THE EFFECTS OF EXPLICIT INSTRUCTION AND CORRECTIVE FEEDBACK ON LEXIS AND COHESION WITH EFL LEARNERS

A Dissertation Submitted
to the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY

By
Junko Yamaai
December 2023

Examining Committee Members

David Beglar, Advisory Chair, Teaching and Learning
Tomoko Nemoto, Teaching and Learning
Edward Schafer, External Member, Ochanomizu University
James Elwood, External Member, Meiji University

`

ABSTRACT

The primary purpose of the present study was to examine the effect of explicit instruction and feedback on the development of lexis and cohesion in second language writing as well as the extent to which these two measures correlated with overall writing quality scores. The participants were two intact classes of first-year Japanese university students who attended class twice a week for 14 weeks. The classes were randomly assigned to an explicit instruction group (n = 34) and a comparison group (n = 31). The explicit group received instruction and feedback on the use of target vocabulary and cohesion, while the comparison group only submitted drafts. Both groups produced three drafts of a comparison-contrast essay for Task 1 and a problem-solution essay for Task 2 during one academic semester. Based on their writing performance, six participants were selected for interviews about their learning experiences to complement the results of the quantitative analyses.

Writing samples from each participant were analyzed with TALLES (Kyle & Crossley, 2015), TAALED 1.41 (Zenker & Kyle, 2021), and TAACO 2.0 (Crossley et al., 2016). Each draft was assessed for the number of target words, three lexical complexity metrics— MATTR 50 content words (MATTR), COCA academic frequency logarithm content words (COCAFrq), and COCA academic trigram MI2 (Trigrams) —two cohesion indices—All connectives and Adjacent overlap between paragraphs content words (AdjacentP)—and overall writing quality produced with a writing rubric. Six raters trained in applied linguistics assessed the participants' essays, and FACETS 3.80.0 (Linacre, 2017) was used to produce interval measures of student ability and rater severity.

Five hypotheses were assessed. Hypothesis 1 predicted that the explicit instruction group would score higher than the comparison group on target word and cohesion use, and overall writing quality. The results supported Hypothesis 1 for both tasks. The explicit instruction group increasingly used target words and cohesion based on descriptive statistics. Both groups significantly improved overall writing quality over time. The explicit instruction group significantly scored higher. The comparison group was slower in learning about writing essays during Task 1. Both groups scored higher in Task 2.

Hypothesis 2, which predicted that both groups would improve on the lexical indices and that the explicit instruction group would exceed the comparison group, was partially supported. MATTR and COCAFrq significantly improved, but Trigram did not improve. There were significant group differences in COCAFrq and Trigrams, but not in MATTR. The comparison group scored higher on COCAFrq, and the explicit instruction group scored higher on Trigram. In Task 2, MATTR improved significantly, but COCAFrq and Trigrams did not. There were no group differences. More effects were found in Task 1 than in Task 2.

Hypothesis 3, which predicted that cohesion would improve over time for both groups and that the explicit instruction group would exceed the comparison group, was mostly supported. The results of Task 1 showed that All connectives and AdjacentP significantly improved. Although there was no group significant difference in All connectives, there was a group difference in AdjacentP, as the explicit instruction group scored higher. The results of Task 2 showed that All connectives did not significantly

improve over time, but AdjacentP did. Global cohesion was affected more than local cohesion.

Hypothesis 4, which predicted that lexis would be positively correlated with overall writing quality for both groups, was supported for Task 1 and not supported for Task 2. In Task 1, MATTR was significantly positively correlated with overall writing quality. COCAFrq was significantly negatively correlated. Trigrams were not correlated. In Task 2, none of lexical indices were correlated.

Hypothesis 5 predicted that cohesion would be positively correlated with overall writing quality. This hypothesis was partially supported in Task 1 and Task 2. All connectives were not correlated with overall writing quality, but AdjacentP was significantly and positively correlated with overall writing quality in Task 2.

The study contributes to the understanding of the development and assessment of lexis and cohesion using computational automated tools. L2 development is a complex phenomenon, so a further examination of assessment indices offers a wealth of research in future studies.

ACKNOWLEDGMENTS

The completion of this dissertation would not have been possible without the support of numerous individuals, and I would like to thank them in this acknowledgement section.

First, I'm very grateful for the incredible support that Dr. David Beglar has given me. I cannot thank him enough for his continued guidance, insightful suggestions, and support to develop this dissertation. I appreciate his thoughtful feedback and his generosity with his valuable time. I truly appreciate that he took his time to help me improve my dissertation study.

I would also like to extend my thanks to Dr. Tomoko Nemoto for constructive and detailed feedback on statistical analyses in her busy days. I would also like to thank Dr. Edward Schaefer and Dr. James Elwood for their comments and feedback, as members of my dissertation committee. Their guidance transformed this study into a much richer final product. I would also like to thank Dr. Katerina Petchko for her feedback in the early days of this study. I would also like to express my gratitude to raters for taking their time to evaluate the participants' essay during their busy schedule. I would also like to express my appreciation to my students. I would also like to thank Satomi Yoshimuta and Naoko Hosoda for their support and encouragements while writing this dissertation.

TABLE OF CONTENTS

PAGE

ABSTRACT iii

ACKNOWLEDGMENTS vi

TABLE OF CONTENTS vii

LIST OF TABLES xii

LIST OF FIGURES xv

CHAPTER

1. INTRODUCTION 1

    Background of the Issue 1

    Statement of the Problem 3

    Purposes and Significance of the Study 5

    The Audience for the Study 6

    Delimitations 7

    Organization of the Study 8

2. REVIEW OF THE LITERATURE 9

    Input 9

    Output 11

    Explicit Instruction 12

    Feedback and Second Language Writing Development 14

    The Effectiveness of Different Vocabulary Learning Activities 18

    Explicit Productive Vocabulary Instruction 21

    Vocabulary Development Studies Using Lexical Indices 23

    Summary of Vocabulary Studies 33

    Cohesion Studies 34

Summary of Cohesion Studies....42
Linguistic Variables and Overall Writing Quality....43
Summary of the Literature Review....47
Gaps in the Literature....48
Purposes of the Study....50
Research Hypotheses....52
3. METHODS....57
Context of the Study....57
Participants....58
Instrumentation....58
Background Questionnaire....59
Writing Tasks....59
Writing Rubrics....59
Interview Protocols....60
Research Design....60
Procedures....61
Writing Task 1: Comparison-Contrast Essay....61
Writing Task 2: Problem-Solution Essay....64
Analyses....66
4. PRELIMINARY ANALYSES....68
FACETS Analyses for Task 1 Draft 1....68
FACETS Analyses for Task 1 Draft 2....74
FACETS Analyses for Task 1 Draft 3....78
FACETS Analyses for Task 2 Draft 1....83
FACETS Analyses for Task 2 Draft 2....87

FACETS Analyses for Task 2 Draft 3 92

5. RESULTS 97

Measures 98

The Moving-Average Type-Token Ratio (MATTR) 98

COCA Academic Frequency (COCAFrq) 98

COCA Academic Trigrams (Trigrams) 98

All Connectives 99

Adjacent Overlap Between Paragraphs (AdjacentP) 99

Overall Writing Quality 99

Results of Analyses for Each Hypothesis 99

Hypothesis 1: Target Vocabulary, Target Cohesion, and Overall Writing Quality 99

The Total Number of Words Written 100

Target Word Use 101

Target Cohesion Use 102

Univariate Normal Assumptions and Descriptive Statistics 104

Overall Writing Quality 104

Initial Group Difference Check for Overall Writing Quality 106

Main Analysis of Overall Writing Quality for Hypothesis 1 107

Summary of the Results for Hypothesis 1 110

Hypothesis 2: The Development of Lexical Complexity 110

Univariate Normal Assumptions and Descriptive Statistics 110

Main Analysis for Hypothesis 2 for Task 1 117

Main Analysis for Hypothesis 2 for Task 2 121

Summary of the Results for Hypothesis 2 124

Hypothesis 3: The Development of Cohesion 125

Univariate Normal Assumptions and Descriptive Statistics.....125

Main Analysis for Hypothesis 3 for Task 1.....129

Main Analysis for Hypothesis 3 for Task 2.....132

Summary of the Results for Hypothesis 3.....135

Hypothesis 4: The Relationship Between Lexical Complexity and Overall Writing Quality.....135

Main Analysis for Hypothesis 4 for Task 1.....135

Main Analysis for Hypothesis 4 for Task 2.....136

Hypothesis 5: The Relationship Between Cohesion and Overall Writing Quality.....137

Main Analysis for Hypothesis 5 for Task 1.....137

Main Analysis for Hypothesis 5 for Task 2.....138

Summary of the Results for Hypothesis 5.....138

6. DISCUSSION.....140

The Effects of Explicit Instruction on Target Vocabulary, Target Cohesion, and Overall Writing Quality.....140

The Development of Lexical Complexity.....144

The Development of Cohesion.....147

Lexical Complexity and Overall Writing Quality.....149

Cohesion and Overall Writing Quality.....150

Pedagogical Implications.....151

7. CONCLUSION.....154

Summary of the Findings.....154

Limitations of the Study.....156

Suggestions for Future Research.....157

Concluding Comments.....158

REFERENCES 159

APPENDICES 173

A. SYLLABUS (JAPANESE VERSION) 174

B. SYLLABUS (ENGLISH VERSION).....177

C. CONSENT FORM FOR THE ENTIRE STUDY (JAPANESE VERSION) 180

D. CONSENT FORM FOR THE ENTIRE STUDY (ENGLISH VERSION).....181

E. CONSENT FORM FOR INTERVIEWEES (JAPANESE VERSION) 182

F. CONSENT FORM FOR INTERVIEWEES (ENGLISH VERSION).....183

G. BACKGROUND QUESIONNAIRE (JAPANESE VERSION) 184

H. BACKGROUND QUESIONNAIRE (ENGLISH VERSION) 185

I. WRITING RUBRICS 186

J. INTERVIEW PROTOCOL (JAPANESE VERSION).....188

K. INTERVIEW PROTOCOL (ENGLISH VERSION) 189

L. TEACHING MATERIALS 190

M. MODEL ESSAYS 194

N. REVISION CHECKLIST (OSHIMA & HUGUE, 2007) 198

O. FACETS COMMAND FILE 199

LIST OF TABLES

Table Page

1. Research Procedures....62
2. Rater Measurement Report for Task 1 Draft 1....70
3. Student Measurement Report for Task 1 Draft 1....71
4. Draft Measurement Report for Task 1 Draft 1....72
5. Criteria Measurement Report for Task 1 Draft 1....73
6. Category Statistics for Task 1 Draft 1....73
7. Rater Measurement Report for Task 1 Draft 2....75
8. Student Measurement Report for Task 1 Draft 2....75
9. Draft Measurement Report for Task 1 Draft 2....77
10. Criteria Measurement Report for Task 1 Draft 2....78
11. Category Statistics for Task 1 Draft 2....78
12. Rater Measurement Report for Task 1 Draft 3....80
13. Student Measurement Report for Task 1 Draft 3....80
14. Draft Measurement Report for Task 1 Draft 3....82
15. Criteria Measurement Report for Task 1 Draft 3....82
16. Category Statistics for Task 1 Draft 3....83
17. Rater Measurement Report for Task 2 Draft 1....83
18. Student Measurement Report for Task 2 Draft 1....85
19. Draft Measurement Report for Task 2 Draft 1....86
20. Criteria Measurement Report for Task 2 Draft 1....87
21. Category Statistics for Task 2 Draft 1....87
22. Rater Measurement Report for Task 2 Draft 2....89
23. Student Measurement Report for Task 2 Draft 2....89

24. Draft Measurement Report for Task 2 Draft 2 91
25. Criteria Measurement Report for Task 2 Draft 2 91
26. Category Statistics for Task 2 Draft 2 92
27. Rater Measurement Report for Task 2 Draft 3 92
28. Student Measurement Report for Task 2 Draft 3 94
29. Draft Measurement Report for Task 2 Draft 3 95
30. Criteria Measurement Report for Task 2 Draft 3 96
31. Category Statistics for Task 2 Draft 3 96
32. Summary of the Final Dependent Measures With Index Names 97
33. The Total Number of Words Written in Task 1 100
34. The Total Number of Words Written in Task 2 101
35. Descriptive Statistics for Target Word Use in Task 1 101
36. Descriptive Statistics for Target Word Use in Task 2 102
37. Descriptive Statistics for Target Cohesion Use in Task 1 103
38. Descriptive Statistics for Target Cohesion Use in Task 2 103
39. Descriptive Statistics for Overall Writing Quality in Task 1 104
40. Descriptive Statistics for Overall Writing Quality in Task 2 105
41. Descriptive Statistics for Overall Writing Quality Difference Scores in Task 1 105
42. Descriptive Statistics for Overall Writing Quality Difference Scores in Task 2 106
43. Descriptive Statistics for MATTR in Task 1 111
44. Descriptive Statistics for COCAFrq in Task 1 111
45. Descriptive Statistics for Trigrams in Task 1 112
46. Descriptive Statistics for MATTR in Task 2 112
47. Descriptive Statistics for COCAFrq in Task 2 113
48. Descriptive Statistics for Trigrams in Task 2 113

49. Descriptive Statistics for MATTR Difference Scores in Task 1 114
50. Descriptive Statistics for COCAFrq Difference Scores in Task 1 114
51. Descriptive Statistics for Trigrams Difference Scores in Task 1 115
52. Descriptive Statistics for MATTR Difference Scores in Task 2 116
53. Descriptive Statistics for COCAFrq Difference Scores in Task 2 116
54. Descriptive Statistics for Trigrams Difference Scores in Task 2 117
55. Descriptive Statistics for All Connectives in Task 1 125
56. Descriptive Statistics for AdjacentP in Task 1 126
57. Descriptive Statistics for All Connectives in Task 2 126
58. Descriptive Statistics for AdjacentP in Task 2 127
59. Descriptive Statistics for All Connectives Difference Scores in Task 1 127
60. Descriptive Statistics for AdjacentP in Task 2 128
61. Descriptive Statistics for All Connectives Difference Scores in Task 2 128
62. Descriptive Statistics for AdjacentP Difference Scores in Task 2 129
63. Pearson Correlations Between Lexical Complexity and Overall Writing Quality in Task 1 136
64. Pearson Correlations Between Lexical Complexity and Overall Writing Quality in Task 2 137
65. Pearson Correlations Between Cohesion and Overall Writing Quality in Task 1 ...138
66. Pearson Correlations Between Cohesion and Overall Writing Quality in Task 2 ...138

LIST OF FIGURES

Figure Page

1. Wright Map for Task 1 Draft 1 69
2. Wright Map for Task 1 Draft 2 74
3. Wright Map for Task 1 Draft 3 79
4. Wright Map for Task 2 Draft 1 84
5. Wright Map for Task 2 Draft 2 88
6. Wright Map for Task 2 Draft 3 93
7. Overall Writing Quality for Task 1 109
8. Overall Writing Quality for Task 2 109
9. MATTR for Task 1....120
10. COCAFrq for Task 1....120
11. Trigrams for Task 1....121
12. MATR for Task 2....123
13. COCAFrq for Task 2....123
14. Trigrams for Task 2....124
15. All Connectives for Task 1....131
16. AdjacentP for Task 1....132
17. All Connectives for Task 2....134
18. AdjacentP for Task 2....134

CHAPTER 1

INTRODUCTION

Background of the Issue

Writing in a second language (L2) is a complex, challenging task (Doró & Pietilä, 2015) that requires a great deal of study and practice to develop. Learners must gain linguistic knowledge about L2 vocabulary, multiword units, and morpho-syntax, and practice combining these elements into larger chunks of discourse to express meaning. In addition to these linguistic elements, generating ideas and thinking critically are necessary aspects of skilled writing (McKeown et al., 2017). Moreover, learners need to understand how ideas are connected to one another, how to use cohesive devices, and how to support topic sentences to convey ideas coherently. In addition to these writing-related processes, learners must also learn to use self-regulation strategies, such as goal setting, planning, and evaluating. Learners become skilled writers by orchestrating these processes and by learning how to communicate their ideas appropriately and precisely.

Vocabulary is a particularly important element of language. Without vocabulary, language comprehension and production are not possible (Treffers-Daller & Milton, 2013). If learners have a large vocabulary, they can understand reading and listening passages better, and they can express their ideas when speaking and writing more easily. Stæhr (2008) found that learners' receptive vocabulary size was strongly associated with their reading and writing abilities and moderately associated with their listening ability. Nation (2015) stated that a rich vocabulary makes language skills such as writing easier to perform.

As learners acquire vocabulary, they can also expand their depth of lexical knowledge by learning to match the form and meaning of words, learn multiple meanings of each word, and become familiar with the different forms that words can take (Nation, 2001; Schmitt, 2008). As they acquire vocabulary over time, they can select more appropriate words that allow them to express their ideas more precisely. This knowledge helps them perform better in writing as well as in other skills. Fluency and confidence in learning the target language also improve as learners increase their vocabulary size. Hence, teaching vocabulary is important, especially for lower-proficiency learners.

In addition to vocabulary, cohesion is a basic element of language (Halliday & Hassan, 1976). Cohesion devices connect ideas, so they help speech and writing flow more smoothly and increase listeners' and readers' comprehension. The effective use of cohesive devices helps learners express their ideas more precisely and clearly, and as such, is related to L2 language proficiency.

Japanese secondary school learners' English-language studies are focused on grammar and reading, as they are the main skills needed for university entrance examinations. As a result, they spend relatively little time developing the productive skills of writing and speaking. In my experience, when first-year Japanese university students write a single paragraph or multiple paragraphs, their limited lexical knowledge inhibits their ability to plan what to write and restricts what they can express. This weakness is important because among the many elements that make up a language, vocabulary is a central unit of language used to express ideas. Learners with larger vocabularies can generally express themselves more precisely and clearly than those with smaller lexicons, and having a large vocabulary often indicates greater L2 proficiency

(Read, 2000). Intermediate or advanced writing skills cannot be developed without a strong foundation of productive vocabulary knowledge.

Notwithstanding the importance of lexical knowledge, little attention has been paid to vocabulary instruction in Japanese university English-language curricula. In addition, because vocabulary acquisition is a relatively slow process (Folse, 2006), lexical interventions are needed. Through teaching learners about vocabulary and cohesive devices explicitly or by using awareness-raising activities, learners can arrive at a better understanding of vocabulary and its role in creating more cohesive texts.

Statement of the Problem

Four problems are addressed in this study. The first problem concerns teaching productive vocabulary and cohesion. Explicit teaching is an effective way to teach and learn second language vocabulary (Hinkel, 2002; Nation, 2013; Read, 2000), but few studies have been conducted with low-proficiency learners (Hennebry et al., 2017; Nation, 2005; Norris & Ortega, 2000). How explicit instruction and feedback influence the development of written productive vocabulary and cohesion in L2 writing with low-proficiency learners is a gap that needs to be addressed.

The second problem concerns lexical development and its assessment in L2 writing over multiple drafts. An increasing number of researchers have investigated lexical development in L2 writing using a variety of lexical indices (Bulté & Housen, 2014; Crossley et al., 2016; Mazgutova & Kormos, 2015). However, few of them have looked at the development of written productive vocabulary over multiple drafts. In one such study, Muncie (2002) reported that when writing several drafts of a composition, learners tended to use more lower-frequency words in their final draft. More studies are

needed to identify effective pedagogical approaches and to investigate which lexical indices are sensitive to lexical development in writing.

The third problem concerns the lack of L2 writing studies in which the development of cohesion has been examined. Cohesion is related to the use of vocabulary and is a key component in producing texts with clearly connected ideas. In studies conducted to examine the development of cohesion in L2 writing (Connor, 1990; Crossley & McNamara, 2012; Crossley et al., 2016; Witte & Faigley, 1981), most studies have shown that higher-proficiency writers use more cohesive devices than lower-proficiency writers (Connor, 1990; Crossley et al., 2016; Witte & Faigley, 1981). However, Crossley and McNamara (2012) showed that more proficient learners in Hong Kong did not use cohesive devices; thus, because the results to date are mixed, this issue needs to be studied further.

The fourth problem concerns how lexical and cohesion metrics are related to overall writing quality scores. Some researchers have analyzed the relationship between syntactic complexity, accuracy, and holistic scores (Connor-Linton & Polio, 2014; Knoch et al., 2015; Marlowe, 2019; Ruiz-Funes, 2015), while others have examined the relationship between syntactic complexity and holistic scores (Bulté & Housen, 2014; Crossley & McNamara, 2014; Marlowe, 2019; Yang et al., 2015). However, a limited number of researchers have examined the relationship among lexis, cohesion, and holistic writing scores with L2 writers (Crossley & McNamara, 2012; Crossley et al., 2016; Garner et al., 2020; Guo et al., 2013).

