this study.

**BENEFITS**
There may be no direct benefit to you in participating in this study. However, your participation will contribute to advancement of the field of Second Language Acquisition by allowing us to investigate the effects of task design on second language learners’ task performance.

**CONFIDENTIALITY**
Every effort will be made to keep any information collected from you confidential. However, it is impossible to guarantee absolute confidentiality. In order to keep information about you safe, the digital data will be kept password-locked in a password-protected personal computer and the paper-based data will be kept in a locked box at the researcher’s office at all times. Only the principal investigator (Shoko Sasayama) will have access to your responses. Once the study has terminated, all information collected from you will be discarded. Your name or other identifiable information will not be included in the publications that result from the research project. Please note, however, that even if your name is not used in publication, the researcher will still be able to connect you to the information gathered about you in this study. The Georgetown University IRB is allowed to access your study records if there is any need to review the data for any reason.

**EXTRA CREDIT/PAYMENT**
You will receive extra credits for Ling 001: Introduction to Language.

**YOUR RIGHTS AS A RESEARCH PARTICIPANT**
Participation in this study is entirely voluntary at all times. You can choose not to participate at all or to leave the survey at any point. If you decide not to participate or to leave the survey, there will be no effect on your relationship with the researcher or any other negative consequences. If you decide that you no longer want to take part in the study, the information already obtained through your participation will be included in the data analysis and final report for this study.

**QUESTIONS OR CONCERNS?**
If you have questions about the study, you may contact Shoko Sasayama at 202-687-6213 or ss2228@georgetown.edu. You may also contact the researcher’s faculty advisor, Dr. John M Norris at norrisj@georgetown.edu. Please call the Georgetown University IRB Office at 202-687-1506 (8:30am to 5:00pm, Monday to Friday) if you have any questions about your rights as a research participant.

379
Title: Effects of cognitive task complexity on second language learners' task performance

CONSENT OF PARTICIPANT

I understand all of the information in this Informed Consent Form.
I have gotten complete answers for all of my questions.
I freely and voluntarily agree to participate in this study.
I understand that I will be audio recorded as a part of this study.

Please indicate whether you agree to be audio recorded as a part of this study.

☐ YES  (If you change your mind about this at any point, please let the researcher know)
☐ NO

________________________________________  ____________________________
Signature of Subject                      Date

________________________________________
Printed Name of Participant

Once you sign this form, you will receive a copy of it to keep, and the researcher will keep another copy in your research record.
Appendix J
Consent Form for Study 1 (English-L2 Speakers)

実験参加同意書

この実験は課題の難易度とパフォーマンスとの関係を調べることを目的としています。

この実験では、英語でイラストを説明する課題、色の変化を見分ける課題、アンケート、そして簡単な英語のテストを受けていただきます。実験は約1時間かかります。一つ目の課題に対する解答は録音されますが、その録音を研究者以外の者が聞くことはなく、その録音はこのプロジェクトが終わり次第破棄されます。また、個人が特定されるような情報が実験結果に反映されることはありません。

この実験への参加は任意です。実験中いつでも参加を取りやめることができます。実験を終える絶えないにかかわらず、謝礼として千円をお支払いいたします。

研究結果は学会での発表、また論文として公表される予定です。その場合、参加者の皆様の氏名などの個人情報が明らかにされる事はありません。

実験内容に関してご不明な点がありましたら、研究者：笈山尚子 (shokosasayama@gmail.com) までご連絡ください。実験に際して、大変お手数をおかけしますが、ご協力よろしくお願いいたします。

ジョージタウン大学 言語学科
博士課程 応用言語学専攻

笈山尚子
上記同意書を読み、理解しました。また、実験に参加することを快諾します。

お名前

日付

上記同意書のコピーをご希望の方は、以下にメールアドレスをご記入下さい。
Appendix K

Consent Form for Study 2 (Native Speakers of English)

Title: Effects of cognitive task complexity on second language learners’ task performance

Georgetown University
Consent to Participate in Research Study

SURVEY

STUDY TITLE: Effects of cognitive task complexity on second language learners’ task performance

PRINCIPAL INVESTIGATOR: Shoko Sasayama
TELEPHONE: 202-687-6213

ADVISOR: Dr. John M Norris

INTRODUCTION
You are invited to consider participating in this research study. Please take as much time as you need to make your decision. Feel free to discuss your decision with whomever you want, but remember that the decision to participate, or not to participate, is yours. If you decide that you want to participate, please sign and date where indicated at the end of this form. If you have any questions, please ask the principal investigator (Shoko Sasayama).

PURPOSE
The purpose of this study is to investigate the effects of task complexity on second language learners’ task performance.

STUDY PLAN
You are being asked to take part in this study because your first language is English. You will be asked to tell a story in English based on five sets of pictures (your story will be audio-recorded), fill out a questionnaire on your perceptions of each task, and complete a background questionnaire. The session should take you approximately 30 minutes to complete.

RISKS
There are no known risks associated with participating in this study.

BENEFITS
There may be no direct benefit to you in participating in this study. However, your participation will contribute to advancement of the field of Second Language Acquisition by allowing us to investigate the effects of task design on second language learners’ task performance.

CONFIDENTIALITY
Every effort will be made to keep any information collected from you confidential. However, it is impossible to guarantee absolute confidentiality. In order to keep information about you safe, the digital data will be kept password-locked in a password-protected personal computer and the paper-based data will be kept in a locked box at the researcher’s office at all times. Only the principal investigator (Shoko Sasayama) will have access to your responses. Once the study has terminated, all information collected from you will be discarded. Your name or other identifiable information will not be included in the publications that result from the research project. Please note, however, that even if your name is not used in publication, the researcher will still be able to connect you to the information gathered about you in this study. The Georgetown University IRB is allowed to access your study records if there is any need to review the data for any reason.