Purposes and Significance of the Study

The first purpose of this study was to examine the effect of explicit instruction and feedback on the development of written productive vocabulary and cohesion in L2 writing. The results can shed light on this issue and thereby help researchers and educators understand how vocabulary and cohesion contribute to writing development. In most EFL contexts, opportunities to write are limited, so providing effective instruction and corrective feedback are two ways that educators can support L2 learners' attempts to develop writing skills while gaining a better understanding of L2 writing development.

The second purpose is to investigate the characteristics of productive vocabulary development by low-intermediate EFL learners over three drafts. Previous empirical results have shown that the use of lexis in writing tasks can be improved. For instance, Storch and Tapper (2009) showed that significantly more academic words were in students' essays at the end than at the beginning of an EAP course. Frear and Bitchener (2015) reported a significant increase in lexical complexity measured by the mean segmental type-token ratio, and Bulté and Housen (2014) found an increased use of lower-frequency words in an investigation of lexical and syntactic complexity changes over four months with upper-intermediate ESL learners. Mazgutova and Kormos (2015) assessed lexical density, diversity, and sophistication and found that lower-proficiency groups made significant improvements on all measures of lexical diversity. Muncie (2002) reported that when writing several drafts of a composition, learners used more lower-frequency words on their final draft. Notwithstanding the knowledge provided by the above studies, this issue needs to be investigated in a wider variety of educational contexts to fully understand the nature of written productive vocabulary learning.

The third purpose is to track changes in the participants' ability to produce cohesive texts over one academic semester. Research has generally shown that higher-proficiency writers use more cohesive devices than lower-proficiency writers (Connor, 1990; Witte & Faigley, 1981), though not all investigations have supported that conclusion (e.g., Crossley & McNamara, 2012). In this study, because the focus is on low-proficiency EFL learners, a group that has been under-researched, the results can shed light on the development of cohesion in L2 writing with such learners.

The fourth purpose is to examine the degree to which the use of lexis and cohesion is related to overall writing quality. This purpose can shed light on (a) the relationship between overall writing quality and the use of lexis and cohesive devices and (b) which lexical indices and cohesive devices are most strongly related to overall writing quality. Understanding these issues can help teachers and researchers when assessing learners' essays.

The Audience for the Study

The results will benefit L2 writing researchers by providing results concerning the effects of explicit instruction and explicit feedback on the development of productive vocabulary and cohesion by Japanese EFL learners. The results will also allow L2 researchers to better understand the development of lexis and cohesion by Japanese EFL learners over multiple drafts and will provide data concerning how lexis and cohesion are related to L2 writing proficiency.

The results can also help curriculum developers and classroom instructors produce writing curricula focused on improving learners' L2 writing skills by considering the role of explicit instruction and feedback on the development of lexis, cohesion, and

overall writing quality. Curriculum designers can see one way to include instruction and feedback concerning the development of productive vocabulary in course syllabi. If educators and researchers better understand the development of productive vocabulary and cohesion in L2 writing, they can design more effective instructional materials that can help learners increase the diversity of their written productive vocabulary and their use of cohesive devices.

Delimitations

Delimitations describe the degree to which the results of the present study can and cannot be generalized to other contexts. Given that definition, this study is subject to four main delimitations. The first delimitation concerns the research design. This study is a quasi-experimental study because of a lack of random sampling and group assignment, so the results should be generalized to other contexts with caution. Second, the participants' L1 was Japanese, so the results might not be applicable to learners with different L1s, particularly those genetically related to English, given that Japanese is unrelated to English (Eckstein & Ferris, 2017). This situation creates numerous learning difficulties specific to Japanese learners of English. Third, the learners were approximately 20 years of age, so generalizing the results to much younger or older learners should be done with caution, as they could differ in a number of important ways such as working memory capacity and motivation to learn (DeKeyser, 2007; Kellogg, 2001). Fourth, the participants' English proficiency was relatively low, so the results from relatively low learner groups can be applied to beginner level learners, but cautiously to learners with higher English proficiency.

Organization of the Study

In Chapter 2, Review of the Literature, I review literature concerning issues such as input and output, explicit instruction, feedback and second language writing development, the effectiveness of explicit vocabulary learning activities, explicit productive vocabulary instruction, vocabulary development studies using lexical indices, cohesion studies, linguistic variables and writing quality scores. At the end of the Chapter, I list the gaps in the literature, the purposes of the study, and the research hypotheses that guide this study. In Chapter 3, Methods, I describe the context of the study, the participants, instrumentation, research design, procedures, coding schemes, and analyses used to answer the research hypotheses. In Chapter 4, Preliminary Analyses, the writing quality data were analyzed using many-facets Rasch measurements with the FACETS software. In Chapter 5, Results, the results for each hypothesis are presented. In Chapter 6, Discussion, the results are interpreted and theoretical and pedagogical implications are presented. In Chapter 7, Conclusion, I summarize the results, present the limitations of the study, make suggestions for further research, and provide concluding comments.

CHAPTER 2

REVIEW OF THE LITERATURE

In this chapter, I review the literature concerning input and output, explicit instruction, feedback and second language writing development, the effectiveness of different explicit vocabulary learning activities, explicit productive vocabulary instruction, vocabulary development studies using lexical indices, cohesion studies, and writing quality scores. At the end of the chapter, I identify the gaps in the literature, present the purposes of the study, and posit the research hypotheses.

Input

Nation (2013) stated that the first strand in his four strands model is *meaning-focused input*. Input is central to successful language acquisition (Barcroft, 2015), with written input being a key source of positive evidence for language learning in formal contexts (McCarthy & Carter, 1997). For example, written texts provide learners with opportunities for acquiring vocabulary incidentally and intentionally (Hulstijn, 2002). Texts written by native speakers of the target language can provide rich information concerning how lexis is used to express ideas. Moreover, input processing can lead to intake (Barcroft, 2015), which is defined as the subset of the input that learners perceive and process (Schmidt & Frota, 1986; VanPatten, 1990). Chaudron (1985) characterized intake as involving several stages from perception, to comprehension, to assimilation into an interlanguage grammar, which he placed on a continuum from preliminary intake to final intake. Intake leads to the development and restructuring of learners' interlanguages, which is the developing linguistic system used to produce output (Barcroft, 2015). To

learn a word productively, learners must encode the word's phonological and/or orthographic forms into memory, activate an appropriate semantic representation, and map the word form onto an appropriate semantic representation (Barcroft, 2015).

For input to become intake, attention is necessary (Schmidt, 1990; VanPatten, 1990), as attention mediates access to conscious experience and plays a role in the processing of the informational content of input and in the processing of linguistic forms in memory (VanPatten, 1990). Allocating attention is a necessary and sufficient condition for encoding a stimulus into long-term memory; however, efficient retrieval depends on both the quantity and quality of attention at the time of encoding (Logan, 1988). If there is sufficient attentional space for learners to process incoming information in working memory, they can pay more attention to linguistic features. For this reason, noticing is important for learning.

Craik and Lockhart (1972) proposed that mental activity strongly affects memory formation and that there is a close relationship between depth of processing, which has been a contentious idea in the psychological literature, and the retention of information. Nonetheless, some researchers have maintained that when information is processed deeply, that is, more meaningfully, the probability of retention is higher. In particular, a semantic or cognitive analysis leads to a greater depth of processing more than physical or sensory features. In L2 studies, building on Craik and Lockhart's theory, Laufer and Hulstijn (2001) emphasized the importance of involvement in learning and proposed the involvement load hypothesis in L2 vocabulary studies. According to the hypothesis, need, search, and evaluation contribute to learners' depth of vocabulary processing and when all three are present, the likelihood of acquisition is increased. For this reason, it is

important that learners pay attention to new lexis and engage in some form of cognitive effort when learning L2 vocabulary.

In addition to the above arguments, Krashen (1985) argued that meaningful input processing is necessary to convert input to acquired knowledge, by which he meant implicit knowledge. Learners are primarily concerned with meaning and not with form, suggesting that their attention is likely to be drawn to vocabulary, particularly content words. In addition, content words are often stressed, so they are relatively easily heard and remembered.

VanPatten (1990) investigated whether learners could attend to form and meaning simultaneously by comparing four groups: Group 1 listened to a passage for content only, Group 2 listened to the passage for content and simultaneously noticed vocabulary items, Group 3 listened to the passage for content and simultaneously noted definite articles, and Group 4 listened to the passage and simultaneously noted verbal morphemes. The results showed that Group 2 performed similarly to Group 1 and those two groups performed better than Groups 3 and 4. The results suggested that learners have difficulty in attending to content and grammar simultaneously.

Output

In addition to input, pushed output is useful for vocabulary learning (Nation, 2013). This type pf output can occur when learners are pressured to make their production more comprehensible, more accurate, or more pragmatically correct. Swain (1985, 1995, 2005) proposed that input processing alone is not sufficient for developing productive skills; it is also necessary for learners to produce language, as production promotes the use of particular L2 acquisitional processes. Swain (1985) described three

functions of output: noticing, hypothesis testing, and metalinguistic reflection. The first function, noticing, encourages learners to pay attention to the meanings and linguistic expressions needed for verbal communication. This knowledge can lead them to recognize the gap between the output of proficient target language users and their interlanguage forms. In other words, learners can notice L2 linguistic forms they cannot produce or cannot produce accurately and recognize linguistic forms that are at variance with native speaker norms. Their attention can then be selectively directed to relevant input (Swain, 1998). The second function of output is hypothesis testing. Learners test target forms by using them in oral communication or writing. By producing output, learners can receive explicit or implicit feedback. If their hypothesis is incorrect, they can modify it and try again. The third function of output, metalinguistic reflection, occurs when learners reflect on their hypotheses by verbalizing them to another individual, an act that can raise their awareness of linguistic form and facilitate L2 development (Swain, 1995). Producing output in contextualized practice can help learners develop productive L2 proficiency (Muranoi, 2007). Muranoi stated that producing output promotes noticing, hypothesis formulation and testing, conscious reflection, and automatization. Automatization leads to more fluent performances (de Bot, 1996), which can free up working memory capacity and allow learners to attend to other features of the language.

Explicit Instruction

A considerable body of research has shown that explicit instruction is effective for acquiring a foreign language (Goo et al., 2015; Norris & Ortega, 2000). Two benefits of explicit instruction are that it directs learners' attention to linguistic form, thereby promoting noticing, and it provides clear, specific information about target language

forms and their use (Ellis, 2008). It is more effective and efficient than incidental learning (Schmitt, 2008). Although incidental learning occurs, the amount of learning is small (Webb et al., 2023). As learners focus conscious attention on target linguistic elements in explicit learning, they can convert input into intake by integrating the elements they have noticed and processed into their developing interlanguage rule system (Schmidt, 1990).

When discussing language-focused learning, a synonym for form-focused learning, Nation (2013) stated that numerous studies have shown that explicit vocabulary instruction is effective (Hinkel, 2002; Nation, 2013; Read, 2000), as it clarifies the forms and meanings of vocabulary more effectively than implicit approaches. In vocabulary learning, explicit learning is often contrasted with incidental learning where learners acquire language forms without an intention to do so through reading or listening. Incidental learning occurs gradually when learners focus on communication (Schmitt, 2000). If learners are to reach a vocabulary size threshold that allows incidental vocabulary learning to occur, explicit teaching is necessary (Schmitt, 2000). Various explicit teaching activities including dictionary work, word unit analysis, mnemonic devices, semantic elaboration, producing collocations and lexical phrases, and oral production have been suggested in the literature. In this regard, Sökmen (1992) listed key principles of explicit vocabulary teaching. Teachers need to help learners build a large sight vocabulary, integrate new words with old ones, encounter target words frequently, engage in deep processing, use imaging, connect new words to their personal world, use various techniques, and use independent learning. It is helpful for learners when teachers provide exercises to increase their receptive vocabulary, arrange for them to encounter

target vocabulary multiple times, make visual images of words, use the words productively, use various mnemonic techniques, and adopt independent learning.

Yousefi and Biria (2018) conducted a meta-analysis on the effectiveness of L2 vocabulary instruction to identify moderator variables. Sixteen studies published from 2004 to May 2014 were included in the analyses. The studies were focused on the acquisition of receptive vocabulary. The results showed that the effect size of the overall effectiveness of L2 vocabulary instruction was medium to large (d = 0.80). A small to medium effect (d = 0.53) was found for second language settings, while a large effect was found for foreign language settings (d = 0.96). However, there was no statistically significant difference between the two. In addition, a small effect size was found for advanced (d = 0.53) and elementary levels (d = 0.54), while a large effect size (d = 0.95) was found for intermediate L2 learners. However, again, the differences were not statistically significant. Moreover, the effect size for abstract words (d = 0.92) was larger than that for concrete words (d = 0.65), though the difference was not significant.

Feedback and Second Language Writing Development

In addition to explicit instruction, feedback is a common part of L2 writing classes, and it is provided to promote learners' awareness of their strengths and weaknesses on various aspects of their writing. Research has shown the importance of feedback and subsequent revision to composing processes in general and specifically for the development of academic literacy (Leki, 2006; Storch & Tapper, 2009). In addition, feedback helps learners notice mismatches between their interlanguage and the target language and provides learning opportunities that can result in the development of L2 knowledge. Feedback helps learners attend to and correct erroneous L2 forms, speeds up

the rate of development, helps learners develop explicit linguistic knowledge, increases their awareness of writing conventions, and contributes to the development of greater self-efficacy (Pajares, 2005).

Biber et al. (2011) conducted a meta-analysis of 306 quantitative and qualitative feedback studies from 1980 to 2007 to review the effectiveness of feedback on L1 and L2 writing development. The types of feedback were teacher, peer, and computer feedback. Twenty-three studies published from 1982 and 2007 were included for the second phase of the meta-analysis. The results showed that many instructors used teacher feedback with university students performing pedagogical tasks. In more than half of the studies, feedback was provided on form and content, while only linguistic form was the focus in 27% of the studies. Biber et al. showed that two common outcome types included writing proficiency scores (accuracy or quality) versus a measure of revising, and most outcome measures used were either holistic quality or grammatical accuracy. The results indicated that vocabulary and cohesion were less studied. The results with L2 learners showed that the effect size (Cohen's *d*) for teacher feedback was 0.96, and was 0.19 for other types of feedback in pretest-posttest designs. The effect size for teacher feedback was 0.53, and for other types of feedback was 0.52 in treatment-control designs. In addition, the effect size for teacher feedback was 0.80, and that for other types of feedback was 1.96 in pretest-posttest designs. Gains were shown in most of the outcome measures. For L2 learners, the effect size for holistic ratings of quality was 1.23, 0.51 for content, and 1.19 for grammar/form in pretest/posttest design. The effect size for holistic rating quality was 0.56, 0.86 for content, and .80 for grammar/form in treatment-control designs. Furthermore, lower-proficiency learners made greater gains from feedback than higher-

proficiency students. The effect size for low/beginning learners was 1.35, while it was 0.75 for high/advanced learners in pretest-posttest designs. The effect size for low/beginning students was 0.98, and was 0.46 for high/advanced students in treatment/control designs.

Many L2 writing researchers have investigated the effectiveness of different types of teacher feedback by comparing direct, indirect, meaning-focused, code, or metalinguistic feedback (Ferris & Roberts, 2001; Robb et al., 1986; Van Beuningen et al., 2012). Direct feedback refers to teachers' explicit correction of learners' written texts, while with indirect feedback, problems are indicated by underlining errors or using error codes. Many teacher feedback studies have been focused on grammatical accuracy. Van Beuningen et al. (2012) reported that both direct and indirect teacher feedback groups showed more improvements in syntactic accuracy than a self-editing group and a group that only engaged in writing practice without corrective feedback. Other researchers have examined metalinguistic explanations, which is a type of direct feedback (Shintani & Ellis, 2013), focused feedback on selected linguistic items, and unfocused feedback on a wide variety of linguistic forms (Bitchener & Knoch, 2010; Ellis et al., 2008; Sheen et al., 2009). Seijas and Spino (2023) examined teacher-provided direct or indirect feedback on vocabulary to L2 Spanish learners. Researchers have also examined learner perceptions of feedback (Matsuo, 2019; Vasu et al., 2016) or teacher perceptions of feedback (Bastola & Hu, 2023), and the effects of feedback on the number of revisions (Merkel, 2018; Min, 2006). Generally, learners improve their writing after receiving feedback (Bitchener, 2008; Bitchener & Knoch, 2010; Ferris & Roberts, 2001; Sheen et al., 2009).

Vocabulary has been the focus of feedback studies relatively rarely. Mazgutova and Kormos (2015) investigated the development of lexical and syntactic complexity in a one-month intensive academic English program at a British university with 39 Chinese L1 participants. The participants wrote two argumentative essays and received feedback. Explicit instruction was not provided, but the participants received feedback on various linguistic items from tutors. The content of the feedback was not explained in the study. The researchers found that feedback on essays contributed to improvement in the use of target language vocabulary. The Academic Word List (Coxhead, 2000) vocabulary was significantly different for the higher-proficiency group. In the lower-proficiency group, all five measures of lexical diversity differed significantly. In addition, the lower-proficiency group showed more improvement on the lexical measures.

There are relatively few feedback studies on cohesion, even though cohesion and coherence are essential elements of effective writing (Witte & Faigley, 1981). This situation might be due to the complexities of assessing cohesion (Knoch, 2007). Earlier researchers examined teacher feedback on essay revision and accuracy (Lee et al., 2021). Relatively few studies have been focused on cohesion and coherence. Kawamoto (2017) investigated whether teacher feedback affected the use of cohesive devices with lower-proficiency Japanese EFL learners. They wrote four 250- to 300-word expository essays. The results showed that corrective feedback did not significantly affect the use of cohesive devices in both the short- and long-term. Chen and Cui (2022) compared the effectiveness of automated written evaluation feedback and peer feedback on cohesion and coherence. Chinese learners wrote two drafts and completed questionnaires. Coh-

Metrix 3.0 was used to assess the use of cohesive devices. The results showed that peer feedback was preferred to the automated written evaluation for cohesion and coherence.

The Effectiveness of Different Vocabulary Learning Activities

Two ways of analyzing vocabulary activities have been proposed: the involvement load hypothesis (Laufer & Hulsitjn, 1995) and technique feature analysis (Nation & Webb, 2011). The involvement load hypothesis includes three elements: need, search, and evaluation. Each element can be either absent or present in a task. Need concerns whether learners need to understand a word or not. Search concerns whether they have to retrieve or search for new information concerning a word, and evaluation involves comparing words and making choices. As such, the involvement load hypothesis is related to depth of processing.

Nation and Webb's (2011) technique feature analysis involves examining goals, learning conditions, signs, and design features to improve vocabulary teaching activities. Nation and Webb produced a checklist for technique feature analysis that includes criteria for assessing motivation, noticing, retrieval, generation, and retention. Nation and Webb also listed vocabulary learning activities including learning vocabulary from word cards, learning vocabulary using the keyword technique and word parts, and guessing word meaning from context.

Instructional activities can increase learners' awareness of vocabulary and their ability to use it productively. Young-Davy (2014) suggested that explicit vocabulary instruction exerts positive effects on L2 classroom vocabulary learning. For example, it provides increased opportunities for exposure to target vocabulary and raises learners' awareness of the need to expand their vocabularies. As a result, learners are more aware

of, recognize, think about, and produce new words. As learners consciously process new words, they can gradually gain confidence using them and use them creatively in new contexts. Examples of useful vocabulary activities include identifying words in reading texts they can use in writing assignments (Hinkel, 2015) and completing fill-in-the gap exercises (Folse, 2006; Laufer & Hulstijn, 2001). Such decontextualized vocabulary teaching can be beneficial, as learners can gain familiarity with how words function in sentence contexts before applying them to longer writing tasks.

Researchers have compared different types of vocabulary activities. Barcroft (2002, 2004, 2006) compared sentence writing, answering questions about word meaning, making pleasantness ratings about words, counting the number of letters in words, and copying target words during intentional vocabulary learning. Barcroft (2002) used pleasantness ratings about words and examined the effects of semantic and structural elaboration on L2 lexical acquisition with low-intermediate L2 Spanish learners in three conditions. He defined semantic elaboration as the increased evaluation of an item with regard to its meaning. Structural elaboration refers to increased evaluation of an item related to its phonemic or graphemic structure. One group rated the pleasantness of each word based on previous experiences, another group counted the number of letters in each word, and a third group had no elaboration. The dependent variables were free recall and cued recall of the target words. ANOVA results showed a significant main effect on free recall in Spanish, $F(1, 47) = 20.07$, $p < .001$, $\eta^2 = .299$. The structural elaboration group scored significantly higher than the semantic elaboration group and significantly higher than the no-elaboration group in free recall in Spanish. The ANOVA results showed a significant main effect on free recall in English. The semantic elaboration

group was significantly higher than the structural elaboration condition and the no-elaboration condition. Both results indicated that increased semantic processing can inhibit learners' amount of free recall of Spanish, although it was more facilitative for English.

Barcroft (2004) compared the effects of writing new words in sentences, word-picture repetition learning, and no-sentence writing (repetition only) with 45 L2 Spanish learners. The participants viewed Spanish nouns with their picture on a screen. In Experiment 1, in the no-sentence writing condition, the participants viewed four repetitions of each word. In the sentence writing condition, they viewed one repetition of each word and were asked to write the word in a Spanish sentence. In the picture repetition group, they only watched the screen. The results showed that writing new words in a sentence had a negative effect on an L2 productive vocabulary measure and the negative effects were maintained over time. The results indicated that writing words in sentences, which had increased semantic processing, had negative effects. The design of Experiment 2 was the same as Experiment 1, but the number of repetitions in the no-sentence writing was controlled. The participants were 22 Spanish learners of English. The results again showed that writing new L2 words in sentences negatively affected L2 lexical learning. Barcroft suggested that sentence writing can inhibit word form learning in the initial stages of L2 lexical acquisition.

The effectiveness of common written exercises for L2 vocabulary learning was investigated by Folse (2006), who compared one fill-in-the blank, three fill-in-the-blank, and original sentence writing conditions. The meaning and usage of the target words were tested with 154 university intensive English students. A repeated-measures ANOVA

showed that the three groups differed significantly from one other. The three types of exercises differed significantly, $F(2, 306) = 87.01$, $p < .001$ with the three fill-in-the-blank group outperforming the other two groups. The effect size using Cohen's *d* (Cohen, 1992) of comparisons between the one-fill-in-the blank task and the three fill-in-the blank task was 1.01, that between the one fill-in-the blank condition and original sentence writing condition was .09, and that between the three fill-in-the blank condition and original sentence writing was .91. Folse stated that multiple retrievals of the target words were a facilitative factor in L2 vocabulary learning.

Explicit Productive Vocabulary Instruction

Most English teachers in Japan have observed that Japanese EFL learners generally have larger receptive vocabularies than productive vocabularies, as their productive vocabularies are limited because of the lack of opportunities to speak and write English (Waring, 1997). Productive vocabulary does not necessarily develop in parallel with receptive vocabulary (Laufer 1998). When learners write an essay, they have trouble expressing themselves clearly and sufficiently, which can impact their language and academic performances. A lack of vocabulary also creates difficulty when solving the rhetorical and conceptual problems that are part of writing (McKeown et al., 2017). A larger vocabulary is associated with higher language proficiency. Researchers have reported that vocabulary is related to writing quality (Doró & Pietilä, 2015; McKeown et al., 2017), so developing lexical knowledge and fluency in productive vocabulary are crucial.

The effects of explicit instruction on vocabulary use in writing was investigated with 65 secondary school multi-grade and multi-L1 intermediate-proficiency ESL

learners attending a public secondary school in Vancouver by Lee (2003). The participants first took a vocabulary test, completed reading instruction and comprehension exercises on bull fighting, and wrote a composition titled *A Cruel Sport*. The results showed that 13.19% of recognized target vocabulary was used in a post-reading writing task. After explicit target vocabulary instruction, this figure increased to 63.63% in an immediate writing task, indicating that explicit vocabulary instruction contributed to an increase in productive vocabulary. In a delayed writing task, there was no significant loss in receptive and productive vocabulary, but a significant loss was observed in the newly learned vocabulary. Explicit vocabulary instruction helped the learners develop productive vocabulary in an immediate writing task and increased retention. Newly learned vocabulary was used productively in an immediate writing task.

Lee and Muncie (2006) looked at the effect of instruction leading up to a piece of writing based on the story of the Titanic. The participants watched a movie, completed a cloze exercise, read a closely related text, completed a reading comprehension activity, engaged in the interactive elicitation of vocabulary, wrote the first draft of an essay, discussed vocabulary based on a writing frame, wrote the second draft, and then received feedback. The researchers found that a large number of target words were produced in the writing task and that most of them were used in a delayed re-writing task two weeks later. The amount of exposure to the words in the pre-writing tasks was substantial and most of it involved a deliberate lexical focus. A writing task at the end of a unit, for example, in linked skills tasks, can help learners acquire productive vocabulary knowledge. These findings support the earlier study by Lee (2003).

Dimas (2011) conducted an action research study and explored explicit vocabulary instruction in an L2 English content-area course with a group of university student teachers. The researcher taught vocabulary explicitly through class activities and the students' completed vocabulary entries in a Word Study notebook. The data collection procedures included the teacher´s lesson plans, their Word Study entries, and the students' opinions about both their vocabulary learning and their notebook completion through a final interview. The findings indicated that the Word Study notebook contributed to the students' learning because they acquired a grasp of the vocabulary used in the class in terms of participating in the instructional conversations led by the professor and in-class activities.

Vocabulary Development Studies Using Lexical Indices

Definitions of lexical complexity vary, and researchers have used various indices to assess this construct. Skehan (2009) suggested including measures of lexical complexity, which he defined as diversity, sophistication, and variety, as part of the complexity, accuracy, lexis, and fluency framework. Wolfe-Quintero et al. (1998) defined lexical complexity as the use of a wide variety of basic and sophisticated words, and Bulté and Housen (2014) defined lexical complexity as lexical diversity, richness, and sophistication, which were all subsumed under lexical diversity. Lexical diversity is defined as variation in the use of word types.

Lexical diversity indices have also been used as a measure of lexical richness or lexical complexity in L2 writing assessment research (Cumming et al., 2005; Wolfe-Quintero, et al., 1998; Zenker & Kyle, 2021). Lexical richness refers to variation in and the number of word types used. Lexical sophistication is defined as the learners' use of

sophisticated words, which are less frequent words or so-called advanced words. Furthermore, Mazgutova and Kormos (2015) used a lexical diversity measure that consisted of rarity, variability, and disparity based on studies by Bulté and Housen (2014) and Jarvis (2013). Disparity is the degree of differentiation between lexical types. Some researchers have found that lexical variety is a useful measure of academic writing development (Crossley & McNamara, 2014; Staples & Reppen, 2016).

Researchers have used Guiraud's index (Guiraud, 1960) to assess lexical diversity (Bulté & Housen, 2014; De Jong et al., 2012; Verspoor et al., 2012). Guiraud's index, a transformation of the type-token ratio, is calculated by dividing the number of types by the square root of the number of tokens. In this way, text length is taken into consideration (Vermeer, 2000). De Jong et al. (2012) used this index and found large complexity effects for lexical diversity using spoken data. Complex tasks elicited more diversity than simple tasks. Bulté and Housen (2014) argued that the Guiraud index measures both diversity and productivity (the number of words produced), and they used it for written data. Bulté and Housen found that in a one-month study, the Guiraud index was the best predictor of subjective ratings with intermediate learners. Verspoor et al. (2012) used the Guiraud index and found that it was a strong discriminator among different proficiency levels.

An increasing number of researchers have used different lexical indices and have shown that second language writers' vocabulary improves over time. These results are based on advances in automated computational tools, as shown by a variety of lexical indices (Bulté & Housen, 2014; Crossley et al., 2016; Frear & Bitchener, 2015; Mazgutova & Kormos, 2015). Crossley and McNamara (2011) compared intermediate-

and advanced-proficiency learners as well as L1 writers and L2 writers. The researchers showed that lexical indices can be used to predict learners' writing proficiency by analyzing 100 L2 essays from the International Corpus of Learner English (Granger et al., 2002). L1 essays were from undergraduate university students who wrote argumentative essays. The researchers investigated lexical, syntactic, and cohesive features in the texts. The linguistic variables included cohesion (connectives, word overlap, semantic co-referentiality), lexical sophistication (word concreteness, word imageability, lexical diversity, and word frequency), and syntactic complexity (syntactic similarity and number of words before the main verb). The indices included measures of lexical co-referentiality, semantic co-referentiality (Latent Semantic Analysis), word frequency, lexical diversity, word information from the MRC Psycholinguistic database, hypernymy and polysemy values from WordNet, spatial cohesion, causal cohesion, temporal cohesion, and syntactic complexity. A discriminant function analysis was performed to identify the linguistic indices that showed significant differences between the L2 groups and the L1 group. The results showed that word hypernymy, word polysemy, lexical diversity, and stem overlap were significant. Four linguistic features—hypernymy, polysemy, stem overlap, and lexical diversity—demonstrated intergroup homogeneity between four groups of L2 writers with different L1 backgrounds. In addition, the L1 writers' vocabulary was more specific than that of the L2 writers; the L2 writers used less sophisticated, more general words, and less ambiguous words. The L2 writers' word choices also contained fewer senses. The L1 writers used more word repetition or stem repetition than the L2 writers, and while the L1 writers focused on

cohesion, the L2 learners focused on word choice. The L2 writers produced fewer overlapping stems.

Crossley et al. (2011a) investigated which lexical indices best characterize lexical competence with 100 L2 writing samples analyzed with Coh-Metrix. The researchers divided the essays into basic, intermediate, and advanced levels. A discriminant function analysis was performed to predict the participants' lexical competence using the following lexical indices: breadth of lexical knowledge (word frequency, lexical diversity), depth of lexical knowledge (hypernymy, polysemy, semantic co-referentiality, and word meaningfulness), and access to core lexical items (word concreteness, familiarity, and imageability). The results showed that the strength of various breadth of knowledge indices and indices that evaluate lexical accessibility predicted the L2 learners' language proficiency, especially at the beginning and advanced levels. Word imageability, word frequency, lexical diversity, and word familiarity were the strongest predictors of proficiency. Beginner writers used more imageable words and advanced writers produced the least imageable words, suggesting that that beginning language learners used more concrete, high-frequency words that were easily accessible. The imageability index had a large effect size, $\eta^2 = .459$, while the second strongest index, word frequency, had a medium effect size, $\eta^2 = .290$. Advanced L2 writers used more infrequent words than beginning L2 writers who produced more frequent words. The third strongest index, word familiarity, had a medium effect size, $\eta^2 = .207$. Lower-proficiency learners used more familiar words, while advanced-proficiency learners produced less familiar words. The results showed that as proficiency increased, L2

writers produced less imageable, less familiar, and more infrequent words. In contrast, as proficiency levels increased, so did lexical diversity.

Bulté and Housen (2014) investigated syntactic complexity and lexical development in L2 writing and how these measures were related to subjective ratings over one semester of an EAP course with 45 intermediate-level ESL learners. The researchers defined lexical complexity as diversity, density, richness, and sophistication. They used D (Malvern et al., 2004), which is an index of lexical diversity, as well as the Guiraud G index (Guiraud, 1960) to assess lexical richness, the variation in and number of word types used. They used the advanced Guiraud index (Daller et al., 2003) to assess lexical sophistication (i.e., advanced or less frequent words). The participants wrote essays at the beginning and end of the semester. The results showed that only the lexical sophistication measure, the advanced Guiraud index, increased, but this increase was not statistically significant. The scores on the lexical diversity (D) and lexical richness (G) measures decreased slightly and non-significantly. The subjective ratings for vocabulary increased significantly from Time 1 to Time 3 and the effect sizes were large. The largest effect size was for the vocabulary rating scale of writing quality rating rubrics (d = 1.098). The learners displayed an increased use of less frequent words; however, although lexical sophistication increased, the learners' writing did not become more lexically diverse, rich, or sophisticated.

Frear and Bitchener (2015) investigated the effects of task complexity on lexical and syntactic complexity with 34 non-native speakers of English studying at language schools in New Zealand. The intermediate-proficiency participants produced three letter-writing tasks of varying levels of task complexity. The researchers analyzed the

relationship between increases in cognitive task complexity and writing quality. A repeated-measures ANOVA showed a statistically significant difference, $F(2, 32) = 3.40$, $p = .04$ among the three tasks. The η_p^2 effect size was .176, which was considered large. Lexical variety increased between each task, but each increase was not statistically significant. A Bonferroni corrected post hoc pairwise comparison with an alpha level of .017 was conducted, and the only statistically significant mean score difference was between Task 1 (low complexity) and Task 3 (high complexity). Significant findings were found for increases in lexical complexity, while syntactic complexity decreased.

Mazgutova and Kormos (2015) investigated the characteristics of lexical and syntactic complexity development in a one-month intensive academic English program at a British university with 39 Chinese L1 participants who were planning to study various disciplines upon completion of the EAP course. Two groups participated in the study: 14 intermediate-proficiency students who were preparing for undergraduate university studies in the United Kingdom and 25 upper-intermediate participants who were planning to continue their studies at the postgraduate level. The participants wrote two argumentative essays and received feedback. In order to examine differences from Time 1 to Time 2, a Wilcoxon signed-rank test, a non-parametric equivalent to the paired sample *t*-test, was conducted. Mazgutova and Kormos assessed lexical diversity by considering rarity (i.e., lexical richness), variability (i.e., diversity), and disparity (the degree of differentiation between lexical types). Five measures were used: the measure of textual lexical diversity (MTLD), the log frequency of content words, latent semantic analyses (LSA), the Academic Word List, and squared verb variation. MTLD was computed using Coh-Metrix 3 (Graesser et al., 2011; Graesser et al., 2004). The log

frequency of content words was estimated with Coh-Metrix 2.0 (Graesser et al., 2011; Graesser et al., 2004) and was based on the CELEX lexical database corpus. The LSA index was calculated with Coh-Metrix 2.0. The Academic Word List measure was computed using Vocabprofiler BNC (Nation & Heartley, 1994). Squared verb variation was computed with Synlex (Lu, 2010).

When comparing Time 1 and Time 2 for the two groups, the Wilcoxon signed-rank tests showed that squared verb variation and the Academic Word List (Coxhead, 2000) were significantly different for the higher-proficiency group. The effect sizes were medium for both squared verb variation (Z = -3.123, p < .0022, r = -.44) and the Academic Word List (Z = - 2.222, p < .026, r = -.31). In the lower-proficiency group, all five measures of lexical diversity differed significantly. The effect sizes were large for MTLD (Z = -3.296, p < .001, r = -.52) and squared verb variation (Z = -2.731, p < .006, r = -.31), and medium for the remaining measures of lexical diversity. In addition, the lower-proficiency group showed more improvement on the lexical measures. Mazgutova and Kormos found that learners improved lexis more than syntactic features and that they used more words from the Academic Word List. In particular, the lower-proficiency group made significant improvements on all measures of lexical diversity and showed the largest gains on measures of lexical variability, which was one of the most sensitive indicators of L2 proficiency. The researchers concluded that feedback on essays can contribute to improvement in the use of target language vocabulary.

Kim et al. (2018) suggested indices for L2 proficiency, L2 lexical proficiency, and L2 writing development. The researchers examined interconnected components of lexical sophistication by aggregating 424 indices of lexical sophistication in TAALES

(tools for the automated analysis of lexical sophistication) (Kyle & Crossley, 2015). They investigated whether the macro-feature components predicted L2 writing proficiency levels, holistic lexical proficiency scores, and longitudinal lexical growth. They used the Younsei English Learner Corpus (Rhee & Jung, 2014), the Crossley written lexical proficiency corpus (Crossley et al., 2011b), and the Salisbury longitudinal spoken corpus (Salsbury et al., 2011). The Younsei English Learner corpus had a large sample size and essay quality scores. It consists of 6,572 texts with 1,085,828 words written by 3,286 Korean high school graduates (or equivalents). The participants were Korean high school graduates who had written one narrative and one argumentative essay. They were divided into nine proficiency levels based on the Common European Framework of Reference for Language, ranging from A1 (beginner) to C2 (proficient). There were 3,031 argumentative essays with a total of 770,511 words among the essays with more than 100 words. The Crossley written lexical proficiency corpus was chosen because it had L1 and L2 essays from various levels in the intensive language program of a university that were scored for lexical proficiency. The Salisbury corpus was a year-long corpus.

Regression analyses showed that bigram (e.g., *look into, of the, in the*) and trigram (*this sentence for, with an example*) strength of directional association, content word properties, bigram mutual information, bigram and trigram proportions, and word specificity explained 16.1% of the variance of L2 writing proficiency and 31.0% of the variance in lexical proficiency. Additionally, word acquisition properties and content word frequency explained 8.5% of the variance of L2 writing proficiency. In addition, six lexical components including bigram and trigram proportions, word acquisition properties, content word frequency, bigram frequency and range, content word properties,

and function word frequency and range demonstrated significant development over the academic year.

Zenker and Kyle (2021) investigated text length effects for several lexical diversity indices in a corpus of 45,442 L2 argumentative essays ranging in length from 50 to 200 tokens from the International Corpus Network of Asian Learners of English (ICNALE; Ishikawa, 2013). Various kinds of type-token ratios were used. Indices used included the type-token ratio (Johnson, 1944), Root TTR (Guiraud, 1960), LogTTR (Herdan, 1960), Maas' index (Maas, 1972), MATTR (Covington & McFalll, 2010). HD-D (McCarthy & Jarvis, 2007), MTLD Original (McCarthy & Jarvis, 2010), MTLD-MA-BI (McCarthy & Jarvis, 2010), and MTLD-MA-Wrap (McCarthy & Jarvis, 2010). Root TTR is a kind of type-token ratio called Guiraud's index. LogTTR is also a kind of type-token ratio, bilogarithmic TTR, and it is called Herdan's C. Maas' index is a more complicated transformation of TTR that matches a logarithmic curve (Zenker & Kyle, 2021). MATTR is moving-average TTR to minimize text length effects. HD-D is the hypergeometric distribution diversity index. MTLD is the measure of textual lexical diversity (MTLD Original), and it is another kind of TTR. MTLD-MA-BI is the moving-average bidirectional MTLD. MTLD-MA-Wrap is a moving-average wrapped MTLD. Proficiency levels were streamed based on the Common European Framework of Reference for Languages (Counsil of Europe, 2001) using the Vocabulary Size Test (Nation & Beglar, 2007). The results indicated that MATTR was the most stable index for estimating lexical diversity with texts ranging from 50 to 200 words. Two types of MTLD, MTLD Original and MA-Wrap, were the second-best indices, as they were also relatively stable. Zenker and Kyle (2021) recommended that researchers use MATTR

(Covington & McFall, 2010) and MTLD (McCarthy, 2005; McCarthy & Jarvis, 2010) as lexical diversity indices.

A limited number of researchers have examined vocabulary development over multiple drafts. Muncie (2002) investigated whether a process writing approach would improve the vocabulary of 30 Japanese university students. The researcher analyzed timed writing texts and the first and final drafts of a composition using the Lexical Frequency Profile (Laufer & Nation, 1995), which compared the words used in the essays to lists of the first and second 1,000 most frequent words of English. Ratios were produced. The percentage of words used in the text were shown in these two categories, as was the percentages of words that did not belong to these two categories (i.e., words above the 2,000 most frequent words in English). First, the participants wrote a timed self-introduction composition. As the lessons proceeded, the learners were taught to generate and structure ideas and wrote two drafts about friendship inside and outside of class in order to have enough time to finish the composition. The students also exchanged peer feedback. The entire process took four months and each participant produced three drafts. The results showed that few differences were found between the vocabulary used in the timed composition, and the three drafts of the composition. There was a tendency for lower-frequency words to increase between the first and final drafts and between the timed composition and the final draft of the friendship essay. When the learners revised their compositions after receiving feedback, the final drafts contained a higher percentage of more sophisticated vocabulary than the other drafts. Muncie suggested that teaching vocabulary explicitly in the pre-writing stage might promote even greater vocabulary development. Writing experiences provided the learners with more time and resources

and with experience using more sophisticated words than those used in the timed writing task. More research is needed with different language settings to investigate this issue further.

Summary of Vocabulary Studies

Although numerous explicit teaching strategies have been suggested, the number of investigations into the development of written productive vocabulary is limited. Mazgutova and Kormos (2015) indicated that instruction and feedback promote vocabulary learning, but little attention has been paid to the effects of feedback on the acquisition of vocabulary and writing skills. Providing input and pushing output in the form of writing helps learners process target vocabulary repeatedly. In addition, explicit instruction and feedback help them focus on and be more aware of vocabulary. When learners process words deeply, they are more likely to acquire them.

Researchers have investigated learners' vocabulary use and development utilizing various lexical indices. Studies have shown that vocabulary improves over time. Frear and Bitchner (2015) found that lexical variety increased, while Bulté and Housen, (2014) found that although the lexical sophistication as assessed by the Advanced Guiraud index showed a tendency to increase, learners' vocabularies did not become more diverse, richer, and more sophisticated over one month of instruction. Mazgutova and Kormos (2015) found that in lower-proficiency learners, MTLD, squared verb variation, log frequency of content words, the Academic Word List vocabulary, and LSA differed significantly. Zenker and Kyle (2021) recommended the use of MATTR and MTLD. Crossley and McNamara (2011) found that word hypernymy, word polysemy, lexical diversity, and stem overlap were significantly different for L1 and L2 writers. In addition,

L1 writers used more word repetition than L2 writers. Crossley et al. (2011) found that word imageability, word frequency, lexical diversity, and word familiarity were the strongest predictors of proficiency using Coh-Metrix and that as proficiency increased, L2 writers produced less imageable, less familiar, and more infrequent words. In contrast, as proficiency levels increased, so did lexical diversity. Kim et al. (2018) reported that bigram and trigram explained 16.1% of the variance of L2 writing proficiency and 31.0% of the variance in lexical proficiency. Word acquisition properties and content word frequency explained 8.5% of the variance in L2 writing proficiency. Bigram and trigram proportions, word acquisition properties, content word frequency, bigram frequency and range, content word properties, and function word frequency and range demonstrated significant development over one year. In sum, the following lexical indices help to explain L2 proficiency and development: lexical diversity, bigram and trigram, content word properties, word acquisition properties, content word frequency, function word frequency and range, and word repetition.