EXTRA CREDIT
You will receive an extra 5% homework extra credit for Ling 001: Introduction to Language.

YOUR RIGHTS AS A RESEARCH PARTICIPANT
Participation in this study is entirely voluntary at all times. You can choose not to participate at all or to leave the survey at any point. If you decide not to participate or to leave the survey, there will be no effect on your relationship with the researcher or any other negative consequences. If you decide that you no longer want to take part in the study, the information already obtained through your participation will be included in the data analysis and final report for this study.

QUESTIONS OR CONCERNS?
If you have questions about the study, you may contact Shoko Sasayama at 202-687-6213 or ss2228@georgetown.edu. You may also contact the researcher’s faculty advisor, Dr. John M Norris at norrisj@georgetown.edu. Please call the Georgetown University IRB Office at 202-687-1506 (8:30am to 5:00pm, Monday to Friday) if you have any questions about your rights as a research participant.
CONSENT OF PARTICIPANT
I understand all of the information in this Informed Consent Form.
I have gotten complete answers for all of my questions.
I freely and voluntarily agree to participate in this study.
I understand that I will be audio recorded as a part of this study.

Please indicate whether you agree to be audio recorded as a part of this study.
☐ YES (If you change your mind about this at any point, please let the researcher know)
☐ NO

_________________________________________ __________________________
Signature of Subject Date

_________________________________________
Printed Name of Participant

Once you sign this form, you will receive a copy of it to keep, and the researcher will keep another copy in your research record.
Appendix L
Consent Form for Study 2 (English-L2 Speakers)

IRB #2014-0197

脅迫名: Effects of cognitive task complexity on second language learners’ task performance

ジョージタウン大学
実験参加同意書

研究課題名: Effects of cognitive task complexity on second language learners’ task performance

研究責任者: 筆山 尚子
指導教授: DR. JOHN NORRIS

はじめに
本同意書をよく読み、実験へ参加いただけるかどうかを熟慮してください。その際、誰かに相談していただ
いても構いませんが、参加するか否かの最終決定はご自身でしていただくよう、お願いいたします。参加
に同意していただきましたら、本同意書の最後にご署名ください。質問等ございましたら、研究責任者
（筆山尚子）までお尋ねください。

目的
本研究は、英語課題のデザインと第二言語学習者のパフォーマンスとの関係について調べることを目的と
しています。

研究計画
英語を現在学んでいる、または英語の学習経験がある方に実験への参加をお願いしています。この実験で
は、英語でイラストを説明する課題、アンケート、そして簡単な英語のテストをしていただきます。課題
の中には、回答が録音されるものがあります。実験は全部で1時間15分程度かかる予定です。

リスク
この実験に参加するにあたって、リスクはありません。

利益
この実験に参加することで、英語を話す機会が得られます。また、本実験への参加は、第二言語習得の分
野への貢献につながります。

守秘義務
実験中に集めた情報はすべて厳守しないよう、十分に配慮いたします。そのため、デジタルデータにはロ
ックをかけ、パスワードで保護されたコンピュータに保管します。また、紙上のデータは、研究者のオ
フィス内の鍵をかけたロッカーに保管されます。同データを見ることができるのは、研究責任者（筆山尚
子）のみです。研究が終了次第、全ての情報は破棄されます。また、氏名や個人が特定できるような情報
が公開されることはありません。なお、もし何らかの理由で、データを審査する必要がある場合、ジョー
ジタウン大学倫理委員会は参加者の皆様の情報をお知らせします。

謝礼
実験参加に対する謝礼として、実験終了後に千五百円お支払いいたします。

実験参加者としての権利
この研究に参加されるかどうかは、皆様の自由です。実験に参加しないという選択をすることもできます
し、実験中において参加を取りやめることができます。実験に参加しないという選択をした場合でも、
何ら不利益を受けることはありません。ただし、その時点までにすでに集められたデータは、研究データ
の一部として扱われます。

質問等
この研究について何か質問がありましたら、筆山尚子 [ss2228@georgetown.edu] までご連絡ください。ま
た、研究責任者の指導教授である DR. JOHN NORRIS (norrisj@georgetown.edu) に連絡していただいても構いま
せん。実験参加者としての権利について質問がある場合には、ジョージタウン大学倫理委員会にお電話く
ださい。

ジョージタウン大学倫理委員会
電話番号: +1-202-687-1506
受付時間: アメリカ東部標準時 月曜日から金曜日、午前 8:30 から午後 5 時まで
ジョージタウン大学
実験参加同意書

ジョージタウン大学 竜山 尚子 殿
研究課題名「課題の難易度が第二言語学習者のパフォーマンスに与える影響についての研究」

上記研究課題の内容について、以下の四点を確認いたします。

本実験参加同意書にある全ての項目を理解いたしました。
本研究について、不明な点はありません。
本研究に自分の意思で参加いたします。
実験の中で、自分の回答が録音されるとの告知を受けました。
実験の中で、回答を録音することに同意していただける方は、以下の「はい」をチェックしてください。（同意を撤回したい場合には、いつでも研究者にお知らせください。）
□ はい
□ いいえ

本研究に参加することに同意し、以下に署名します。

______________________________
署名（サイン）

______________________________
氏名

この同意書に署名いただけましたら、同意書の写しをお渡しいたします。こちらでも研究記録として同じ写しを保管いたします。

__________ 年 __ 月 __ 日
同意年月日
Appendix M
Task Instructions Sheet for the Dual-Task Condition (English version)

Task Instructions (A)

In this experiment, you will do (a) five story-telling tasks and (b) a color detection task.

(a) Story-telling tasks
For each story-telling task, you will see a 6-framed cartoon strip on the computer screen. Based on the pictures, tell a story in English. You have 5 minutes for each story.