Cohesion Studies

Cohesion and coherence are essential elements of effective writing (Witte & Faigley, 1981). Vocabulary conveys meaning, connects ideas through cohesive devices and semantic similarity, and creates coherence in texts through lexical repetition. Cohesion is based on vocabulary usage, and as such, plays an important role in writing. Cohesion refers to words or phrases used to connect ideas between different parts of a text. The presence of cohesive devices such as the repetition of the same subject, the use of synonyms, and pronoun reference help writers connect ideas in ways that allow readers to move through a text smoothly. Another type of cohesion is grammatical cohesion,

which is created through sentence structures. Cohesive devices refer to conjunctive or connective expressions (Williams, 2012), which include true grammatical conjunctions (e.g., *and, but, so*), subordinating conjunctions (e.g., *because, although*, *when*), and conjunctive adverbials (e.g., *however*, *therefore*, *moreover*). Their use creates more textual coherence, which allows readers to comprehend the texts more easily. Some previous researchers have shown that higher-proficiency writers use more cohesive devices than lower-proficiency writers (Connor, 1990; Witte & Faigley, 1981).

Coherence refers to readers' understanding of a text (Crossley et al., 2016). It is based on a number of features, such as cohesion cues, prior knowledge, and reading skill (O'Reilly & McNamara, 2007). An essay can be coherent when it is appropriate in a certain context, such as the writer's purpose and readers' knowledge and expectation (Witte & Faigley, 1981). It is one of the elements of successful writing. Cohesion is a part of coherence. It can contribute to coherence, but cohesive texts can be incoherent (Witte & Faigley, 1981).

Relatively little is known where low-proficiency EFL learners are concerned. Cohesive devices use vocabulary to connect ideas and produce more coherent, easier-to-read texts. However, there are mixed findings about the relationship between learners' writing proficiency and their use of cohesive devices. As noted above, most previous studies have shown that higher-proficiency writers use more cohesive devices than lower proficient writers (Connor, 1990; Witte & Faigley, 1981), but Crossley and McNamara (2012) showed the opposite result with Hong Kong Chinese students.

Crossley et al. (2016) investigated the development of local, global, and text cohesion in 57 L2 writers in a semester-long course. They examined the relationship

between cohesive devices and human judgements of writing quality. The researchers used 30 indices from Coh-Metrix (Graesser et al., 2004) and TAACO (Crossley et al., 2016). The participants wrote a 30-minute descriptive essay at the beginning, midpoint, and end of the course. The corpus included 171 essays with an average length of 335.351 words. Two expert raters assessed the essays for content, organization, vocabulary, language use, and mechanics. In addition, automated cohesion indices for local, global, and text cohesion were used. Local cohesion indices included connectives (*because*, *so*, *although*, *whereas*, *moreover*, *and*, *or*, *first*, *until*, *also*, *however*, *but*), lexical overlap between sentences (noun overlap, argument overlap, stem overlap, and content word overlap), semantic similarity between sentences, synonym overlap between sentences, and syntactic cohesion. Global cohesion indices included semantic similarity between paragraphs, lexical overlap between paragraphs, and synonym overlap between paragraphs. The text cohesion indices included casual cohesion, spatial cohesion, givenness, lexical diversity, and semantic similarity across a text. Within-subjects ANOVAs were conducted using the Coh-Metrix and TAACO indices focusing on the initial, middle, and final essays to identify indices that demonstrated significant linear growth patterns. The indices that demonstrated significant linear growth patterns were then entered into a Support Vector Machine (supervised learning algorithm) to predict whether an essay was written at the beginning or end of the semester.

Statistical analyses were conducted to examine whether there was significant writing growth. The results of a repeated-measures ANOVA revealed a significant effect for time (initial, middle, and final essay) on the writing quality score. Follow-up pairwise comparisons showed that the scores on the initial essays were significantly lower than

those for the middle and final essays. The middle and final essays were not significantly different. A repeated-measures ANOVA also showed a significant effect for time on organization scores. Follow-up pairwise comparisons showed that the scores on the initial essays were lower than those for the middle and final essays, which were not significantly different.

Second, repeated-measures ANOVAs were performed on the Coh-Metrix and TAACO indices to examine differences between initial, middle, and final essays. The L2 writers showed growth in their use of cohesive devices in local, global, and text cohesion; 13 out of 30 indices showed significant growth. Three of them concerned global cohesion. These variables were related to lexical overlap (adjacent overlap between two paragraphs [nouns] [$\eta^2 = .25$]), semantic similarity (Latent Semantic Analysis, LSA initial to final paragraphs [$\eta^2 = .17$]), and synonym overlap (synonym overlap between paragraphs [nouns] [$\eta^2 = .25$]). Six variables concerning local cohesion were related to connectives (all positive connectives, incidence all positive causal connectives [$\eta^2 = .14$]) and lexical overlap (adjacent overlap two sentences [all words] [$\eta^2 = .19$]), adjacent overlap between two sentences [function words] [$\eta^2 = .15$]), adjacent overlap among three sentences [function words] [$\eta^2 = .15$]), and verb overlap sentences [$\eta^2 = .09$]). Four indices concerning text cohesion were related to lexical diversity (bigram TTR [$\eta^2 = .14$]), content lemma TTR [$\eta^2 = .21$]), semantic similarity (LSA between all sentences and paragraphs [$\eta^2 = .21$]), and syntactic cohesion (syntactic similarity [$\eta^2 = .11$]). In addition, pairwise comparisons between initial, middle, and final essays showed that the final essays had significantly higher levels of cohesive devices than the initial essays. The middle essays included significantly higher levels of cohesive devices than the initial

essays (except the indices for LSA initial to final paragraphs and noun synonym overlap between paragraphs). However, there were no significant differences between the middle and final essays in terms of cohesion.

Multiple regression analyses were performed to examine the degree to which the cohesive devices predicted human ratings of essay quality. The results showed that the cohesive devices were predictors of human judgments of text organization and holistic ratings. With regards to writing quality, four cohesive devices were predictors of human ratings: two global indices (adjacent overlap between paragraphs for both function words [r = .59] and nouns [r = .44]) and one local index (adjacent overlap between sentences for function words [r = .42]). The last index was text cohesion (pronoun to noun ratio [r = -.15]), which was a negative predictor of the human ratings. As for essay organization, two indices that assessed function word cohesion predicted human judgments of organization, global cohesion (adjacent overlap between paragraphs [function words]) (r = .54), text cohesion (function word TTR) (r = -.42), local cohesion (incidence of coordinating conjuncts [r = -.16]), and adjacent overlap within one sentence (pronouns) (r = -.18). The regression model indicated R = .603, R^2 = .364; thus, the cohesion indices accounted for 36% of the variance in the organization ratings of the essays.

Crossley and McNamara (2012) investigated the role of linguistic features in L2 writing proficiency by examining the linguistic features of writing samples produced by L2 writers at different proficiency levels to find out how they related to human ratings of L2 writing. They assessed text cohesion and linguistic sophistication and included linguistic measures using Coh-Metrix (Graesser et al., 2004; Graesser et al., 2011). Cohesion included logical operators, lexical overlap, temporal cohesion, semantic co-

referentiality, causality, connectives, and lexical diversity. Linguistic sophistication included psycholinguistic word ratings, hypernymy, word frequency, and syntactic complexity. The linguistic measures included conceptual knowledge, causality, temporality, and given/new information. The data were two hundred 500-word essays written by Hong Kong senior high school students for the Hong Kong Advanced Level Examination. The prompts were the popularity of comic books, wearing brand name fashion, responding to a letter of complaint, and writing a letter welcoming an exchange student.

The results showed that surface code, text base, and situation model measures related to cohesion, and linguistic sophistication predicted the essay scores. There were also large effect sizes (r = .50) between the selected lexical features and the essay scores. Multiple regression analyses showed that lexical diversity, word frequency, word meaningfulness, aspect repetition, and word familiarity significantly predicted L2 writing proficiency. These variables accounted for 30% of the variance in the multiple regression model and the lexical diversity variable (as assessed by *D*) accounted for 18% of the variance. Unexpectedly, the higher-proficiency learners used fewer cohesive devices, but they produced more linguistically sophisticated essays. In addition, higher-scored essays had less lexical overlap than lower-scored essays. Lower-proficiency writers showed more content word overlap and higher semantic similarity scores than the higher-proficiency writers. The higher-proficiency writers used fewer positive logical connectors (e.g., *and, also, in sum*), used less aspect repetition than lower-proficiency writers, and produced less given information.

Guo et al. (2013) examined whether lexical sophistication, syntactic complexity, cohesion, and text information predicted writing proficiency on the TOEFL iBT integrated and independent writing tasks. The results showed that linguistic features predicted scores on both essays, but the tasks elicited different scoring patterns. Regression analyses showed that text length was the strongest predictor of essay quality for the integrated tasks, accounting for 26.4% of the variance in human scores. The next strongest predictor for integrated essays was past participle verbs, accounting for 15.6% of the variance. Learners with higher ratings used more past participle verbs, which are often used in passive voice. The third strongest predictor was word familiarity as a measure of lexical sophistication, accounting for 8.5% of the variance. Test-takers who used less frequent and less familiar words received higher writing scores. The next strongest predictor, the use of third-person singular verbs, was present in the higher-quality essays. This variable accounted for 4.1% of the variance. The index of semantic similarity assessed through LSA, accounted for 1.4% of the variance, and verbs in the base form accounted for 1.3%. For independent tasks, text length was the most significant predictor of the writing quality scores, accounting for 47.8% of the variance. The next two strongest predictors were average syllables per word and noun hypernymy values, accounting for 9.0% and 4.8% of the variance, respectively. More proficient test-takers used more sophisticated words. The next strongest predictor was past participle verbs, accounting for 2.5% of the variance, followed by conditional connectives, a cohesion index, which accounted for 1.0% of the variance. Test-takers who used fewer conditional connectives received higher scores. In addition, test-takers who used less aspect repetition and content word overlap obtained lower scores. For similarities between the

two tasks, higher-rated test-takers often wrote longer essays, used more past participle verbs, and used less frequent words. For the differences between the two tasks, semantic similarity was a significant predictor in the integrated task, but not for the independent task. A greater number of third-person singular present forms and fewer verbs in the base form were significant predictors for the integrated tasks, but not for the independent tasks. Cohesive devices negatively affected writing quality in the independent task, but they correlated positively with writing quality.

Crossley et al. (2019) evaluated the validity of TAACO 2.0, a freely available tool for the automatic analysis of cohesion that uses the newspaper and popular magazine sections of the Corpus of Contemporary American English (COCA; Davies, 2008). TAACO 2.0 calculates the average similarity between progressive adjacent sentences or paragraphs in a text. It includes local, global, and overall text cohesion indices based on connective lists, words, lemmas, and synonym overlap, along with type-token indices. It also includes new indices related to local and global cohesion at the semantic level, a source overlap feature assessed by latent semantic analysis, Dirichlet allocation, and semantic similarity (word2vec). Dirichlet allocation assesses semantic similarity features, and it perceives related concepts that have similar topic probabilities on the basis of underlying cooccurrence patterns (Crossley et al., 2019). Word2vec is also a semantic similarity measure in which words co-occurring in similar contexts are represented as being closer and dissimilar words as being farther apart (Crossley et al., 2019).

The researchers conducted two studies using TAACO 2.0. In the first study, 35 students wrote two essays on the two prompts in 25 minutes and were provided an additional 15 minutes to develop their essays. The researchers examined how semantic

overlap features at the local and global levels could be used to assess writing quality. Expert writers added cohesion in each essay by adding word overlap across sentences and paragraphs. The results showed that the second essay had higher scores than the first essay and that the original essays with added cohesion by expert writers scored higher than both the original and elaborated essays. Essays that had greater global cohesion, as found in semantic similarity (word2vec), were scored higher. The results showed that global semantic similarity reported by word2vec was a predictor of coherence ratings.

The second study examined the extent to which source overlap between the speaking samples in the TOEFL-iBT integrated speaking tasks and the responses by test-takers predict human ratings of speaking proficiency. The results showed that the similarity between the source document and the response were significant predictors of speaking quality. Pearson correlations indicated that all semantic similarity source overlap variable had a weak effect size. Only semantic similarity (the word2vec variable) was retained because it had the strongest correlation with the speaking score (r = .369). The final linear mixed effect model showed a marginal R^2 of .016 and a conditional R^2 of .655.

Summary of Cohesion Studies

Cohesion and coherence are essential elements of effective writing (Witte & Faigley, 1981). Relatively little is known about the use of cohesion. Previous studies have shown that higher-proficiency learners generally use more cohesive devices (Connor, 1990; Witte & Faigley, 1981). Kawamoto (2017) investigated the effects of teacher feedback on cohesion with lower-proficiency Japanese EFL learners and found that there was no effect of teacher feedback on cohesive devices. With improved automated

computational tools, an increasing number of studies have been conducted on cohesion (Crossley et al, 2016; Guo et al. 2013; Crossley et al. 2019). Crossley et al. (2016) showed that local, global, and text cohesion improved over time. Crossley and McNamara (2012) found that higher-scored essays used fewer cohesive devices, and less lexical overlap than lower-scored essays. Lower-proficiency writers showed more content word overlap. Guo et al. (2013) found that cohesive devices were negatively related to overall writing quality. Crossley et al. (2019) found that when researchers added cohesion to learners' essays, their writing quality scores were higher. Essays that had greater global cohesion were also scored higher.

Linguistic Variables and Overall Writing Quality

Generally, the results of writing quality measures improve even in short-term studies (Bulté & Housen, 2014). For example, Storch (2009) showed that writing quality measured with three criteria—fluency, content, and form—improved from Time 1 to Time 3 after one semester of study at an Australian University. Crossley et al. (2016) found texts improving significantly from the initial draft to the mid draft and plateauing from the mid draft to the final draft. Other researchers have shown that learners' texts improve after three to six months of instruction (Green, 2005; Larsen-Freeman, 2006).

Several researchers have investigated the relationship between linguistic performance and writing quality scores. For instance, researchers have investigated the relationship between accuracy, complexity, and overall writing quality scores (Connor-Linton & Polio, 2014; Knoch et al., 2015; Ruiz-Funes, 2015; Yang et al., 2015), while other researchers have examined the relationship between syntactic complexity and writing quality ratings (Bulté & Housen, 2014; Crossley & McNamara, 2014; Friginal &

Weigle, 2014; Yang et al., 2015). Polio and Shea (2014) reviewed studies of accuracy measures in various journals and applied selected accuracy measures to the MSU data set to determine which measures were reliable and valid, their ability to show change over time, and their relationship with holistic measures. The researchers found significant correlations between the subjective ratings for Vocabulary and Language Use and writing accuracy measures.

An increasing number of researchers have investigated vocabulary development and writing quality. Numerous studies have shown that lexis is positively related to L2 writing proficiency. For instance, Staples and Reppen (2016) found that greater lexical variety was related to higher writing quality scores consisting of language and organization. Crossley and McNamara (2011) demonstrated that lexical variety is a measure of L2 writing development. In addition, essays that include lower-frequency words tend to receive higher scores (Guo et al., 2013, Laufer & Nation, 1995). Kim et al. (2018) found that lexical sophistication was significantly correlated with quality scores on argumentative essays. N-gram frequencies were positively correlated with proficiency and quality scores.

Staples and Reppen (2016) investigated how lexico-grammatical patterns were used in first-year writing classes across three L1s—Chinese, Arabic, and English—and two genres, Rhetorical Analysis and Long Argument. The researchers also examined the relationship of these patterns to writing quality scores. The corpus consisted of texts collected from a required first-year writing course at an American university. The researchers found that greater lexical variety, greater use of attributives, greater use of *noun + that* clauses, and less use of causative adverbial clauses were related to higher

language ratings. The strongest correlation between language rating and linguistic features was for the type/token ratio ($r = .42$, $p < .0001$), a finding that indicated that first-year writers who used a greater variety of vocabulary in their essays received higher ratings.

Bulté and Housen (2014) investigated syntactic complexity and lexical development and how these measures were related to subjective ratings of L2 writing in a one-semester EAP course with 45 intermediate-level ESL learners. The researchers defined lexical complexity as diversity, density, richness, and sophistication. They used D (Malvern et al., 2004), an index of lexical diversity, the Guiraud G index (Guiraud, 1960), and the advanced Guiraud index (Daller et al., 2003) for assessing lexical sophistication. The advanced Guiraud index is a measure of lexical sophistication that assesses advanced words (i.e., less frequent words). The participants wrote essays at the beginning and end of the semester. The results showed that only the advanced Guiraud index increased but not to a significant degree. The scores on the lexical diversity (D) and lexical richness (G) measures decreased slightly and non-significantly. The subjective rating scales for vocabulary increased significantly from Time 1 to Time 3. The strongest effect size was for the vocabulary rating scale ($d = 1.098$). Bulté and Housen (2014) found that lexical richness as assessed by the Guiraud index and the use of longer units at the clausal and sentential level were indicators of higher writing quality as shown in clausal subordination, mean length of (finite) clauses, sentences, and T-units, and the proportion of simple sentences, predicted subjective ratings. However, the syntactic complexity indices that showed significant development over time did not coincide with

the measures that correlated significantly with the subjective ratings. This finding is similar to those reported by Yang et al. (2015) and Crossley and McNamara (2014).

Verspoor et al. (2012) examined objective measures of syntactic and lexical complexity to assess L2 learners' written texts from a dynamic usage-based perspective. They assessed 437 texts written by 489 Dutch learners of L2 English that were holistically coded for proficiency level. Each text was hand coded for 64 variables distilled from the syntax and vocabulary literature at the sentence, phrase, and word levels. The results showed that word length was not a good discriminator of proficiency, while the Guiraud index was a strong discriminator, and there was a significant difference between all adjacent levels. Lexical changes took place between levels 1 and 2, syntactic changes occurred between levels 2 and 3, and both lexical and syntactic changes appeared between levels 3 and 4. The transition between levels 4 and 5 was characterized by lexical changes in particles, compounds, and fixed phrases.

Garner et al. (2020) investigated the relationship between bigram and trigram association measures and human ratings of L2 writing quality by beginner- and intermediate-proficiency L1 Korean learners of English. This corpus was made up of 6,572 written responses to the Yonsei English Placement Test (YEPT) written by 3,286 Korean EFL students. A written corpus of essays by 3,286 Korean EFL students during an English placement test was analyzed using TAALES (Kyle & Crossley, 2015), an automatic text analysis program that includes indices for lexical sophistication such as word frequency, range, psycholinguistic word information, and academic language for individual words and n-grams. The indices were taken from the Corpus of Contemporary American English (COCA; Davies, 2008) as well as the BNC as the reference corpora.

The researchers also included indices based on the five subsections of the COCA: Mutual Information (MI), Mutual Information Squared (MI^2), *t*-score, the collexeme score, and ΔP. MI is the strength of attractions between two words, and it is focused on low-frequency words. The *t*-score assesses the certainty of how two words can form a bigram, and measures high-frequency words (Garner et al., 2020). These two measures presume normal distributions for languages and do not account for the directionality of word combinations (Gries & Ellis, 2015). MI^2, the collexeme score, and ΔP and the collexeme score are solutions to these problems. The results showed that 25 out of the 30 bigram and trigram association indices differed significantly between proficiency levels. In addition, advanced learners produced more strongly associated word pairs with low-frequency words and the intermediate learners used strongly associated high-frequency words. A logistic regression model using eight bigram and trigram association indices differentiated three proficiency levels. Three measures were based on the academic subsection of the COCA and five were based on the spoken subsection. A post-hoc analysis using five bigram and trigram association strength indices differentiated between the high-beginner and high-intermediate groups. Two were bigram indices based on the academic subsection of the COCA, and three were trigram indices based on the spoken subsection of the COCA. Bigram and trigram association strength are important factors in human judgments of L2 writing proficiency.

Summary of the Literature Review

Although researchers have shown that learners' writing improves over time, little is known about the effects of explicit instruction on vocabulary and cohesion in L2 writing over multiple drafts with lower-proficiency L2 learners. In addition, there are

relatively few feedback studies on vocabulary and cohesion. There were also mixed results about whether lower-proficiency learners use more cohesive devices or not. Moreover, an increasing number of studies have been conducted to examine how vocabulary and cohesion relate to overall writing quality using advances in automated computational tools. The results showed that these are related to overall writing quality, but the participants in these studies were with higher-proficiency learners. Researchers also need to examine lower-proficiency learners concerning these issues.

Gaps in the Literature

The first gap addressed by this study concerns the effects of explicit instruction and feedback on vocabulary and cohesion development in L2 writing. Researchers have suggested that explicit teaching is effective for L2 vocabulary learning (Hinkel, 2002; Nation, 2013; Read, 2000; Schmitt, 2000), as it clarifies the forms and meanings of vocabulary more effectively than implicit approaches. Schmitt (2000) stated that explicit teaching is necessary to reach the vocabulary size threshold for incidental vocabulary learning to occur. Nation (2013) stated that the meaning of vocabulary and vocabulary use should be learned or taught explicitly. Although explicit vocabulary instruction has been investigated (Barcroft, 2002, 2004, 2006; Dimas, 2011; Folse, 2006; Yeung et al., 2020), little attention has been paid to the effects of explicit instruction on vocabulary and its effect on cohesion development in L2 writing (Lee, 2003; Muncie, 2002; Lee & Muncie, 2006). Moreover, although feedback is important in L2 writing (Leki, 2006; Storch & Tapper, 2009), it has not been systematically studied in L2 writing with lexis.

The second gap concerns the characteristics of vocabulary development by low-intermediate Japanese EFL learners over multiple drafts. Few researchers have

investigated the development of written productive vocabulary over multiple drafts. Muncie (2002) reported that when writing several drafts of a composition, learners tended to use more lower-frequency words on their final draft. However, more studies are needed with Japanese EFL learners. In recent years, an increasing number of researchers have investigated lexical development in L2 writing using a variety of lexical indices (Bulté & Housen, 2014; Frear & Bitchener, 2015; Kim et al., 2018; Mazgutova & Kormos, 2015, Zenker & Kyle, 2021). Among these studies, Frear and Bitchener (2015) reported a significant increase in lexical complexity (mean segmental type-token ratio), and Bulté and Housen (2014) found an increased use of less frequent, sophisticated words in an investigation of lexical and syntactic complexity changes over four-months with upper-intermediate ESL learners. The lexical diversity and lexical richness measures decreased slightly and non-significantly. The results of these studies have shown that lower-proficiency groups made significant improvements in all measures of lexical diversity. Further studies are needed to examine lexical development with EFL learners over multiple drafts and to investigate which lexical indices are sensitive to writing development shown by low-proficiency learners.

The third gap concerns the development of cohesion in L2 writing. Crossley et al. (2016) showed that L2 writers improved the use of cohesive devices using Coh-Metrix (Grasser et al., 2004; Grasser et al., 2011) and TAACO (Crossley et al., 2016). Among the indices that showed significant growth included connectives and lexical diversity. However, there are mixed findings about the relationship between learners' writing proficiency and their use of cohesive devices. Previous studies have shown that higher-proficiency writers use more cohesive devices than lower-proficiency writers (Connor,

1990; Witte & Faigley, 1981). Crossley and McNamara (2012) showed the opposite results with Hong Kong Chinese students. In addition, the number of studies is limited, so more research is needed, particularly with lower-proficiency L2 writers.

The fourth gap concerns the relationships among lexical metrics, cohesion, and overall writing quality scores with low-proficiency learners. Previous studies have shown that overall writing quality scores generally improve over time (Bulté & Housen, 2014; Crossley et al., 2016; Storch, 2009). In addition, an increasing number of researchers have examined the relationship between lexis and overall writing quality scores, and the results have shown that vocabulary is related to overall writing quality with ESL learners (Crossley et al., 2011; Crossley, 2012; Kim et al., 2018; Verspoor et al., 2012). Garner et al. (2020) showed that bigram and trigram association measures are related to human ratings of L2 writing quality. Furthermore, with regards to cohesion, a limited number of researchers have examined lexis, cohesion, and overall writing quality (Crossley & McNamara, 2011, 2012; Crossley et al., 2016; Guo et al., 2013). Previous studies have generally shown that higher-proficiency writers use more cohesive devices than lower-proficiency writers (Connor, 1990; Witte & Faigley, 1981). However, it is unclear which indices reflect vocabulary development for lower-proficiency EFL learners, and whether these indices are related to overall writing quality.

Purposes of the Study

The present study was conducted to examine the effect of explicit instruction and feedback on the development of L2 writing focusing on lexis, cohesion, and overall writing quality over time. In addition, how these variables relate to each other was examined. The first purpose concerned the effect of explicit instruction and feedback on

L2 writing (Guo et al., 2013; Norris & Ortega, 2000). The effect of explicit instruction and feedback was assessed for target vocabulary, cohesion, and overall writing quality over time. In most EFL contexts, input is limited, so finding effective instruction and feedback can help educators support L2 learners' attempts to develop writing skills.

The second purpose was an investigation of the characteristics of productive vocabulary development by low-intermediate EFL learners over multiple drafts by filling a gap. Muncie (2002) reported that learners used more lower-frequency words on their final draft; however, relatively little is known about how productive vocabulary develops over multiple drafts. L2 writing studies generally have shown that the use of lexis improves (Bulté & Housen, 2014; Frear & Bitchener, 2015; Mazgutova & Kormos, 2015; Storch & Tapper, 2009), but this issue needs to be investigated in a variety of different educational settings.

The third purpose involved tracking changes in the participants' ability to produce cohesive texts over one academic semester over multiple drafts. Although few studies have been conducted on changes in the use of cohesive devices, researchers have shown that higher-proficiency writers used more cohesive devices than lower-proficiency writers (Connor, 1990; Witte & Faigley, 1981) with one exception (Crossley & McNamara, 2012). More research is needed, particularly with lower-proficiency L2 writers. The results would fill a gap regarding how lower-proficiency learners produce cohesion.

The fourth purpose is to examine the degree to which the use of lexis and cohesion using computational automated tools correlates with overall writing quality. Few researchers have examined the relationship between lexis, cohesion, and overall

writing quality (Bulté & Housen, 2014; Garner et al., 2020; Storch & Tapper, 2009). The examination of these issues allows practitioners and assessment researchers to assess the development of lexis, cohesion, and L2 writing. By understanding how lexis and cohesion relate to overall writing quality, more detailed assessment rubrics can be made to evaluate learners' performances.

Research Hypotheses

In this section I present the research hypotheses that guide this study. Hypothesis 1 concerns the effect of explicit instruction and feedback on target vocabulary, target cohesion, and overall writing quality through working on multiple essay drafts by Japanese learners of English. Hypotheses 2 and 3 concern improvements in lexis and cohesion, while hypotheses 4 and 5 are focused on the degree to which lexis and cohesion indices correlate with overall writing quality scores.

Hypothesis 1: The explicit instruction group will outperform the comparison group on the use of instructed vocabulary, use of instructed cohesion, and overall writing quality scores.

Although explicit instruction on vocabulary has been investigated (Barcroft, 2002, 2004, 2006; Dimas, 2011; Folse, 2006; Yeung et al., 2020), little attention has been paid to its effect on vocabulary and cohesion development in L2 writing (Lee, 2003; Muncie, 2002; Lee & Munice, 2006). Researchers have suggested that explicit teaching is effective and practical for L2 vocabulary learning (Hinkel, 2002; Nation, 2013; Read, 2000, Schmitt, 2000). Researchers that have examined explicit vocabulary teaching have shown that learners improve their use of vocabulary (Dimas, 2011; Prince, 1996; Yeung,

et al., 2020). Research has generally shown that learners improve writing quality over time (Bulté & Housen, 2014; Storch, 2009). In addition, studies have shown that feedback groups generally perform better than non-feedback groups. Mazgutova and Kormos (2015) reported that feedback contributes to vocabulary development, although they did not investigate it systematically. Hence, it is hypothesized that the explicit instruction group will perform better than the comparison group.

Hypothesis 2: Productive vocabulary will improve over multiple essay drafts and the explicit instruction group will exceed that of the comparison group.

L2 writing studies have shown that the use of lexis improves over time (Bulté & Housen, 2014; Frear & Bitchener, 2015; Mazgutova & Kormos, 2015; Storch & Tapper, 2009). For instance, Muncie (2002) reported that when writing several drafts of a composition, learners used more lower-frequency words on their final draft. Although Storch and Tapper (2009) did not investigate the vocabulary development over multiple essay drafts, they showed that significantly more academic words were in students' essays at the end than at the beginning of an EAP course. Bulté and Housen (2014) also found an increased use of lower-frequency words over four months with upper-intermediate ESL learners. Mazgutova and Kormos (2015) reported that learners used more words from the Academic Word List. In terms of lexical diversity measures, Frear and Bitchener (2015) reported a significant increase in lexical complexity measured by the mean segmental type-token ratio. Mazgutova and Kormos (2015) found that lower-proficiency groups made significant improvements on all measures of lexical diversity. However, Bulté and Housen (2014) found that the Guiraud index decreased slightly and

non-significantly over four months with upper-intermediate ESL learners. Mazgutova and Kormos (2015) also found that the learners' writing did not become more lexically diverse, rich, or sophisticated. In terms of subjective ratings for vocabulary, Bulté and Housen (2014) also found that the subjective rating scales for vocabulary increased significantly from Time 1 to Time 3 and the effect sizes were large. Notwithstanding the knowledge provided by the above studies, this issue needs to be investigated in a wider variety of educational contexts to fully understand the nature of written productive vocabulary learning. Hence, it is hypothesized that explicit instruction can draw learners' attention to vocabulary use, so all the lexical indices will improve over multiple drafts.

Hypothesis 3: The use of cohesion, as measured by local and global cohesion will develop over multiple essay drafts and the explicit instruction group will exceed that of the comparison group.

Researchers have reported mixed findings about the relationship between learners' writing proficiency level and the use of cohesive devices. Crossley et al. (2016) showed that L2 writers improved their use of cohesive devices in local, global, and text cohesion. Local cohesion included connectives and adjacent overlap between sentences, and global cohesion included lexical overlap between two paragraphs, semantic similarity, and synonym overlap between paragraphs. Studies have shown that higher-proficiency writers used more cohesive devices than lower-proficiency writers (Connor, 1990; Witte & Faigley, 1981). Crossley and McNamara (2012) showed the opposite results with Hong Kong Chinese students. It is hypothesized that explicit instruction can

draw learners' attention to cohesive device use so that even low-proficiency learners can improve their use of both local and global cohesion over multiple drafts.

Hypothesis 4: The lexical complexity of the explicit instruction group will be positively related to overall writing quality and will exceed that of the comparison group.

An increasing number of studies with ESL learners have shown that vocabulary is related to overall writing quality (Bulté & Housen, 2014; Crossley et al., 2011; Crossley, 2012; Kim et al., 2018; Verspoor et al., 2012). As for lexical diversity measures, the Guiraud index (a type of type-token ratio) was related to overall writing quality in a one-month study (Bulté & Housen, 2014; Mazgutova & Kormos, 2015; Verspoor et al., 2012), while Crossley and McNamara (2012) showed that lexical diversity was related to L2 writing proficiency. Staples and Reppen (2016) found that learners who used a greater variety of vocabulary in their essays received higher ratings. Hence, it is hypothesized that lexical diversity measures are positively related to overall writing quality. As for the use of low-frequency words, lower-frequency words tend to receive higher scores (Guo et al., 2013; Laufer & Nation, 1995). Crossley et al. (2011) showed that word frequency and lexical diversity were related to L2 proficiency. Guo et al. (2013) also found that higher rated essays included less frequent words. Kim et al. (2018) found that word acquisition measures loaded positively in a principal component analysis, while Kim et al. (2018) found that content word frequency loaded positively in a principal component analysis. Moreover, Garner et al. (2020) showed that bigram and trigram association measures are related to human ratings of L2 writing quality. Kim et al. (2018) found that N-gram frequencies were positively correlated with proficiency and writing quality scores. Hence,

it is hypothesized that all lexical indices, including bigrams and trigrams, are positively related to overall writing quality.

Hypothesis 5: The use of local and global cohesion as measured by cohesive devices and adjacent overlap between paragraphs, is positively related to overall writing quality and the explicit instruction group will exceed that of the comparison group.

A limited number of researchers have examined cohesion and overall writing quality scores on L2 writing development (Crossley & McNamara, 2012; Guo et al., 2013). Crossley and McNamara (2012) found that the higher-proficiency learners used fewer cohesive devices. Other studies showed the opposite results. Guo et al. (2013) found that cohesive devices negatively affected writing quality in the independent task, but they correlated positively with essay quality. Crossley et al. (2016) showed that cohesion was a predictor of human judgments of text organization and holistic ratings. These cohesion measures were two global indices (adjacent overlap between paragraphs for both function words and nouns) and one local index (adjacent overlap between sentences for function words). Crossley et al. (2019) showed that when researchers added word overlap across sentences and paragraphs, the essays were rated higher. Beginner L2 writers are learning how to write coherent essays, so it is hypothesized that the use of both local and global cohesion helps them produce higher quality texts.

CHAPTER 3

METHODS

Context of the Study

The participants were attending a private co-educational university located in a suburban area near Tokyo that offers various majors including humanities, sports, sciences, medicine, economics, law, and engineering. The students were placed in classes using an Internet-based GTEC test (Benesse Corporation), which they completed at home before entering the university.

The participants took four required courses taught by both Japanese and native English-speaking teachers: Reading and Writing, Listening and Speaking, Academic English, and Global Skills. In the first-year Reading and Writing course, the students engage in reading and writing tasks using the *UnLock* series (Westbrook, 2014) published by Cambridge University Press. They also complete one required process writing assignment with multiple drafts. One half of class time is devoted to reading activities using readings in the *UnLock 3* (Westbrook, 2014) textbook. In the second year, the students write a longer essay and read longer passages. The classes meet for 100 minutes twice a week for 14 weeks in language classrooms.

In this course, the participants engaged in reading activities for half of the class and writing activities in the other half. They used the same textbook (See Appendix A for the Japanese course syllabus and Appendix B for the English translation). The writing instruction for academic essays was focused on organization, and as such, the learners

were taught to write an introduction (i.e., a hook, background information, and thesis statement), body paragraphs, and a conclusion.

Participants

The participants were 65 first-year university students with various majors. The explicit instruction group (n = 34) was made up of 20 female and 14 male students, and the comparison group (n = 31) had 14 female and 17 male students. Before entering the university, they had studied English for six years in junior and senior high school. Most of the participants had not written English essays in secondary school, while a few had written single paragraphs.

The participants' general proficiency level was the B1 (Intermediate) in the Common European Framework Reference for Languages (CEFR). This level is equivalent to a TOEFL iBT score of 42–71, a TOEIC (Reading and Listening) total score of 550– 780, and an IELTS score of 4.0 or 5.0.

The participants were fully informed of the purpose of the study, and those who were willing to participate signed a written consent form (See Appendix C for the Japanese version of the consent form and Appendix D for the English version). Interviewees signed a written consent form (See Appendix E for the Japanese version of the consent form and Appendix F for the English version).

Instrumentation

Four instruments were used in the study: a background questionnaire, writing tasks, writing rubrics, and interview protocols.

Background Questionnaire

The participants completed a background questionnaire concerning English writing experiences and language proficiency tests (See Appendix G for the Japanese version and Appendix H for the English version). They completed the Japanese version of the questionnaire in order to eliminate any language-related misunderstandings.

Writing Tasks

Two writing tasks were administered to assess vocabulary use, cohesion, and writing quality: a comparison and contrast essay and a problem-solution essay. The learners were instructed not to use dictionaries when completing these tasks. They wrote three drafts for each essay outside of class. The topic for Task 1, the comparison and contrast essay, was to compare and contrast living in a big city and living in a small town. The topic for Task 2, the problem-solution essay, concerned transportation in Tokyo. The writing samples were evaluated by six raters using five criteria: content, organization, vocabulary, cohesion, and grammar. Rater training was conducted before assessing essays, and discrepancies were discussed. Rater 1 rated 1–65, Rater 2 rated essays 1–20, Rater 3 rated essays 10-30, Rater 4 rated essays 20–40, Rater 5 rated essays 30–50, and Rater 6 rated essays 40–65. In this way, overlap was maintained between the different raters. This overlap is necessary when analyzing the data using many-facet Rasch measurement with the FACETS software.

Writing Rubrics

The analytic writing rubric used in this study was made by a university curriculum committee to assess writing quality (See Appendix I). This rubric was adopted for use in this study because it was designed to evaluate this particular group of

participants. As noted above, the rubric covers five aspects of writing—content, organization, vocabulary, cohesion, and grammar—two of which, vocabulary and cohesion, were the main variables investigated in this study. Each criterion was scored on a six-point scale. FACETS 3.80.0 (Linacre, 2017) was used to produce overall writing quality measures. Six raters—four Japanese female raters and two female native-English speakers—trained in Applied Linguistics and experienced in teaching writing rated the essays.

Interview Protocols

Interviews were conducted to complement the quantitative data and to better understand the overall picture of the learning processes and outcomes on vocabulary, cohesion, and general writing quality and to answer the research hypotheses. Six participants were interviewed using the interview protocol (See Appendix J for the Japanese version and Appendix K for the English version of the protocol). The interview, which was focused on the interviewees' writing experiences, was conducted by me in Japanese for 15 minutes in Week 14. The interviews were recorded with an Olympus Voice-Trek DS-51 IC recorder.

Research Design

The present study addresses the effectiveness of explicit instruction and feedback on vocabulary and cohesion in L2 writing. A mixed-methods design was used by including quantitative and qualitative data to allow for a more complete understanding of the quantitative results. The quantitative analysis of learners' writing performance provides a limited understanding of how participants perform on writing tasks and

perceive the pedagogical interventions. The qualitative results complement the quantitative results. After collecting quantitative data, I collected the qualitative data.

The participants completed two tasks with three drafts. Quantitative data involved using these essays that the participants wrote over the semester. Overall writing quality scores were assessed by the six raters. Lexical and cohesion scores were assessed by TALLES (Kyle & Crossley, 2015), TAALED 1.41 (Zenker & Kyle, 2021), and TAACO 2.0 (Crossley et al., 2016). These scores were overall writing quality, MATTR, COCAFrq, Trigrams, All connectives, and AdjacentP. The independent variable was Group (the explicit instruction and the comparison groups). Qualitative interviews were conducted at the end of the study to help explain the quantitative results and to provide a more complete understanding of the participants' writing experiences and their use of vocabulary and cohesion use in L2 writing (See Appendices J and K).

Procedures

Writing Task 1: Comparison-Contrast Essay

Both groups wrote each draft on the same schedule. Table 1 shows the research procedures. In Week 3, before the participants worked on the first draft for Task 1, pre-writing skills were taught to both groups including brainstorming, process writing, essay structure, how to write introduction, body, and conclusion paragraphs, and formatting (See Appendix L).

In Week 4, both groups had one week to write the first draft using a minimum of 150 words. In this task they compared and contrasted living in a big city and living in a small town. The explicit instruction group wrote the first draft with no intervention concerning vocabulary and cohesive devices.

Table 1. *Research Procedures*

Week	Explicit instruction group	Comparison group
1	——	——
2	——	——
3	Pre-writing skills (essay structure)	Pre-writing skills (essay structure)
4	Task 1, Draft 1: Write the first draft	Task 1, Draft 1: Write the first draft
5	Task 1, Draft 2: Read the model essay, complete activities, and write the second draft	Task 1, Draft 2: Write the second draft using a checklist
6	——	——
7	Task 1, Draft 3: Receive feedback on writing quality, vocabulary and cohesive devices, underline the target vocabulary and cohesive devices, and write the final draft	Task 1, Draft 3: Receive feedback on content and formatting, and write the final draft using a checklist
8	——	——
9	Task 2, Draft 1: Write the first draft	Task 2, Draft 1: Write the first draft
10	Task 2, Draft 2: Read the model essay, complete activities, and write the second draft	Task 2, Draft 2: Write the second draft using a checklist
11	——	——
12	Task 2, Draft 3: Receive feedback on writing quality, vocabulary and cohesive devices, underline the target vocabulary and cohesive devices, and write the final draft	Task 2, Draft 3: Receive feedback on content and formatting, and write the final draft using a checklist
13	——	——
14	——	——

In Week 5, different instructions were provided. For the explicit instruction group, explicit teaching was provided in the form of a model essay containing target vocabulary and cohesive devices (See Appendix M) and exercise materials. Two activities concerning vocabulary and cohesive devices using the model essay were given to the explicit instruction group; thus, they were exposed to the target vocabulary and cohesive

devices twice. I taught how to use the target vocabulary and cohesive devices using the model essay. Twelve target words were chosen because I thought that the participants would be able to use them in this essay. The target words were *located, opposite, activities, famous, climate, popular, pace, crowded, comfortable, destination, developed,* and *atmosphere*. Eight target cohesive devices were taught: *the first point*, *in addition*, *on the other hand*, *the second point*, *but*, *however*, *so*, and *in conclusion*. In the two activities using the model essay, the students underlined the target words and cohesive devices in the model essay (See Appendix M). They also completed a fill-in-the-blank vocabulary and cohesive device exercise for the model essay with the first letter of the Japanese translations provided. After completing these exercises, they underlined these target words and cohesive devices that they used from the model essays in their first draft for Task 1. This approach allowed the participants to clearly know which target words and phrases they had and had not used. The comparison group was not provided any instruction. Instead, they were given a checklist to use to revise their essays' organization, content, grammar, vocabulary, and cohesive devices (See Appendix N). Both groups were instructed to write a second draft using a minimum of 200 words by Week 6. The explicit instruction group was told to increase their use of the target words and target cohesive devices, particularly if they did not use them or used them sparingly in their first draft.