• First, you have 30 seconds to prepare your story. (Note-taking not allowed)
• After 30 seconds of preparation, the pictures will disappear.
• Start telling the story when you see the pictures again.
• As soon as you finish the story, hit the “Q” key.
• Then, you will see “Answer the questionnaire. The task number is XXX,” so please answer the questionnaire.
• When you finish the questionnaire, hit “return” to move to the next story-telling task.
• Your story will be assessed based on its completeness (i.e., whether all six pictures are included in the story), effectiveness (i.e., whether the order of the pictures is clear to the listener), and creativity; so, please do your best in telling the stories.
• There will be a practice session (with the questionnaire) at the beginning.

(b) Color-detection task
At the same time, you will be asked to do a color detection task. The task is to hit the “space” key when you detect a color change in the letter “A” presented on the computer screen. Hit “space” ONCE, only when the color changes from BLACK to RED (but not from red to black).

There will be a color detection task without a story-telling task after the practice session.

Practice session Color detection task 30 sec. preparation time Story telling in English Hit “Q” Answer questionnaire Hit “return” to start the next task

At the same time, Hit “space” on black ➔ red

Repeat 4 times

Any questions?
I'd appreciate it if you could keep the content of this experiment confidential.
Thank you for your participation!
Appendix N
Task Instructions Sheet for the Dual-Task Condition (Japanese version)

課題の説明 (A)

英語でイラストを説明する課題と色の変化を見分ける課題が5問（練習課題1問と本課題4問）あります。

英語でイラストを説明する課題

コンピューターのスクリーンに6枚のイラストが表示されますので、そのイラストをもとにストーリーを考え、英語で説明してください。それぞれの課題につき、5分の時間があります。

- まず、30秒間でストーリーを考えてください。（ノートはとれない）
- 30秒間の準備のあと、イラストが一気に消えます。
- 再びイラストが現れたたら、準備したストーリーを話し始めしてください。
- ストーリーを話し終えたら、できるだけ早く“Q”のキーを押してください。
- 次に、「アンケートに答えしてください。課題番号は○○番です。」という画面が現れるので、アンケートに答えてください。
- アンケートに答え終わったら、“return”キーを押して、次の課題に進んでください。
- あなたのストーリーは、6枚すべてのイラストがストーリーに含まれているか、6枚のイラストの順序が聞き手に明確に分かるかどうか、創造性があるか、について評価されますので、ストーリーを話のにベストを尽くしてください。

- 実験のはじめに、練習課題（アンケートも含む）があります。

色の変化を見分ける課題

同時に、色の変化を見分ける課題もしていただきます。コンピューターのスクリーンに表示された“A”の文字の色が黒から赤に変わるたびに、“space”キーを押してください。（赤から黒に戻る時には押す必要はありません）

練習課題の直後に、色の変化を見分けるだけの課題があります。

実験に参加していただき、ありがとうございます。
質問があれば、お知らせください。
また、この実験の内容は他の方には伏せていただけますよう、よろしくお願い致します。
Appendix O
Task Instructions Sheet for the Single Task Condition (English version)

Task Instructions

In this experiment, you will engage in five story-telling tasks.

For each story-telling task, you will see a 6-framed cartoon strip on the computer screen. Based on the pictures, **tell a story in English**. You have 5 minutes for each story.

- First, you will have **30 seconds to prepare** your story. (Note-taking not allowed)
- After 30 seconds of preparation, the pictures will disappear.
- Start telling the story **when you see the pictures again**.
- **As soon as you finish the story**, hit the “Q” key.
- Then, you will see “Answer the questionnaire. The task number is XXX,” so please answer the questionnaire.
- When you finish the questionnaire, hit “return” to move to the next story-telling task.
- **Your story will be assessed based on its completeness (i.e., whether all six pictures are included in the story), effectiveness (i.e., whether the order of the pictures is clear to the listener), and creativity:** so, **please do your best in telling the stories.**
- There will be a **practice** session (with the **questionnaire**) at the **beginning**.

Any questions?

*I'd appreciate it if you could keep the content of this experiment confidential.*

*Thank you for your participation!*

389
Appendix P
Task Instructions Sheet for the Single Task Condition (Japanese version)

課題の説明

英語でイラストを説明する課題が5問（練習課題1問と本課題4問）あります。

コンピューターのスクリーンに6枚のイラストが表示されますので、そのイラストをもとにストーリーを考え、英語で説明してください。それぞれの課題につき、5分の時間があります。

・まず、30秒間でストーリーを考えてください。（ノートはとれません）
・30秒間の準備のあと、イラストが一度消えます。
・再びイラストが現れたら、準備したストーリーを話し始めてください。
・ストーリーを話し終えたら、できるだけ早く“Q”のキーを押してください。
・次に、「アンケートに答えてください。課題番号は〇〇番です。」という画面が現れるので、アンケートに答えてください。
・アンケートに答え終わったら、“return”キーを押して、次の課題に進んでください。
・あなたのストーリーは、6枚すべてのイラストがストーリーに含まれているか、6枚のイラストの順序が聞き手に明確に分かるかどうか、創造性があるか、について評価されますので、ストーリーを話すのにベストを尽くしてください。
・実験のはじめに、練習課題（アンケートも有）があります。

実験に参加していただき、ありがとうございます。
質問があれば、お知らせください。
また、この実験の内容は他の方には伏せていただきますよう、よろしくお願いいたします。
Appendix Q
Open-Ended Responses Coding Protocol

Open-Ended Responses Coding Protocol

Instructions
You will be asked to categorize 12 participants’ comments about four tasks. In this experiment, participants engaged in four picture-based narrative tasks. They were then asked to rate the degree of task difficulty and mental effort for each task, using 9-point Likert scales (see Appendix A for the task difficulty/mental effort questionnaire). In this questionnaire, they were also asked to explain their ratings. It is these explanatory comments that you are asked to categorize.

Please read the description of categories below to familiarize yourself with the coding categories. Then, open the excel file entitled “EnglishL2 open-ended responses analysis coder 2” and choose the “responses” tab. There you will find 12 participants’ explanatory comments for their ratings of task difficulty and mental effort for the four tasks. The column “Task 1 TD” contains each learner’s comments for Task 1 in response to their task difficulty ratings. Similarly, “Task 1 ME” contains learners’ comments for Task 1 in response to their mental effort ratings. I think it’s easier to code all task difficulty comments for Task 1, before moving on to Task 2 TD, and so on. (So, code the comment in the Task 1 TD column provided by Participant #2, then code the comment in the same column provided by Participant #7…etc.) In coding Task 1 TD comments, open the sheet entitled “Task 1 TD” in the same excel sheet and copy/past each comment to the appropriate category. I have included an example in the top row. As you categorize each comment, indicate (a) factors that made the task easy or less effort consuming in blue and (b) factors that made the task hard or more effort consuming in red. If it is hard to judge solely from the comments themselves, please refer to their task difficult/mental effort ratings (provided in the sheet entitled “Ratings” in the same excel document) and judge based on them. If that’s still not possible, please leave a comment by inserting “New Comment.” For any other ambiguous cases too, please feel free to leave comments (as I have done in the example).

Coding categories
There are four over-arching categories: (A) input, (B) code complexity, (C) performance factors, and (D) other.

A. Input is concerned with the task prompt (i.e., picture sets), and it is further divided into (i) storyline/picture quality, (ii) simple input making task difficult, (iii) number of characters, (iv) logic, (v) familiarity/relatability, and (vi) background. For the actual task prompts, see Appendix B.

B. Code complexity is about linguistic aspects of task performance.

C. Performance factors are further divided into two sub-categories: (i) task-instruction-related factors and (ii) learner-generated factors.

Below, I will explain each category more in depth with some examples.

A. Input
i. Storyline/picture quality: Comments about (a) clarity/complexity of storyline and (b) clarity of input (pictures) (Note. It is not necessary to distinguish between these two sub-categories.)

Example comments
• 絵がシンプルで説明しやすかった [The pictures were simple, and it was easy to describe them.] (Task 1)
• 絵が細かくて、コマ目で男の人が何を釣り上げているのか判別できなくて焦った [The picture was so fine that I couldn’t figure out what the man was catching in the third frame and I got panicked.] (Task 2)
• 4枚目と5枚目のイラストを違うように説明できなかった [I could not describe the fourth and fifth frames differently.] (Task 3)
ii. **Simple input making task difficult**: A special case of the first category (i.e., storyline/picture quality). Comments where learners argue that a simple input made the task difficult to deal with are categorized into this group.

*Example comments*
- As the pictures were simple and plain, I felt that I had to create a story on my own.] (Task 1)
- There were a lot of frames that didn’t entail much action, like the frame where a man is simply sleeping; so, it was difficult to understand the storyline right away.] (Task 1)

iii. **Number of elements**: Comments about the number of elements (e.g., characters, events) involves in the story, and consequences for having multiple characters.

*Example comments*
- There are a lot of characters, and it is difficult to distinguish among them. (Task 4)
- If we assume that the boy and the girl are the mother’s children, I wondered how I should address them. (Task 3)

iv. **Logic**: Comments about knowing or not knowing why things happen the way they do in the picture strips.

*Example comments*
- I couldn’t figure out why the man looked outside. (Task 1)
- I didn’t know where the bowling ball came from and why, so I didn’t know how to describe the pictures. (Task 3)

v. **Familiarity/relatability**: Whether the storyline depicted by each task is familiar or relatable to them. This category does not include word familiarity. Note that a comment that specifically mentions code-related issues (even in relation with familiarity), for example a comment like “日常生活が描かれているイラストだったので、普段からよく使う英単語で説明できたから [The illustration depicted a daily life, so I was able to explain with words that I normally use],” should be categorized as CODE (see below), not familiarity/relatability.

*Example comments*
- With regards to the size of the fish, I didn’t have enough knowledge to judge whether they were big or small, so it was hard to create a story. (Task 2)

vi. **Background**: Whether the background changes throughout the picture strips.

*Example comments*
- The story took place in the same scene. (Task 1)

B. **Code complexity**
Comments about linguistic aspects (e.g., general proficiency, grammar, lexis, complexity, accuracy, fluency). Any comments that explicitly address issues with language (code) should be included here.

*Example comments*
- I cannot speak English. (Task 3)
- Because I tried to paraphrase the words that I didn’t know] (Task 2)

C. **Performance factors**

i. **Task-instruction-related factors**: Comments that are related to (a) creativity (創造性があるか), (b) completeness (6枚全てのイラストがお話に含まれているか), (c) effectiveness (絵の順序が聞き手に分
Prior to engaging in tasks, participants were asked to pay particular attention to these three aspects when telling a story, in a task-instruction sheet (see Appendix C). *(Note. It is not necessary to distinguish between these two sub-categories.)*

**Example comments**

- 創造性を考えるのが難しい [I was difficult to make my story creative.] (Task 1)
- イラストの順序を的確に説明しようとしたから [I tried my best to make the order of the pictures clear.] (Task 3)

ii. **Learner-generated factors:** Comments that reflect (a) participants’ strategies to deal with high cognitive activity and (b) their attempts to perform well.

**Example comments**

- 分からない単語は考えても出てくるので、半分あきらめが入っている [I’m half giving up because I cannot think of words that I don’t know.] (Task 1)
- うまく表現しようと思って、余計に負担がかかるた [I gave myself extra burden for trying to describe well.] (Task 2)
- 一つ前の課題では、時制を現在形にしてとても混乱したので過去形で統一しようとしたのが良かった [I was really confused when I was telling the previous story because I decided to use the present tense, so I decided to stick to the past tense this time, and this strategy worked well.] (Task 3)

D. **Other**

Any comments other than described above.