Feedback was provided to the two groups in Week 7. The explicit instruction group received feedback concerning their use of the target words and cohesive devices, content, organization, grammar, and formatting. Instead of intensive feedback, general feedback was provided. The comparison group also received feedback, but on content

and format only. Both groups wrote the final draft based on feedback received from me. The participants had one and half weeks to write the final draft at home using at least 250 words.

Writing Task 2: Problem-Solution Essay

Task 2, the problem-solution essay, proceeded in the same manner as Task 1, the comparison-contrast essay. In Week 9, before the participants worked on the first draft, the essay structure for the problem-solution essay was taught to both groups. I explained that a problem-solution essay consisted of an introduction, a problem paragraph, a solution paragraph, and a conclusion. The participants were told to explain a problem in detail in the second draft and solutions to the problem in the third draft. The students wrote the first draft of the problem-solution essay on the topic of transportation. They had one week to write at least 150 words outside of class.

In Week 10, different instructions were provided to the two groups as in Task 1. For the explicit instruction group, explicit teaching was provided with a model essay that included target vocabulary and cohesive devices and exercise materials (See Appendix L). In total, two activities about vocabulary and cohesive devices used in the model essay on transportation were given to the explicit instruction group; thus, they were exposed to the target words and cohesive devices twice. I taught them how to use the target vocabulary and cohesive devices using the model essay. Thirteen target words were chosen because they were potentially easy to use in the essay. The target words were *majority, urban, restricted, solution, proving, fee, public, regular, congestion, decreased, council, excessive,* and *alleviated*. The five target cohesive devices were as follows: *one of the most serious issues, the main reason, so, to conclude,* and *transportation*. In the

two activities using the model essay, the students underlined the target vocabulary and target cohesive devices (See Appendix M), received explanations and Japanese translations about them, and completed fill-in-the-blank exercises with the first letter of the Japanese translations provided. In addition, the participants underlined the target vocabulary and cohesive devices for each draft they wrote because this allowed them and me to clearly know which target words and phrases had been used and not used. In Week 12, both groups wrote the second draft outside of class using at least 200 words. The explicit instruction group was told to use more of the target words and target cohesive devices in this draft, particularly if they did not use them or used infrequently in their first draft. The comparison group did not receive instruction concerning the target vocabulary and cohesive devices. Instead, they wrote their second draft using a checklist (See Appendix N).

Feedback was provided to the two groups in Week 12. The explicit instruction group received feedback concerning the use of target words and cohesive devices in their essays, content, organization, grammar, and formatting. Instead of intensive feedback, general feedback was provided. The comparison group also received feedback, but only on content and format. Both groups had one and a half weeks to write a final draft of at least 250 words outside of class. Both groups wrote the final draft based on feedback received from me. Finally, six participants were interviewed about their experiences writing the essays and the treatments they received. Table 1 shows the procedures followed in this study.

Analyses

Before conducting the statistical analyses, the quantitative data were screened to identify outliers and check for the normality of each dependent variable. In order to answer the five hypotheses, the following analyses were conducted.

Hypothesis 1 stated that the explicit instruction group would perform better than the comparison group on target vocabulary, target cohesion, and writing quality scores. This hypothesis was addressed using one mixed-design MANOVA. The independent variable was group, which included two levels: the explicit instruction group and the comparison group. The dependent variable was Overall Writing Quality for the first, second, and third drafts for the two essays. Two tasks were analyzed together. Target word use and Target cohesion use were analyzed based on descriptive statistics.

Hypothesis 2 stated that lexical indices—as measured by MATTR, COCAFrq, Trigrams—would improve from the first to the final drafts. It was also hypothesized that the explicit instruction group would score higher than the comparison group. Two mixed-design MANOVAs with one repeated-measures variable for each task were used to assess the hypothesis. The independent variable was group: the explicit instruction group and the comparison group. The dependent variables were the first, second, and third drafts analyzed using MATTR, COCAFrq, and Trigrams for each essay.

Hypothesis 3 stated that the use of cohesion, as measured by local (All connectives) and global (AdjacentP) cohesion indices, would develop over time and the explicit instruction group would exceed that of the comparison group. It was also hypothesized that the explicit instruction group would improve more than the comparison group. Two mixed-design MANOVAs were conducted. The independent variable was

group, which had two levels, the explicit instruction group and the comparison group. One repeated-measures variable was Draft. The dependent variables were the first, second, and third drafts of the local (All connectives) and global cohesion (AdjacentP) measures. Each task was analyzed separately.

Hypothesis 4 stated that lexical complexity would be positively related to overall writing quality and the explicit instruction group will exceed that of the comparison group. A Pearson product moment correlation was conducted to estimate the degree to which the lexical indices were correlated with overall writing quality. The two tasks were analyzed separately. Overall writing quality were produced using FACETS 3.80.0 (Linacre, 2017). I rated all of the essays, Rater 2 assessed essays written by participants 1–20, Rater 3 assessed essays written by participants 10–30, Rater 4 assessed essays written by participants 20–40, Rater 5 assessed essays written by participants 30–50, and Rater 6 assessed essays written by participants 40–65.

Hypothesis 5 stated that cohesion will be positively related to overall writing quality. A Pearson product moment correlation was conducted to estimate the degree to which the cohesion indices were correlated with overall writing quality and the explicit instruction group will exceed that of the comparison group. The two tasks were analyzed separately.

Finally, interviews were conducted to complement the quantitative data. Based on the participants' writing performance, six participants were selected for interviews about their learning experiences to complement the results of the quantitative analyses. In each group, one participant showing high performance, one displaying a mid-level performance, and one producing a low performance on the essays were selected.

CHAPTER 4

PRELIMINARY ANALYSES

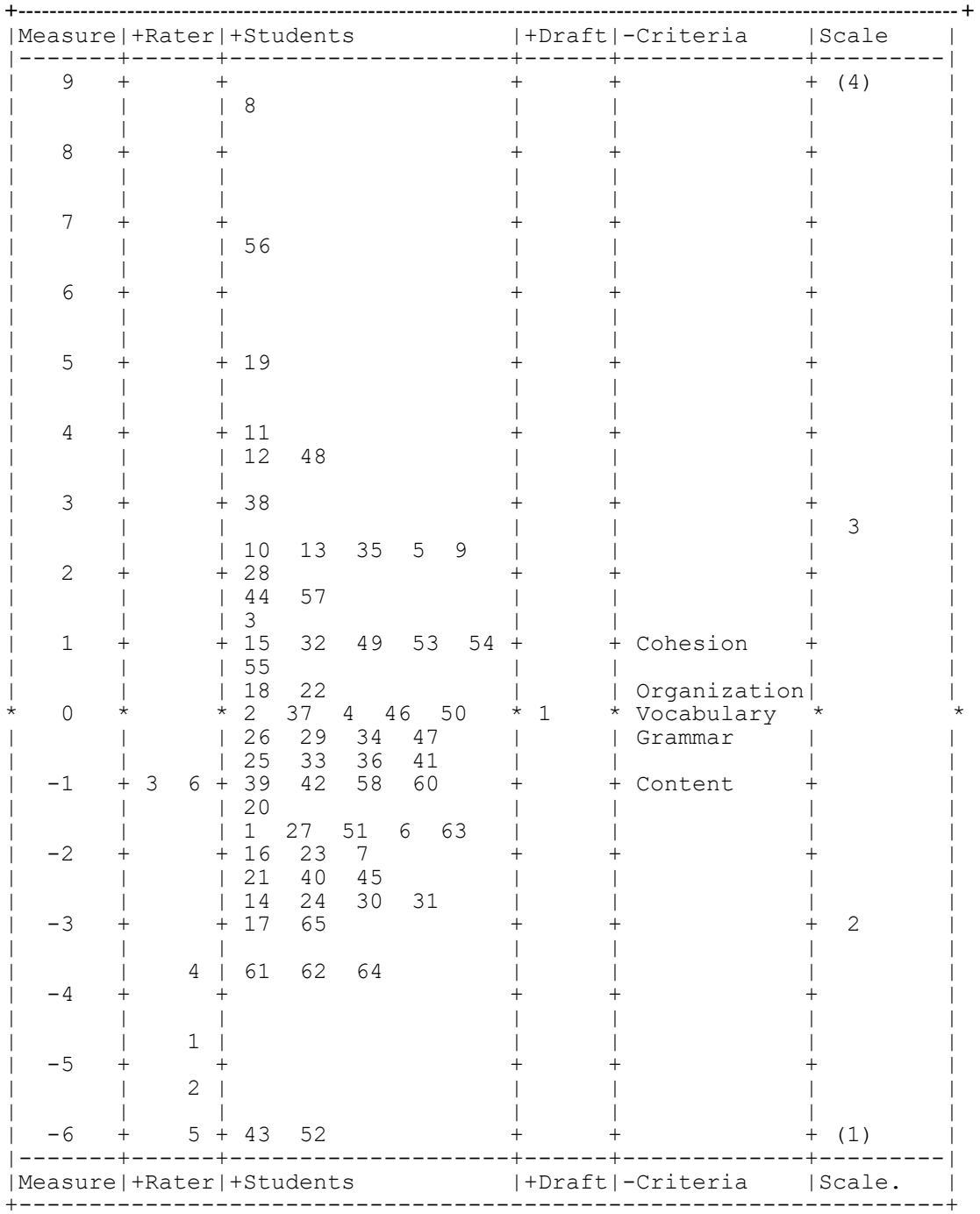
Many-facet Rasch measurement (MFRM) was used to analyze the two writing tasks. The Rasch measurement model was employed because it provides a way to construct interval item and person measures, calculates item and person fit indices for each facet in the analysis, and provides item and person reliability estimates (Bond & Fox, 2007; McLean et al., 2015). The ratings provided by six raters were first entered into an Excel sheet, and exported to and calibrated using the FACETS software 3.80.0 (Linacre, 2017).

Technical quality was evaluated by inspecting Rasch standardized item weighted mean square fit statistics as estimated with the 65 participants. The infit and outfit MNSQ criterion adopted in this study for criteria was 0.50–1.50, while infit and outfit MNSQ statistics for persons were considered acceptable if they were < 2.00. In addition, a standardized infit value of > + 2.00 was adopted as the criterion for determining underfitting criteria to the Rasch model. Technical quality was also assessed through identifying overfitting criteria (Bond & Fox, 2007). Overfit to the Rasch model was defined as items showing standardized infit and outfit values > -2.00. Such items could indicate that the assumption of local independence has been violated.

FACETS Analyses for Task 1 Draft 1

Figure 1 shows the Wright map for Task 1 Draft 1 (See Appendix O). The four facets—participants, raters, drafts, and criteria—are placed on the same logit scale, which is shown in the Measure column. The Rater column shows the distribution of the

Figure 1. *Wright Map for Task 1 Draft 1*



six raters, with more severe raters toward the top of the figure and more lenient raters toward the bottom. The writing ability estimates for the 65 participants are shown in the Students column. More able participants are located toward the top of the Figure and less

able participants toward the bottom. The Draft column shows the difficulty of the draft, which was anchored to 0.00 logits, because of the generic element, so this would exclude it from estimation, but include it for fit statistics. The Criteria column shows the five rating criteria with cohesion being the most difficult and content being relatively easy. This was their first time to write a comparison and contrast essay, so most of the scores were centered around 2, 3, and 4, with no participants receiving a 5 or a 6.

Table 2 shows the Rasch statistics for the six raters for Task 1 Draft 1. The infit MNSQ and outfit MNSQ statistics were inspected and only Rater 4 did not meet the 0.50-1.50 MNSQ criterion. The most severe rater was Rater 5, and the most lenient rater was Rater 3. The point measure correlations for the raters ranged between .57 and .84. The Rasch separation statistic for the raters was 9.58 and the Rasch reliability estimate was .99, which was excellent.

Table 2. *Rater Measurement Report for Task 1 Draft 1*

Rater	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt. measure correlation
3	2.28	-0.98	0.24	0.67	-2.20	1.23	1.30	0.74
6	2.26	-1.07	-0.21	1.17	1.30	1.23	1.03	0.76
4	1.91	-3.55	0.22	1.74	4.80	1.99	5.00	0.57
1	1.74	-4.57	0.13	0.87	-1.70	0.81	-1.90	0.75
2	1.52	-5.50	0.25	0.64	-2.50	0.61	-1.80	0.84
5	1.44	-5.84	0.25	0.92	-0.50	0.90	-0.30	0.58

Note. RMSE = 0.22; *SD* = 2.13; Separation = 9.58; Strata = 13.10; Reliability = .99; Chi-square = 450.3; *df* = 5; $p < .01$.

The Rasch descriptive statistics for the participants for Task 1 Draft 1 are displayed in Table 3. The infit MNSQ statistics ranged from 0.08 to 2.42, and the outfit MNSQ statistics ranged from 0.28 to 3.00. Only student 22 misfit the Rasch model. The

point measure correlation ranged from .57 to .84. The Rasch separation statistic was 5.05 and the Rasch reliability estimate was .93.

Table 3. *Student Measurement Report for Task 1 Draft 1*

Student	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
8	3.34	8.60	0.85	0.76	-0.10	0.80	0.00	.00
56	3.03	6.68	0.74	0.46	-1.50	0.36	-1.30	.63
19	2.84	5.03	0.57	0.73	-0.90	0.64	-1.00	.73
11	2.62	3.98	0.62	0.98	0.00	1.00	0.10	.58
48	2.55	3.71	0.65	1.58	1.20	1.91	1.40	.77
12	2.52	3.58	0.65	0.75	-0.40	0.62	-0.60	.72
38	2.39	3.05	0.61	1.27	0.70	1.26	0.60	.55
35	2.21	2.25	0.65	0.87	-0.10	0.89	0.00	.54
10	2.21	2.23	0.67	0.28	-1.80	0.20	-1.90	.86
13	2.21	2.23	0.67	1.03	0.20	1.07	0.30	.54
5	2.20	2.19	0.86	0.08	-2.10	0.07	-2.10	.00
9	2.20	2.19	0.86	0.08	-2.10	0.07	-2.10	.00
28	2.18	2.06	0.63	1.88	1.70	2.04	1.70	.32
44	2.10	1.56	0.61	0.77	-0.50	0.70	-0.50	.67
57	2.09	1.52	0.75	1.60	1.20	1.50	0.90	.12
3	2.09	1.48	0.82	0.82	-0.10	0.99	0.20	-.29
15	2.03	1.05	0.59	0.66	-1.00	0.57	-1.00	.36
53	2.03	0.98	0.72	0.89	-0.10	0.76	-0.30	.76
54	2.03	0.98	0.72	1.43	1.00	1.20	0.50	.45
55	2.03	0.98	0.72	0.86	-0.20	0.72	-0.40	.42
32	2.03	0.97	0.64	1.00	0.10	1.06	0.20	.61
49	2.01	0.87	0.57	1.04	0.20	0.98	0.00	.71
22	1.97	0.49	0.61	2.42	2.70	3.00	2.90	.26
18	1.96	0.39	0.56	0.64	-1.40	0.57	-1.40	.75
59	1.91	-0.02	0.70	0.45	-1.70	0.40	-1.60	.00
46	1.90	-0.08	0.56	1.42	1.40	1.54	1.50	.38
50	1.90	-0.08	0.56	0.80	-0.60	0.74	-0.70	.80
2	1.89	-0.12	0.67	0.98	0.00	0.95	0.00	.33
4	1.89	-0.12	0.67	1.05	0.20	1.05	0.20	.23
37	1.89	-0.14	0.58	1.55	1.60	1.84	1.80	.78
26	1.88	-0.24	0.59	0.56	-1.30	0.48	-1.30	.22
29	1.88	-0.24	0.59	1.59	1.50	1.82	1.60	-.14
47	1.85	-0.40	0.57	1.60	1.80	1.83	2.00	.56
34	1.84	-0.47	0.58	0.95	0.00	0.86	-0.20	.47
25	1.82	-0.58	0.59	1.13	0.40	1.26	0.70	.67
41	1.79	-0.74	0.59	1.02	0.10	1.07	0.20	.67
33	1.78	-0.80	0.57	1.16	0.60	1.32	0.90	.71
36	1.78	-0.80	0.57	1.42	1.40	1.81	2.00	.51
58	1.74	-1.03	0.73	1.43	1.00	1.53	1.00	-.02
60	1.74	-1.03	0.73	0.62	-0.90	0.50	-0.90	.55
42	1.72	-1.09	0.61	0.92	-0.10	0.92	0.00	.89

Table 3 (continues)

Table 3 (continued)

Student	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
39	1.71	-1.13	0.58	0.84	-0.50	0.79	-0.50	.57
20	1.67	-1.34	0.52	1.33	1.10	1.38	1.00	.71
1	1.62	-1.52	0.73	1.07	0.30	0.96	0.00	.27
6	1.62	-1.52	0.73	0.54	-1.00	0.55	-1.00	.75
51	1.61	-1.58	0.76	0.63	-0.70	0.52	-0.70	.68
63	1.61	-1.58	0.76	0.45	-1.30	0.34	-1.30	.76
27	1.60	-1.61	0.59	1.12	0.40	1.02	0.10	.76
23	1.51	-1.97	0.60	1.12	0.40	0.94	0.00	.47
16	1.49	-2.04	0.68	0.47	-1.30	0.34	-1.40	.85
7	1.47	-2.11	0.83	1.27	0.60	2.25	1.50	-.20
21	1.42	-2.34	0.62	0.92	-0.10	0.78	-0.30	.60
45	1.39	-2.43	0.74	1.14	0.40	0.93	0.10	.82
40	1.39	-2.45	0.58	0.91	-0.10	0.87	0.00	.78
14	1.37	-2.54	0.73	0.51	-0.90	0.38	-0.90	.85
31	1.34	-2.66	0.70	1.12	0.40	1.51	0.80	.21
24	1.33	-2.73	0.64	0.79	-0.40	0.70	-0.50	.69
30	1.30	-2.83	0.61	0.74	-0.60	0.56	-0.50	.74
65	1.30	-2.86	0.83	1.63	1.10	1.32	0.60	.46
17	1.25	-3.11	0.78	0.21	-2.00	0.14	-1.30	.93
61	1.17	-3.57	0.84	0.42	-1.30	0.28	-0.80	.86
62	1.17	-3.57	0.84	0.42	-1.30	0.28	-0.80	.86
64	1.17	-3.57	0.84	0.42	-1.30	0.28	-0.80	.86
52	1.00	(-7.48)	(1.90)	Minimum				.00
43	1.00	(-7.49)	(1.90)	Minimum				.00

Note. RMSE = 0.75; *SD* = 2.64; Separation = 3.54; Strata = 5.05; Reliability = .93; Chi-square = 774.2; *df* = 64; $p < .01$.

The Rasch statistics for Draft 1 (Task 1) are shown in Table 4. The infit MNSQ and the outfit MNSQ statistics were below the 2.00 criterion, at 0.99 and 0.98, respectively. The draft was anchored at 0.00 logits.

Table 4. *Draft Measurement Report for Task 1 Draft 1*

Draft	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt. measure correlation
1	1.91	0.00	0.08	0.99	-0.10	0.98	-0.30	.78

The Rasch statistics for the five rating criteria for Task 1 Draft 1 are displayed in Table 5. The infit MNSQ statistics ranged from 0.85 to 1.10 and the outfit MNSQ statistics ranged from 0.77 to 1.13. Both were within the 0.50–1.50 criterion. The point

measure correlations ranged from .71 to .81, the Rasch separation statistic was 3.70, and the Rasch reliability estimate was .93. Cohesion was the most difficult criterion to score highly on and content was the easiest.

Table 5. *Criteria Measurement Report for Task 1 Draft 1*

Criterion	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt. measure correlation
Cohesion	2.03	1.01	0.18	1.02	0.20	1.03	0.20	.78
Org	1.94	0.22	0.18	1.09	0.80	1.13	0.90	.71
Vocab	1.92	0.09	0.18	0.85	-1.40	0.77	-1.70	.81
Grammar	1.84	-0.46	0.19	0.91	-0.80	0.86	-0.80	.81
Content	1.77	-0.85	0.19	1.10	0.80	1.09	0.50	.73

Note. Sample RMSE = .18; Adjusted *SD* = .68; Separation = 3.70; Strata = 5.27; Reliability = .93; Chi-square = 59.00; *df* = 4; $p < .01$. Org = Organization; Vocab = Vocabulary.