I did my best to make these four categories as distinct as possible, but they are inevitably related and overlapping. To facilitate the coding process, below I will provide some additional guidelines as to how to deal with comments that touch upon multiple categories.

i. If a comment includes elements of both (a) number of elements and (b) code complexity, code it as number of elements. For example, a comment like “二人の男を話す中で、どのように言い分ける（区別）のが難しかった [In talking about two men, it was difficult to distinguish between them.]” (Task 2)” should be coded as number of elements (input). Although the consequence lies in high code complexity, the source of difficulty stems from the fact that there were two similar men in the input.

ii. Otherwise, comments about linguistics aspects should be categorized as ‘code complexity.’

iii. If a comment includes elements of both (a) input and (b) performance factors, code it as input. For example, if you see a comment like “4枚目と5枚目のイラストを違うように説明できなかった [I could not describe the fourth and fifth frames differently.]” (Task 3),” code it as input, although this difficulty also stems from the fact that they were told to include all six pictures in their story in the task instructions.

iv. If a comment includes elements of both (a) code complexity and (b) performance factors, categorize it as ‘code complexity.’ A comment like, “なるべく分かりやすく相手に伝えたいと思い、イディオム等を必死に考えたから” (Task 3)” should be categorized as code complexity, despite the likely possibility that the participant was conscious of the listener because of the kind of task instructions they received (i.e., Your story will be assessed based on its effectiveness – whether the order of the pictures is clear to the listener).

v. In other words, prioritize INPUT, then CODE COMPLEXITY, followed by PERFORMANCE FACTORS (INPUT > CODE > PERFORMANCE FACTORS).

Additionally:

vi. A comment concerning a relationship among picture frames should be coded as ‘input.’ For example, “2と3と4のイラストのストーリーをつなげるのに少し頭を使いました [I used my brain to some extent to connect the second, third, and fourth picture frames]” (Task 4)” should be coded as input.

vii. In coding comments for Task 3 (the baby task), a comment about the ambiguity revolving around a bowling ball that was replaced with a baby should be coded as input, not code complexity. *(e.g., “赤ちゃんの代わりに置いた人形をどう表現すべきか分からなかった。[I didn’t know how to describe the doll put in place of the baby.]”)*
Appendix A
Task Difficulty and Mental Effort Questionnaire

実験用ID番号（）
課題番号（）

<table>
<thead>
<tr>
<th>1. 6枚のイラストを見てストーリーを考え、それを元で解明するのは簡単でしたか？</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>
</tr>
</thead>
<tbody>
<tr>
<td>2. それはなぜですか？なるべく詳しく書いてください。</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. あなたが書いたストーリーの長さはどれくらいだったと思いますか？</td>
<td>約（）分（）秒</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. この課題をこなすのに、どのくらい頑を見せましたか？</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>5. それはなぜですか？なるべく詳しく書いてください。</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix B
Task Prompts

Task 1: Alarm Clock Task (Adapted From Hill, 1960, pp. 48-49)

Task 2: Going Fishing Task (Adapted From Hill, 1960, pp. 10-11)

(A picture prompt was inserted here in the original version; however, it could not be reprinted due to its copyright.)
Task 3: Baby Task (Adapted From Hill, 1960, pp. 60-61)

(A picture prompt was inserted here in the original version; however, it could not be reprinted due to its copyright.)

Task 4: Dog Task (Adapted From Elder & Iwashita, 2005, p. 238)

(A picture prompt was inserted here in the original version; however, it could not be reprinted due to its copyright.)
Appendix C
Task Instructions

課題の説明

英語でイラストを説明する課題が5問（練習課題1問と本課題4問）あります。

コンピューターのスクリーンに6枚のイラストが表示されますので、そのイラストをもとにストーリーを考え、英語で説明してください。それぞれの課題につき、5分の時間があります。

- まず、30秒間でストーリーを考えてください。（ノートはとられません）
- 30秒間の準備のあと、イラストが一度消えます。
- 再びイラストが現れたら、準備したストーリーを話し始めてください。
- ストーリーを話し終了したら、できるだけ早く“Q”のキーを押してください。
- 次に、「アンケートに答えてください。課題番号は〇〇番です。」という画面が現れるので、アンケートに答えてください。
- アンケートに答え終わったら、“return”キーを押して、次の課題に進んでください。
- あなたのストーリーは、6枚すべてのイラストがストーリーに含まれているか、6枚のイラストの順序が間違っているか、創造性があるか、について評価されますので、ストーリーを話すのにベストを尽くしてください。
- 実験のはじめに、練習課題（アンケートも含）があります。

実験に参加していただき、ありがとうございます。
質問があれば、お知らせください。
また、この実験の内容は他の方には知らせていませんよう、よろしくお願いいたします。

7
Appendix R
Second Coder Training Manual: Complexity and Accuracy Measures
Second Coder Training Manual (English)

An arrow (→) indicates an action that coder takes

Please code the following units of analysis in order

Notes: In the examples below, disfluency markers (i.e., reformulations, repetitions, false starts, and replacement) are bracketed; words in non-target language (i.e., in Japanese) are italicized; inaudible words are marked as XXX.

(1) **T-unit:** A main clause plus all subordinate clauses and non-clausal structures attached to or embedded in it.

→ **Indicate a T-unit by hitting the return key twice.**

*Example*

**Segment of text before T-unit parsing**
[and and a boy] and [the boy *etto* the boy] the boy [put] put a ball [on the] in the bed
[when the mother] when the mother woke up [she found she found that she found
that *etto* there is] she found the baby [was not] was not in the bed and instead of the baby [she she found a ball] she found the ball in the basket

**Segment of text after T-unit parsing**
[and and a boy] and [the boy *etto* the boy] the boy [put] put a ball [on the] in the bed
[when the mother] when the mother woke up [she found she found that she found
that *etto* there is] she found the baby [was not] was not in the bed and instead of the baby [she she found a ball] she found the ball in the basket

(2) **Clause:** There are three types of dependent clauses: ‘**coordination**’ (e.g., I went to a store
and bought bread), ‘**subordination**’ (e.g., The dog barked when the criminal broke into the
house, He said that he would be late), and ‘**embedding**’ (e.g., Jane likes the restaurant where
she had her birthday party last year) (Dependent clauses underlined) There is no need to identify which type of clause it is.

→ **Indicate a clause with a double colon (::) at each boundary. Make sure to insert spaces around it.**

*Examples*

[the] then a man [using] using steps [fa] fell into the ground :: and [get] got injured (1 T-unit, 2 clauses; coordination)

while playing with baby [they they thought they thought a] they thought one idea :: [that]
that [they they mm replace] they replaced the children for a ball (1 T-unit, 2 clauses; subordination)
[then] then [her ch] her children :: [who who are] who are the baby’s brother and sister ::
[came into] came into the room :: [and] and [se] said to the mother :: [let’s] let’s play with
us (1 T-unit, 4 clauses; embedding, coordination, subordination)

A clause must include a **finite verb**: hence, the second example of a T-unit above is considered
to include two clauses instead of three. If it were “while they were playing with baby they thought
one idea that they replaced the children for a ball” it would be considered to include three clauses.

Similarly, the following T-unit is considered to contain only one clause, as there is no finite verb
between ‘person’ and ‘standing’:

and [the gir the per] a person standing XXX the field [can] can get fish again and again (1
T-unit, 1 clause)

On the other hand, the following T-unit is considered to contain two clauses:

but the man :: who’s standing in the river :: actually [it’s] cannot catch the water (1 T-unit, 2 clauses)

Note that this complexity analysis does not take into account accuracy of language use.
Therefore, even if a finite verb is missing, if you judge it to be due to a lack of proficiency rather
than learners’ intention to use a non-finite verb, count the element as a clause. For example, I
would think that the following T-unit includes two clauses, not one, although the finite verb ‘is’
is missing:

but [person] a person :: who [stand] standing a field :: actually [can] can get a fish (1 T-unit, 2 clauses)

(3) Complex nominal: Complex nominals include: (a) nouns plus adjective, possessive (e.g., his,
someone’s), prepositional phrase, relative clause, participle, or appositive (e.g., An old
man and his dog, Spot, were walking down the street, I like the idea that we go on a vacation) as
well as compound nouns (e.g., car driver), (b) nominal clauses (e.g., I wonder how
difficult the exam will be, she said that it will rain tomorrow, she said to her friend “let’s go
out for dinner tonight” – reported speech), and (c) gerunds and infinitives in subject
position. Even if a noun is modified by multiple elements (e.g., I found Tim’s tall tree that I
had been looking for.), count as **one** complex noun.

→ **Select the core noun for each complex nominal and put a dollar sign ($) in front of it. Do not insert a space between the dollar sign and the noun. If there is no core noun (as in the third example below), put a dollar sign at the beginning of a complex nominal.**

*Example*

[then] then [her ch] her $children :: [who who are] who are the baby’s $brother and
sister :: [came into] came into the room :: [and] and [se] said to the mother :: [let’s] $let’s play with us
(4) **Accuracy**: Errors are defined as non-target-like use of the language. Errors are defined at the level of T-unit/clause, rather than a discourse or pragmatic level. Hence, it would not be counted as an error when different verb tenses were used within a single task performance or when common nouns were substituted for pronouns, as long as the language use is deemed accurate within a T-unit. Words in brackets will be excluded from the analysis of accuracy. Words in non-target language (i.e., in Japanese; indicated in *italics*) should be considered incorrect.

→ **Identify a clause containing error(s) and:**

(a) **Put a percentage sign and a G (i.e., %G) at the beginning of the clause if it contains grammar errors;**

*Example*

%G while playing with baby [they they thought they thought a] they thought one idea :: %G [that] that [they they mm replace] they replaced the children for a ball

(b) **Put a percentage sign and an L (i.e., %L) at the beginning of the clause if it contains lexical errors;**

*Example*

%L and [the] he tried to stop the timer (< alarm clock)
%L he throw his ee cushion (< pillow)

**If a clause contains errors of multiple categories, include a code for each one.**

Please do insert a space between the sign and the first word in the clause.

As grammatical errors are defined at the level of T-unit/clause, **an incorrect use of connectors at the beginning of a T-unit (e.g., and, but) should not be marked as incorrect.**

Some notes about pauses:
1.  *(long pause)* is inserted when there is an extremely long pause.
2.  Where a sentence boundary could be ambiguous, *(pause)* is inserted to mark it more clearly. When in doubts, however, please listen to the audio recordings.
Tasks

Task 1: Alarm Clock Task (Adapted From Hill, 1960, pp. 48-49)

Example task performance by a native speaker

Uh so there’s a man sitting on his bedroom, setting his alarm clock for the next day. Then he goes over to the window. He looks outside at the moon and the stars and closes the curtains but not quite all the way. Then he gets in the bed and turned off the light, and he goes to sleep for the night. Then he wakes up to his alarm clock, and it’s really loud, and there’s lots of lights shining through the windows because the curtains are slightly opened. But he doesn’t want to get out of the bed quite yet. He’s still very tired. So he throws his pillow at the alarm clock, hoping that that’ll shut it off, and it does. And so uh his nightstand and his alarm clock and the pillow they are all over the floor, but the alarm clock is not ringing anymore, and the man gets to go back to sleep. (NS12)
Task 2: Going Fishing Task (Adapted From Hill, 1960, pp. 10-11)