The Rasch rating scale statistics for Task 1 Draft 1 are shown in Table 6. Each point on the rating scale fit the Rasch model well and the different points were clearly separated from one another, as shown in the two threshold columns. The thresholds were separated between the 1.40–5.00 criterion.

Table 6. *Category Statistics for Task 1 Draft 1*

Score	Count (%)	Average measure	Outfit MNSQ	Rasch-Andrich threshold (*SE*)	Rasch-Thurstone threshold
1	267 (32)	-6.43	0.90	—	Low
2	463 (55)	-3.40	0.90	-5.58 (0.10)	-5.58
3	109 (13)	0.47	1.10	-0.09 (0.15)	-0.09
4	6 (1)	4.60	1.10	5.67 (0.49)	5.66

FACETS Analyses for Task 1 Draft 2

Figure 2 shows the Wright map for Task 1 Draft 2.

Figure 2. *Wright Map for Task 1 Draft 2*

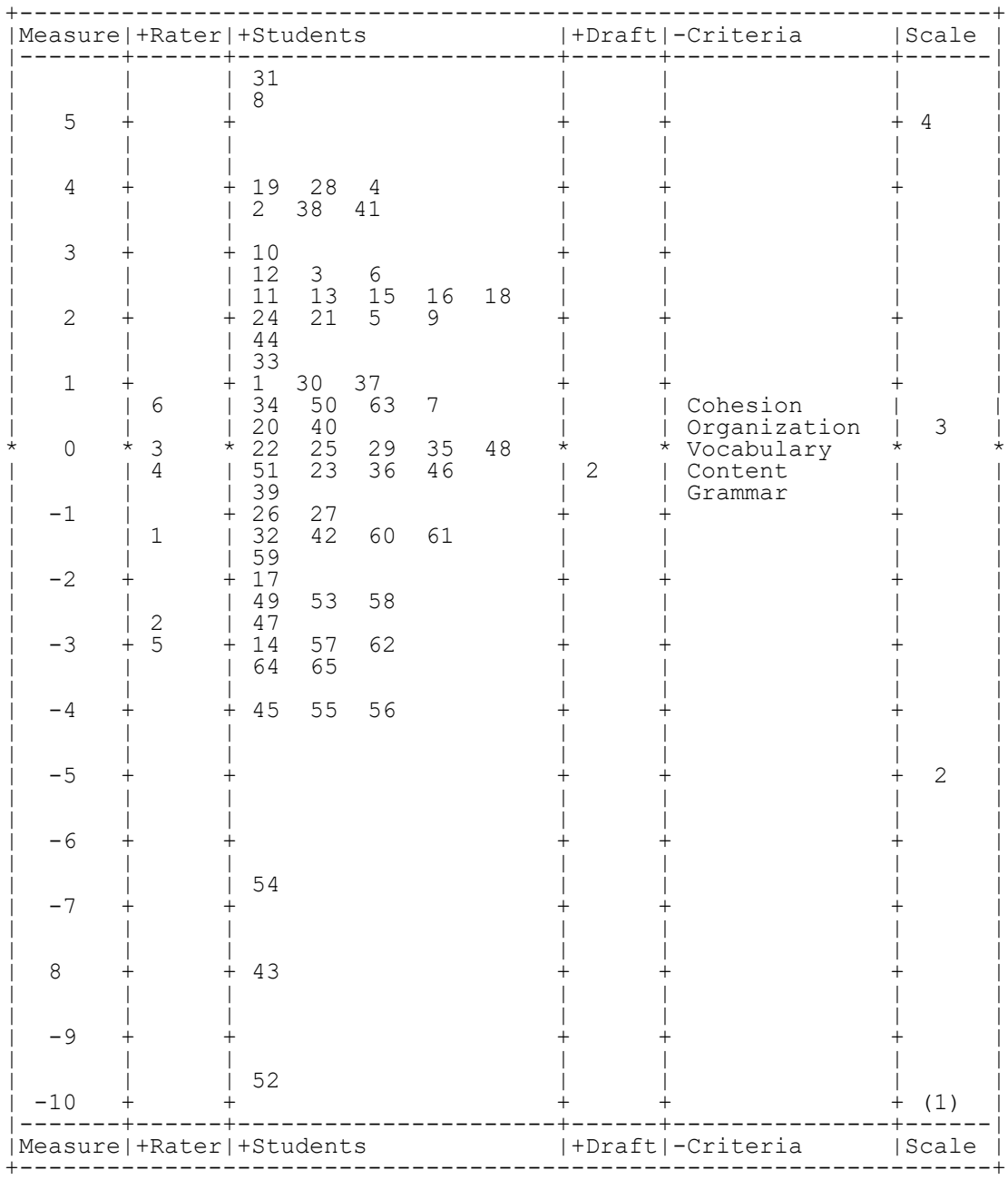


Table 7 shows the Rasch statistics for the six raters 1 for Task 1 Draft 2. In this case, the raters mostly showed an acceptable fit to the Rasch model, while rating in a more severe manner compared with the first draft. The infit MNSQ ranged from 0.55 to 1.65 and the outfit MNSQ statistics ranged from .55 to 1.16. Rater 4 underfit the model. The point measure correlations ranged from .52 to .81, the Rasch separation statistic was 7.46, and the Rasch reliability estimate was .98, which was excellent.

Table 7. *Rater Measurement Report for Task 1 Draft 2*

Rater	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt. measure correlation
6	3.08	0.73	0.19	1.10	0.90	1.16	1.20	.75
3	2.96	0.04	0.21	0.85	-1.10	0.86	-0.90	.62
4	2.92	-0.20	0.22	1.65	3.40	1.71	3.40	.52
1	2.69	-1.31	0.13	0.99	0.00	0.94	-0.60	.80
2	2.37	-2.54	0.23	0.55	-3.10	0.55	-2.70	.67
5	2.26	-3.03	0.23	0.73	-2.10	0.69	-1.80	.81

Note. RMSE = 0.20; *SD* = 1.51; Separation = 7.46; Strata = 10.28; Reliability = .98; Chi-square = 275.00; *df* = 5; $p < .01$.

The Rasch descriptive statistics for the participants for Task 1 Draft 2 are shown in Table 8. The infit MNSQ statistics ranged from 0.32 to 2.80 and the outfit MNSQ statistics ranged from 0.06 to 3.15. Participants 21, 29, 37 and 46 underfit the Rasch model with Infit MNSQ statistics above the 2.00 criterion. The Rasch separation statistic was 4.64 and the Rasch reliability estimate was .96.

Table 8. *Student Measurement Report for Task 1 Draft 2*

Student	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
31	3.95	5.82	0.60	1.23	0.60	1.41	0.90	.25
8	3.88	5.30	0.67	0.72	-0.70	0.67	-0.70	.57
4	3.62	4.06	0.64	0.56	-1.50	0.54	01.50	.83

Table 8 (continues)

Table 8 (continued)

Student	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
31	3.95	5.82	0.60	1.23	0.60	1.41	0.90	.25
8	3.88	5.30	0.67	0.72	-0.70	0.67	-0.70	.57
4	3.62	4.06	0.64	0.56	-1.50	0.54	01.50	.83
28	3.61	4.02	0.58	0.97	0.00	1.04	0.20	.20
19	3.59	3.94	0.55	0.88	-0.20	0.88	-0.10	.47
2	3.52	3.65	0.65	1.13	0.40	1.24	0.60	.79
41	3.50	3.60	0.58	0.86	-0.20	0.95	0.00	.79
38	3.49	3.55	0.55	1.50	1.30	1.53	1.30	.01
10	3.35	3.02	0.56	0.61	-1.10	0.54	-1.20	.65
3	3.27	2.71	0.72	0.45	-1.00	0.42	-1.00	.23
6	3.27	2.71	0.72	0.37	-1.20	0.32	-1.30	.43
12	3.27	2.71	0.56	1.09	0.30	1.04	0.20	.17
11	3.20	2.39	0.56	0.54	-1.20	0.49	-1.30	.54
13	3.20	2.39	0.56	1.06	0.20	1.00	0.10	.06
15	3.20	2.39	0.56	0.32	-2.20	0.27	-2.30	.74
16	3.20	2.39	0.56	0.98	0.00	0.98	0.10	.86
18	3.20	2.39	0.56	0.32	-2.20	0.27	-2.30	.74
24	3.17	2.25	0.53	0.90	-0.20	0.91	-0.10	.26
5	3.16	2.18	0.74	0.06	-2.70	0.06	-2.70	.00
9	3.16	2.18	0.74	0.06	-2.70	0.06	-2.70	.00
21	3.12	1.96	0.54	2.80	3.80	3.05	3.70	-.02
44	3.06	1.64	0.55	0.68	-0.90	0.61	-1.00	.54
33	3.01	1.39	0.56	1.61	1.50	1.76	1.60	.53
1	2.97	1.15	0.69	0.52	-1.00	0.47	-1.10	.62
37	2.96	1.08	0.56	2.01	2.20	2.19	2.30	.43
30	2.92	0.84	0.48	0.94	0.00	0.97	0.00	.56
50	2.91	0.77	0.54	0.99	0.00	1.00	0.10	.41
34	2.91	0.77	0.55	1.46	1.20	1.51	1.20	.81
7	2.89	0.70	0.66	0.77	-0.50	0.76	-0.50	.39
63	2.86	0.54	0.67	0.74	-0.50	0.80	-0.30	.42
20	2.80	0.22	0.49	1.41	1.20	1.48	1.20	.11
40	2.79	0.21	0.48	1.11	0.40	1.15	0.50	.75
35	2.78	0.16	0.56	0.70	-0.80	0.74	-0.60	.83
51	2.77	0.09	0.67	1.59	1.30	1.66	1.30	.01
22	2.75	0.03	0.58	1.42	1.00	1.48	1.00	.01
25	2.75	0.03	0.58	0.32	-2.00	0.26	-2.10	.65
29	2.75	0.03	0.58	2.32	2.40	2.20	2.00	-.25
48	2.72	-0.11	0.55	1.62	1.60	1.49	1.20	.28
23	2.68	-0.30	0.57	1.16	0.50	1.16	0.40	.44
46	2.65	-0.42	0.57	2.83	3.60	3.15	3.60	-.21
36	2.63	-0.47	0.56	0.49	-1.70	0.43	-1.80	.80
39	2.55	-0.79	0.57	0.49	-1.70	0.42	-1.70	.81
26	2.52	-0.92	0.55	1.02	0.10	1.08	0.30	.70
27	2.52	-0.92	0.55	0.41	-2.20	0.38	-2.20	.87
60	2.42	-1.30	0.70	1.11	0.30	1.18	0.50	.33
61	2.42	-1.30	0.70	0.51	-1.10	0.47	-1.20	.76
32	2.38	-1.47	0.60	1.22	0.60	0.99	0.10	.31
42	2.37	-1.80	0.71	0.61	-0.80	0.56	-0.90	.73

Table 8 (continues)

Table 8 (continued)

Student	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
59	2.30	-1.80	0.71	0.61	-0.80	0.56	-0.90	.73
17	2.25	-2.05	0.62	1.67	1.50	1.78	1.30	.30
49	2.20	-2.30	0.65	0.91	0.00	0.67	-0.40	.56
53	2.20	-2.31	0.73	0.54	-1.10	0.44	-1.10	.78
58	2.20	-2.31	0.73	0.36	-1.80	0.30	-1/60	.89
47	2.13	-2.73	0.66	0.69	-0.60	0.63	-0.40	.87
57	2.12	-2.85	0.75	0.82	-0.20	0.60	-0.50	.56
62	2.12	-2.85	0.75	0.81	-0.30	0.59	-0.50	.57
14	2.11	-0.29	0.69	0.52	-1.00	0.32	-1.20	.57
64	2.05	-3.43	0.78	0.53	-1.00	0.35	-0.90	.63
65	2.05	-3.43	0.78	0.64	-0.70	0.43	-0.70	.56
45	2.00	-4.04	0.66	1.28	0.70	1.49	0.90	.51
55	2.00	-4.08	0.83	0.50	-0.80	0.35	-0.90	.40
56	2.00	-4.08	0.83	0.50	-0.80	0.35	-0.90	.40
54	1.59	-6.79	0.76	1.21	0.60	1.53	0.90	.22
43	1.28	-8.10	0.69	0.74	-0.50	0.53	-0.60	.74
52	1.07	-9.75	0.87	1.60	1.30	2.04	1.10	-.05

Note. Sample RMSE = 0.64; *SD* = 2.95; Separation = 4.64; Strata = 6.52; Reliability = .96; Chi-square = 1242.20; *df* = 64; $p < .01$.

The Rasch statistics for Draft 2 (Task 1) are shown in Table 9. The infit MNSQ statistic outfit MNSQ statistics were excellent at 0.99 and 0.98, respectively.

Table 9. *Draft Measurement Report for Task 1 Draft 2*

Draft	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt. measure correlation
2	2.78	0.00	0.08	0.99	-0.10	0.98	-0.30	.78

The Rasch statistics for the five rating criteria for Task 1 Draft 2 are displayed inTable 10. The criteria fit the Rasch model well. Grammar was the most difficult criterion to score highly on, while organization was the easiest. The infit MNSQ ranged from 0.86 to 1.13 and the outfit MNSQ statistics ranged from 0.81 to 1.10. The Rasch separation statistic was 3.72 and the Rasch reliability estimate was good at .87.

Table 10. *Criteria Measurement Report for Task 1 Draft 2*

Criterion	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt. measure correlation
Cohesion	2.87	0.45	0.17	1.05	0.40	1.06	0.40	.76
Org	2.85	0.34	0.17	1.01	0.10	1.02	0.20	.79
Vocab	2.75	0.00	0.17	0.93	-0.60	0.91	-0.60	.79
Content	2.76	-0.06	0.17	0.86	-1.30	0.81	-1.40	.81
Grammar	2.60	-0.73	0.17	1.13	1.10	1.10	0.70	.76

Note. RMSE = .17; *SD* = .43; Separation = 3.72; Strata = 3.72; Reliability = .87; Chi-square = 29.50; *df* = 4; $p < .01$. Org = Organization; Vocab = Vocabulary.

The Rasch rating scale statistics for Task 1 Draft 2 are displayed in Table 11. The distance between the thresholds was outside of the 1.40 to 5.00 range from point 2 to point 3 on the scale.

Table 11. *Category Statistics for Task 1 Draft 2*

Score	Count (%)	Average measure	Outfit MNSQ	Rasch-Andrich threshold (*SE*)	Rasch-Thurstone threshold
1	26 (3)	-8.59	1.30	—	Low
2	265 (30)	-3.24	1.00	-8.19 (0.33)	-8.19
3	453 (52)	0.02	0.90	-2.01 (0.10)	-2.02
4	124 (14)	2.39	1.00	2.52 (0.13)	2.52
5	2 (0)	5.03	1.01	7.68 (0.73)	7.67

FACETS Analyses for Task 1 Draft 3

Figure 3 shows the Wright map for Task 1 Draft 3. The Draft was anchored to 0.00 logits. More participants were spread out toward the top of the Figure, which indicates that they were scoring relatively highly. Organization was the most difficult criterion, followed by cohesion and content. The easiest criterion was vocabulary.

Figure 3. *Wright Map for Task 1 Draft 3*

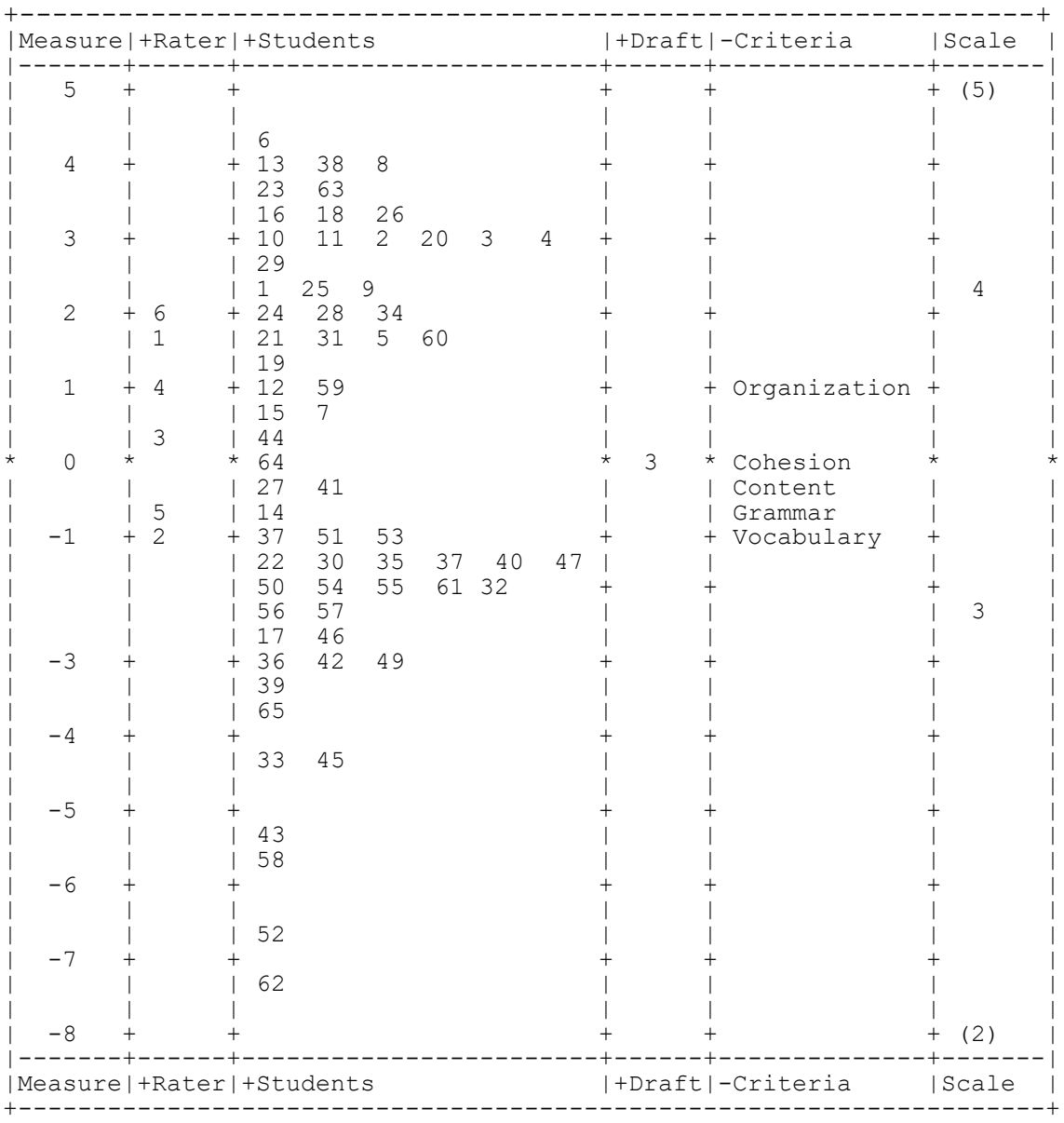


Table 12 shows the Rasch statistics for the six raters for Task 1 Draft 3. The infit MNSQ and the outfit MNSQ statistics were acceptable as they ranged from .67 to 1.34, and 0.63 to 1.45, respectively. The Rasch separation statistic was 6.21 and the Rasch reliability estimate was excellent at .97.

Table 12. *Rater Measurement Report for Task 1 Draft 3*

Rater	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt. measure correlation
6	3.94	2.10	0.19	1.24	1.90	1.21	1.50	0.67
1	3.89	1.79	0.12	0.97	-0.40	0.96	-0.40	0.79
4	3.71	0.93	0.21	1.34	2.40	1.45	2.60	0.72
3	3.54	0.23	0.22	0.99	0.00	0.97	-0.10	0.73
5	3.29	-0.72	0.22	0.75	-1.70	0.69	-1.90	0.78
2	3.25	-0.91	0.22	0.67	-2.30	0.63	-2.30	0.74

Note. RMSE = .20; *SD* = 1.24; Separation = 6.21; Strata = 8.62; Reliability = .97; Chi-square = 224.4; *df* = 5; $p < .01$. Inter-rater agreement opportunities = 780; Exact agreements = 413 (52.9%); Expected = 408.2 (52.3%).

The Rasch descriptive statistics for the participants for Task 1 Draft 3 are displayed in Table 13. The infit MNSQ and the outfit MNSQ statistics ranged from 0.33 to 3.26, and -0.87 to 3.36, respectively. Participants 21 and 46 underfit the Rasch model, and three participants overfit the model. The point measure correlations ranged from .67 to .79. The Rasch separation statistic was 4.65, and the Rasch reliability estimate was .96.