Example task performance by a native speaker
One day a father and his daughter a father took his daughter down to the pond to fish. The dad, Ralph, considered himself to be an expert fisher and had many techniques to teach his daughter, Lucy. He hopped in the pond and began fishing in the pond as opposed to outside of the pond where Lucy had stayed. He was searching for a fish when unexpectedly he saw Lucy grabbed the first fish and he said who said “Look pa! I got a fish before you!” as he raised up a boot he had caught within the pond. In this way, Ralph was determined to catch a much larger fish than his daughter in order to prove his expertise. Uh Lucy had then captured another fish but wasn’t able to um bring it onto shore. Umm Ralph who was still looking for a fish ended up catching his boot this time and tripped himself into the water where Lucy stayed on the shore laughing at him um and remarked “yeah pa, ha ha ha you’re quite the expert!” (NS6)
Task 3: Baby Task (Adapted From Hill, 1960, pp. 60-61)

Example task performance by a native speaker
One day a mother was watching her baby by the fireplace. She was reading quietly while her baby sat on her feet in a little crib. As she was reading, she began to be tired and fell asleep. Then two little kids came up and thought they would play an evil trick on her. They took the baby out of the crib while she was fast asleep and replaced it with a ball looking like the baby’s head, so only the head was sticking out from the blanket. When the mother woke up finally, she realized the baby was missing and that the ball in the basket was not the real baby. The kids were at this time standing right behind her chair, laughing hilariously, thinking about the cruel old trick that they played on this poor mother. (NS23)
Task 4: Dog Task (Adapted From Elder & Iwashita, 2005, p. 238)

Example task performance by a native speaker
One day John was walking down the street with his dog, Fido. Fido got very excited when he saw something across the street, and so all of a sudden, he jumped up, pulled away from John, and ran across the street, barking excitedly. Unfortunately, this caused the car coming down the street to suddenly have to swerve to the right to escape running into Fido. This caused the car to knock someone knock a painter off a ladder and to make two people who were having lunch on the sidewalk frantically have to run away from the oncoming car. All of this meant that a couple of people had to go to the hospital. So when the police came, they wanted to arrest whoever had caused the accident. However, they accidentally arrested the driver who was completely innocent, but luckily enough he was able to convince them that the cause of the problem had been John and Fido, who were trying to walk away without getting into trouble. So in the end, John and Fido ended up in jail because they had caused this terrible accident. (NS10)
### Appendix S
Elicited Imitation Test Scoring Rubric

**SCORING GUIDELINES FOR ELICITED IMITATION TASK**

**L2 ENGLISH**

12/17/98

Noriko Iwashita & Lourdes Ortega


**REVISED by Shoko Sasayama on June 8, 2015**

<table>
<thead>
<tr>
<th>SCORE 0</th>
<th><strong>Criteria</strong></th>
<th><strong>Examples</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nothing (Silence)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Garbled (unintelligible, usually transcribed as XXX)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Minimal repetition, then item abandoned:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Only 1 word repeated</td>
<td>- The-street in... in... street... hmm (16/#2)</td>
</tr>
<tr>
<td></td>
<td>- Only 1 content word plus function word(s)</td>
<td>- ... on the table (2/EL10)</td>
</tr>
<tr>
<td></td>
<td>- Only 1 content word plus function word(s) plus extraneous words that weren't in the original stimulus</td>
<td>- I wish... comfta-portable (19/#1)</td>
</tr>
<tr>
<td></td>
<td>- Only function word(s) repeated</td>
<td>- I watch a movie (9/#22)</td>
</tr>
<tr>
<td></td>
<td>- However, if the item is short and contains only 2-3 idea units, do not apply the above rule. For these short items, as long as 2 or more words (except for articles) are repeated, score 1.</td>
<td>- You don’t... don’t you? (14/#1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- I have to get (1/ EL6) (score 1, not zero)</td>
</tr>
<tr>
<td></td>
<td><strong>NOTE:</strong> with only, just, yet (meaningful adverbs), score 1</td>
<td>- He just finished (15/#23) (Closed word + Adv + lexical word) (score 1)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SCORE 1</th>
<th><strong>Criteria</strong></th>
<th><strong>Examples</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>When only about half of idea units are represented in the string but a lot of important information in the original stimulus is left out</td>
<td>- Cross the cross--cross the street ahead and. (16/#4)</td>
</tr>
<tr>
<td></td>
<td>- Don’t have nap (7/#1)</td>
<td>- I have to hair-hairc (1/#24)</td>
</tr>
<tr>
<td></td>
<td>- The last year (20/#4)</td>
<td>- Would you... the book on the table (26/#7)</td>
</tr>
<tr>
<td></td>
<td>- I have a long piece [peace?] of a nap (&lt;a long, peaceful nap) (7/#4)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- The food was supposed to have ve- good food (&lt;is supposed; meaning changed to past) (12/#4)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SCORE 2</th>
<th><strong>Criteria</strong></th>
<th><strong>Examples</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>When content of string preserves at least more than half of the idea units in the original stimulus; string is meaningful, and the meaning is close or related to original, but it departs from it in some slight changes in content, which makes content inexact, incomplete, or ambiguous</td>
<td>- The gooda friend take care o- chi- children (left out that it was the neighbor’s children, and that they were three) (21/#1)</td>
</tr>
<tr>
<td></td>
<td>- After dinner I have a long piece [peace?] of a nap (&lt;a long, peaceful nap) (7/#4)</td>
<td>- She just finished painting the seaside her apartment (&lt;inside of) (15/#4)</td>
</tr>
<tr>
<td></td>
<td>- The restaurant was supposed to have ve- good food (&lt;is supposed; meaning changed to past) (12/#4)</td>
<td></td>
</tr>
</tbody>
</table>
- I want to big house which... in which... animal can live (left out ‘nice’ ‘my’ and made animal into singular) (13/#4)
- Would you hand me... the books which are on the table (<book; meaning changed to plural) (26/#4)
- It is possible to day tomorrow (from pronunciation problem, it is ambiguous whether ‘rain’ has been understood, but it is possible) (8/#1)

**SCORE 3**

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Original, complete meaning is preserved as in the stimulus. Strings which are quite ungrammatical can get a 3 score, as long as exact meaning is preserved. Some synonymous substitutions are acceptable.</td>
<td>- It is possible that it will be rain tomorrow (8/ EL1)</td>
</tr>
<tr>
<td>• Examples of acceptable substitutions (SCORE 3): hand/give/pass are acceptable synonyms for item 26. Substitutions of and &amp; but are acceptable. A lot of = many, etc.</td>
<td>- Would you pass me the book on the table (26/#21)(Score 3)</td>
</tr>
<tr>
<td>- Anything with or without ‘very’ can be considered synonymous.</td>
<td>- Would you be so kind...to bring... the book...on the table (26/#13)(Score 3)</td>
</tr>
<tr>
<td>- Omission of ‘at all’ in item 30 is acceptable.</td>
<td>- The rest-restaurant is supposed to have good food (12/#11)(Score 3)</td>
</tr>
<tr>
<td>- Omission of adverbs that only express emphasis (e.g., always) is acceptable.</td>
<td></td>
</tr>
<tr>
<td>- That/the substitution on item 12 is acceptable</td>
<td></td>
</tr>
<tr>
<td>- Good/nice substitutions are acceptable</td>
<td></td>
</tr>
<tr>
<td>• Examples of unacceptable substitutions or omissions (SCORE 2):</td>
<td></td>
</tr>
<tr>
<td>- cigar smoking&gt; smoking</td>
<td>- The number of people who smoke ...um is increasing every year (27/#10)(Score 2)</td>
</tr>
<tr>
<td>- apartment &gt;house/room</td>
<td>- He just finished painting... inside of a his house (15/#5)(Score 2)</td>
</tr>
<tr>
<td>- he&lt;&gt;she</td>
<td>- She finished a painting... inside her apartment (15/#7)(Score 2)</td>
</tr>
<tr>
<td>- sense of humor&gt; humor</td>
<td>- The person I'm dading is ...wonderful... humour (17/#11)(Score 2)</td>
</tr>
<tr>
<td>- finished cleaning&gt;cleaned</td>
<td>- Before he get outside...he must clean his room (23/#9)(Score 2)</td>
</tr>
<tr>
<td>- order&gt; eat</td>
<td>- She always eat...meat...nev-never eat vegetable (18/#5)(Score 2)</td>
</tr>
<tr>
<td>- nice,big &gt; big</td>
<td></td>
</tr>
<tr>
<td>- AUX cannot be omitted (can go&gt; go)</td>
<td></td>
</tr>
<tr>
<td>- a lot of Noun&gt; 0 Noun</td>
<td></td>
</tr>
<tr>
<td>- too Adj &gt; 0 Adj</td>
<td></td>
</tr>
<tr>
<td>- Changes in grammar that don’t affect meaning should be scored as 3. For instance, failure to supply past tense (had&gt;have) and missing articles should be considered grammar change only (score 3).</td>
<td></td>
</tr>
<tr>
<td>- By contrast, cases of extra marking or more marked morphology should be considered as meaning change. For example, a present tense repeated as past or as future should be scored as meaning change (score 2).</td>
<td></td>
</tr>
<tr>
<td>- The restaurant was supposed to have ve- good food.(12/#24)(Score 2)</td>
<td></td>
</tr>
<tr>
<td>- After the dinner I will have a long peaceful nap. (7/#8)(Score 2)</td>
<td></td>
</tr>
</tbody>
</table>
Similarly, singular/plural differences between stimulus and repeated string change the meaning, not only the grammar (score 2).

Changes of person (he for she or she for he) change the meaning; but problems of agreement (she...her versus she...his) should be considered grammatical change, not meaning change.

Ambiguous changes in grammar that COULD be interpreted as meaning changes from a NS perspective should be scored as 2. That is, as a general principle in case of doubt about whether meaning has changed or not, score 2.

A contraction (e.g., I’ve > I have, it’s > it is) and its non-contracted form should be treated as synonymous (Score 3). However, if the use of a contraction makes the meaning of the sentence ambiguous (e.g., ‘The red book’s on the table’ or ‘The red books on the table’?), score 2.

**SCORE 4**

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact repetition: String matches stimulus exactly. Both form and meaning are correct without exception or doubt.</td>
<td>- The street in the city is wide (3/#8)(Score 2)</td>
</tr>
<tr>
<td></td>
<td>- She just finished painting ...his room inside (15/#14) (Score 2)(apartment is missing)</td>
</tr>
<tr>
<td></td>
<td>- The streets on the city is wide (3/#23)(Score 2)</td>
</tr>
<tr>
<td></td>
<td>(We can’t know whether the number agreement is just a grammar problem or an interpretation problem, but string is ambiguous in meaning: (a) a generic plural statement or (b) a statement about one street (score 2).</td>
</tr>
<tr>
<td></td>
<td>- It is possible that it’ll rain tomorrow (8/EL36) (Score 3)</td>
</tr>
</tbody>
</table>

**Notes:**
1. If a participant starts repeating a sentence before the beep and do not restart after the beep, score as N/A.
2. Even if a pronunciation of one sound is not native-like, no points will be taken off, unless it severely affects the meaning of the sentence (e.g., light < right). Typically in the Japanese-L1 participants’ responses, these sounds include: /s/ vs. /ʃ/, /ŋ/ vs. /n/ vs. /ŋ/, /ŋ/ vs. /n/, /i/ vs. /i/, /w/ vs. /w/.
REFERENCES


Norris, J. M. (2011, November). Classroom realities and their implication for TBLT research. In M. Bygate (Organizer), *Tasks in classrooms: Developing TBLT as a researched pedagogy*. Colloquium presented at the 4th International Conference on Task-Based Language Teaching, Auckland, NZ.