Table 13. *Student Measurement Report for Task 1 Draft 3*

Student	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
6	4.55	4.43	0.74	0.79	-0.30	0.68	0.40	0.67
13	4.44	3.99	0.57	0.82	-0.40	0.74	-0.60	0.56
38	4.44	3.98	0.56	1.24	0.80	1.28	0.80	0.59
8	4.42	3.89	0.72	0.37	-1.60	0.32	-1.60	0.85
63	4.37	3.73	0.70	0.77	-0.60	0.69	-0.70	0.64
23	4.36	3.68	0.54	1.67	2.30	1.83	2.50	-0.51
26	4.29	3.39	0.54	0.91	-0.20	0.85	-0.40	0.38
16	4.28	3.35	0.57	1.20	0.60	1.24	0.60	0.51
18	4.28	3.35	0.57	0.49	-1.60	0.42	-1.60	0.68
20	4.24	3.15	0.48	0.91	-0.20	-0.87	-0.30	0.57
10	4.21	3.03	0.57	0.51	-1.40	0.43	-1.60	0.58
11	4.21	3.03	0.57	0.99	-0.00	0.98	0.00	0.57
2	4.19	2.91	0.68.	0.93	0.00	1.02	0.20	0.18
3	4.19	2.91	0.68	0.93	0.00	1.02	0.20	0.18
4	4.19	2.91	0.68	0.46	-1.40	0.40	-1.40	0.60
29	4.10	2.79	0.56	1.30	0.90	1.34	0.90	0.42
25	4.10	2.48	0.57	1.15	0.50	1.19	0.50	0.46
1	4.10	2.45	0.67	0.34	-2.00	0.31	-2.00	0.52

Table 13 (continues)

Table 13 (continued)

Student	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
9	4.10	2.45	0.67	0.61	-0.90	0.56	-1.00	0.25
24	4.02	2.14	0.58	0.94	0.00	0.95	0.00	0.55
28	4.05	2.14	0.58	0.38	-1.70	0.32	-1.80	0.57
34	4.04	2.11	0.57	1.05	0.20	1.10	0.30	0.61
21	3.99	1.80	0.59	3.26	3.20	3.36	3.20	0.52
31	3.99	1.79	0.57	0.74	-0.50	0.77	-0.40	0.22
60	3.95	1.56	0.70	0.89	0.00	0.91	0.00	-0.10
5	3.95	1.55	0.67	0.49	-1.40	0.43	-1.40	0.37
19	3.93	1.42	0.56	1.19	0.60	1.23	0.60	0.60
12	3.87	1.10	0.56	1.38	1.00	1.39	0.90	0.28
59	3.86	1.04	0.74	0.65	-0.40	0.69	-0.30	-0.31
15	3.81	0.80	0.55	1.56	1.40	1.70	1.50	0.82
7	3.77	0.62	0.69	0.84	-0.20	0.79	-0.20	0.84
44	3.74	0.46	0.58	0.63	-0.80	0.65	-0.70	0.76
64	3.60	-0.08	0.73	1.22	0.50	1.20	0.50	0.68
41	3.57	-0.22	0.58	0.62	-0.90	0.65	-0.70	0.63
27	3.50	-0.47	0.53	0.49	-2.10	0.46	-2.00	0.82
14	3.43	-0.74	0.56	1.06	0.20	1.08	0.30	0.67
51	3.35	-1.03	0.65	1.23	0.70	1.35	0.90	-0.56
53	3.35	-1.03	0.65	1.08	0.30	1.11	0.40	-0.21
37	3.35	-1.05	0.55	0.77	-0.60	0.69	-0.70	0.55
50	3.32	-1.19	0.56	0.99	0.00	0.97	0.00	0.43
22	3.29	-1.31	0.54	0.97	0.00	1.01	0.10	0.48
35	3.28	-1.36	0.56	0.79	-0.50	0.73	-0.60	0.48
30	3.27	-1.37	0.49	0.81	-0.50	0.72	-0.70	0.45
40	3.26	-1.42	0.47	0.85	-0.40	0.79	-0.50	0.47
54	3.26	-1.44	0.63	1.35	1.10	1.43	1.20	-0.60
55	2.36	-1.44	0.63	1.20	0.70	1.24	0.80	-0.31
61	3.26	-1.44	0.63	1.20	0.70	1.24	0.80	-0.31
47	3.25	-1.50	0.55	0.76	-0.70	0.71	-0.60	0.55
32	3.21	-1.67	0.56	0.62	-1.10	0.52	-1.20	0.56
48	3.18	-1.80	0.55	0.71	-0.90	0.61	1.00	0.54
56	3.11	-2.24	0.64	0.73	-0.80	0.72	-0.80	0.60
57	3.03	-2.77	0.59	1.09	0.30	1.05	0.20	0.51
46	3.03	-2.72	0.56	3.06	4.00	2.85	3.40	-0.12
17	3.03	-2.77	0.59	1.09	0.30	1.05	0.20	0.51
36	2.99	-3.03	0.60	1.91	1.80	2.05	1.80	0.51
42	2.99	-3.05	0.57	1.32	0.90	1.38	0.90	0.47
49	2.99	-3.05	0.57	0.78	-0.50	0.72	-0.60	0.46
39	2.94	-3.39	0.60	1.58	1.30	1.62	1.20	0.73
65	2.89	-3.70	0.77	0.36	-1.10	0.27	-1.30	0.65
45	2.75	-4.44	0.61	0.82	-0.30	0.77	-0.30	0.40
33	2.75	-4.45	0.58	0.33	-2.30	0.28	-2.20	0.84
43	2.57	-5.18	0.61	0.56	-1.10	0.54	-1.00	0.73
58	2.49	-5.51	0.71	0.60	-0.80	0.51	-0.90	0.61
52	2.20	-6.84	0.64	0.95	-0.10	0.94	-0.10	0.29
62	2.14	-7.26	0.66	1.18	0.70	1.16	0.60	-0.08

Note. RMSE = .61; *SD* = 2.84; Separation = 4.65; Strata = 6.54; Reliability = .96; Chi-square = 1428.3; *df* = 64; $p < .01$.

The Rasch statistics for Task 1 Draft 3 are shown in Table 14. The infit MNSQ and the outfit MNSQ statistics were excellent at 1.00 and .99, respectively. All statistics were within acceptable ranges.

Table 14. *Draft Measurement Report for Task 1 Draft 3*

Draft	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt. measure correlation
3	3.62	0.00	0.07	1.00	0.00	0.99	-0.10	.79

The Rasch statistics for the five rating criteria for Task 1 Draft 3 are displayed in Table 15. The infit MNSQ and the outfit MNSQ statistics were acceptable as they ranged from 0.94 to 1.15, and from 0.82 to 1.13, respectively. Organization was the most difficult criterion and vocabulary was the easiest. The Rasch separation statistic was 4.80, and the Rasch reliability estimate was excellent at .92.

Table 15. *Criteria Measurement Report for Task 1 Draft 3*

Criterion	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	*Outfit ZSTD*	Pt. measure correlation
Org	3.83	0.90	0.17	1.15	1.40	1.13	*1.10*	.78
Content	3.66	0.14	0.16	0.86	-1.40	0.82	*-1.50*	.82
Cohesion	3.60	-0.11	0.16	0.94	-0.50	0.90	*-0.80*	.80
Grammar	3.55	-0.27	0.17	1.09	0.90	1.11	*0.90*	.78
Vocab	3.45	-0.66	0.17	0.95	-0.40	0.97	*-0.20*	.75

Note. RMSE = .17; *SD* = .55; Separation = 3.35; Strata = 4.80; Reliability = .92; Chi-square = 48.7; *df* = 4; $p < .01$. Vocab = Vocabulary; Org = Organization

The Rasch rating scale statistics for Task 1 Draft 3 are displayed in Table 16. Each point on the scale showed excellent fit to the Rasch model, and as shown in the two threshold columns, each point showed good separation from the other points on the scale.

Table 16. *Category Statistics for Task 1 Draft 3*

Score	Count (%)	Average measure	Outfit MNSQ	Rasch-Andrich threshold (*SE*)	Rasch-Thurstone threshold
2	41 (5)	-4.25	0.90	—	Low
3	298 (34)	-1.25	1.00	-4.78 (0.20)	-4.87
4	421 (48)	2.10	1.00	0.11 (0.11)	0.09
5	110 (13)	4.49	1.00	4.76 (0.13)	4.76

FACETS Analyses for Task 2 Draft 1

Table 17 shows the Rasch statistics for the six raters for Task 2 Draft 1. Rater 4 slightly misfit the Rasch model with infit and outfit MNSQ values above the 0.50–1.50 range. Rater 5 was the most severe rater, while Rater 6 was the most lenient. The Rasch separation statistic was 6.04 and the Rasch reliability estimate was .97.

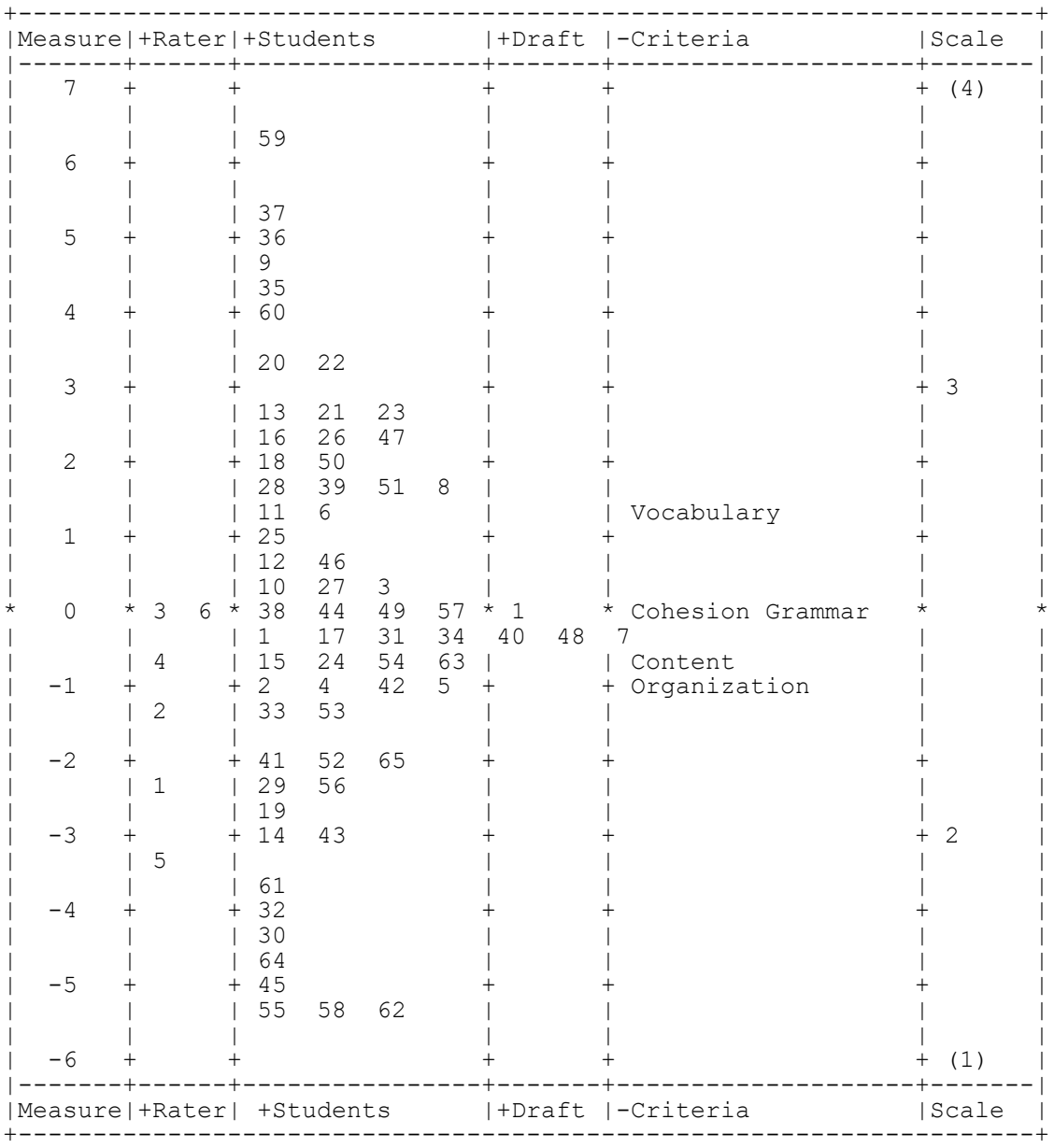
Table 17. *Rater Measurement Report for Task 2 Draft 1*

Rater	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt. measure correlation
6	2.53	0.14	0.20	1.06	0.50	1.10	0.70	.76
3	2.49	-0.01	0.24	1.03	0.20	1.08	0.40	.65
4	2.30	-0.78	0.23	1.52	3.10	1.62	2.90	.69
2	2.20	-1.27	0.24	0.89	-0.70	0.86	-0.60	.66
1	2.06	-2.26	0.14	0.82	-2.20	0.73	-2.50	.80
5	1.95	-3.36	0.25	0.91	-0.50	0.83	-0.70	.78

Note. RMSE = .22; *SD* = 1.34; Separation = 6.04; Strata = 8.39; Reliability = .97; Chi-square = 199.10; *df* = 5; $p < .01$.

Figure 4 shows the Wright map for Task 2 Draft 1. As shown in the Figure, the Draft was anchored to 0.00 logits because it was a generic element. Unlike Task 1, vocabulary and cohesion were the most difficult criteria to score on, and organization was the easiest.

Figure 4. *Wright Map for Task 2 Draft 1*



The Rasch descriptive statistics for the participants for Task 2 Draft 1 are shown in Table 18. The infit MNSQ and the outfit MNSQ statistics ranged from 0.10 to 2.66, and from 0.08 to 2.83, respectively. Student 22 underfit the Rasch model. The Rasch separation statistic was 3.81 and the Rasch reliability estimate was excellent at .94.

Table 18. *Student Measurement Report for Task 2 Draft 1*

Student	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
59	3.32	6.33	0.74	0.58	-1.10	0.45	-1.00	.76
37	3.16	5.46	0.66	1.19	0.50	1.17	0.40	.38
36	3.09	5.01	0.66	0.93	0.00	0.98	0.10	.26
9	3.05	4.64	0.87	0.11	-1.90	0.08	-1.90	.00
35	3.00	4.17	0.63	0.51	-1.30	0.37	-1.40	.48
60	2.98	3.96	0.79	1.00	0.10	0.90	0.00	.48
22	2.93	3.47	0.67	2.66	2.50	2.67	2.10	-.16
20	2.90	3.28	0.57	1.03	0.20	1.13	0.40	.41
13	2.83	2.79	0.62	1.17	0.50	1.43	0.90	.55
21	2.80	2.64	0.63	1.48	1.10	1.72	1.30	-.22
23	2.80	2.64	0.63	1.55	1.30	1.65	1.20	.38
47	2.77	2.48	0.60	1.81	1.90	1.85	1.60	.02
16	2.76	2.42	0.60	0.97	0.00	0.80	-0.30	.37
26	2.73	2.26	0.61	1.20	0.60	1.34	0.80	.60
50	2.70	2.11	0.61	0.85	-0.30	1.00	0.10	.81
18	2.69	2.07	0.59	0.82	-0.40	0.73	-0.50	.52
8	2.63	1.83	0.68	1.22	0.70	1.19	0.60	.16
39	2.60	1.68	0.59	0.56	-1.50	0.46	-1.40	.76
51	1.59	1.64	0.74	0.73	-0.59	0.63	-0.60	.66
28	2.57	1.56	0.58	0.67	-1.00	0.58	-1.10	.67
11	2.53	1.40	0.57	1.36	1.10	1.38	1.00	.18
6	2.52	1.36	0.69	1.61	1.60	1.74	1.60	-.25
25	1.40	0.89	0.57	1.08	0.230	1.00	0.10	.40
12	2.36	0.74	0.58	0.52	-1.80	0.43	-1.60	.79
46	2.31	0.53	0.64	0.45	-1.50	0.31	-1.50	.76
10	2.28	0.29	0.60	0.83	-0.40	0.78	-0.30	.54
3	2.26	0.30	0.78	0.76	-0.30	0.77	-0.10	.43
27	2.25	0.22	0.59	0.54	-1.60	0.44	-1.40	.74
38	2.23	0.12	0.66	3.18	3.20	2.83	2.30	.20
44	2.23	0.12	0.65	0.73	-0.50	0.51	-0.90	.56
49	2.23	0.12	0.65	0.77	-0.40	0.67	-0.40	.81
57	2.21	0.01	0.75	0.69	-0.60	0.52	-0.70	.64
48	2.15	-0.30	0.65	0.64	-0.80	0.51	-0.90	.78
31	2.15	-0.32	0.67	1.51	1.00	1.36	0.70	.62
34	2.15	-0.37	0.64	0.77	-0.40	0.90	0.00	.39
17	2.14	-0.37	0.64	0.77	-0.40	0.90	0.00	.39
1	2.14	-0.37	0.86	0.88	0.00	1.02	0.20	-.11
7	2.14	0.37	0.86	0.72	-0.20	0.58	-0.30	.21
40	2.14	-0.39	0.55	0.69	-0.80	0.57	-0.90	.70
24	2.12	-0.53	0.63	1.42	1.00	1.40	0.80	.45
54	2.12	-0.57	0.78	1.37	0.80	1.65	1.00	.65
63	2.12	-0.57	0.78	0.46	-1.20	0.34	-1.20	.66
15	2.09	-0.80	0.67	0.42	-1.40	0.30	-1.50	.60
42	2.05	-1.12	0.63	0.53	-1.20	0.46	-1.20	.54
2	2.05	-1.15	0.89	0.10	-1.90	0.08	-1.80	.00
4	2.05	-1.15	0.89	0.10	-1.90	0.08	-1.80	.00
5	2.05	-1.15	0.89	0.10	-1.90	0;08	-1.80	.00
53	2.04	-1.19	0.80	0.69	-0.40	0.52	0.60	.28

Table 18 (continues)

Table 18 (continued)

Student	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
33	2.04	-1.23	0.66	1.45	1.00	1.33	0.70	.70
52	1.98	-1.84	0.80	1.31	0.70	1.57	0.90	.26
65	1.98	-1.84	0.80	1.84	1.40	2.05	1.40	.58
41	1.97	-1.90	0.61	1.06	0.20	1.05	0.20	-.05
29	1.93	-2.34	0.69	1.61	1.20	1.52	0.90	.68
56	1.91	-2.47	0.79	1.30	0.70	0.98	0.10	-.61
19	1.89	-2.64	0.65	1.29	0.70	1.17	0.40	.51
43	1.84	-3.01	0.61	0.98	0.00	0.83	-0.20	.49
14	1.83	-3.05	0.63	0.47	-1.60	0.36	-1.50	.65
61	1.72	-3.66	0.76	0.46	-1.40	0.36	-1.30	.76
32	1.67	-3.86	0.60	0.87	-0.30	1.08	0.30	.55
30	1.58	-4.26	0.53	0.53	-1.90	0.42	-1.70	.80
64	1.45	-4.78	0.75	0.48	-1.50	0.39	-1.30	.83
45	1.38	-5.04	0.68	1.47	1.20	1.35	0.60	.67
55	1.31	-5.35	0.76	1.28	0.80	1.92	1.30	.32
58	1.31	-5.35	0.76	0.62	-1.00	0.47	-0.80	.76
62	1.31	-5.35	0.76	1.24	0.70	1.21	0.50	.40

Note. Sample RMSE = .69; *SD* = 2.64; Separation = 3.81; Strata = 5.42; Reliability = .94; Chi-square = 1018.60; *df* = 64; $p < .01$.

The Rasch statistics for Task 2 Draft 1 are shown in Table 19. The draft was anchored to 0.00 logits. The infit MNSQ and the outfit MNSQ statistics were acceptable at 0.99 and 0.96, respectively.

Table 19. *Draft Measurement Report for Task 2 Draft 1*

Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt. measure correlation
2.20	0.00	0.08	0.99	-0.10	0.96	-0.50	.78

The Rasch statistics for the five rating criteria for Task 2 Draft 1 are displayed in Table 20. The infit MNSQ and the outfit MNSQ statistics ranged from 0.83 to 1.02, and from 0.74 to 1.18, respectively. Vocabulary was the most difficult criterion followed by cohesion, while content was the easiest criterion. Each criterion showed acceptable fit to the Rasch model.

Table 20. *Criteria Measurement Report for Task 2 Draft 1*

Criterion	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt. measure correlation
Vocab	2.53	1.43	0.18	0.84	-1.50	0.77	-1.60	.76
Cohesion	2.22	0.10	0.19	0.83	-1.40	0.74	-1.80	.80
Grammar	2.20	0.00	0.19	1.17	1.40	1.18	1.10	.77
Org	2.10	-0.73	0.19	1.09	0.80	1.07	0.50	.77
Content	2.09	-0.80	0.19	1.02	0.10	1.03	0.20	.75

Note. RMSE = .19; *SD* = .88; Separation = 4.73; Strata = 6.64; Reliability = .96; Chi-square = 95.00; *df* = 4; $p < .01$. Vocab = Vocabulary; Org = Organization.

The Rasch rating scale statistics for Task 2 Draft 1 are shown in Table 21. The scale showed good fit to the Rasch model, and each point on the scale was clearly separated from the other points as shown by the two threshold columns.

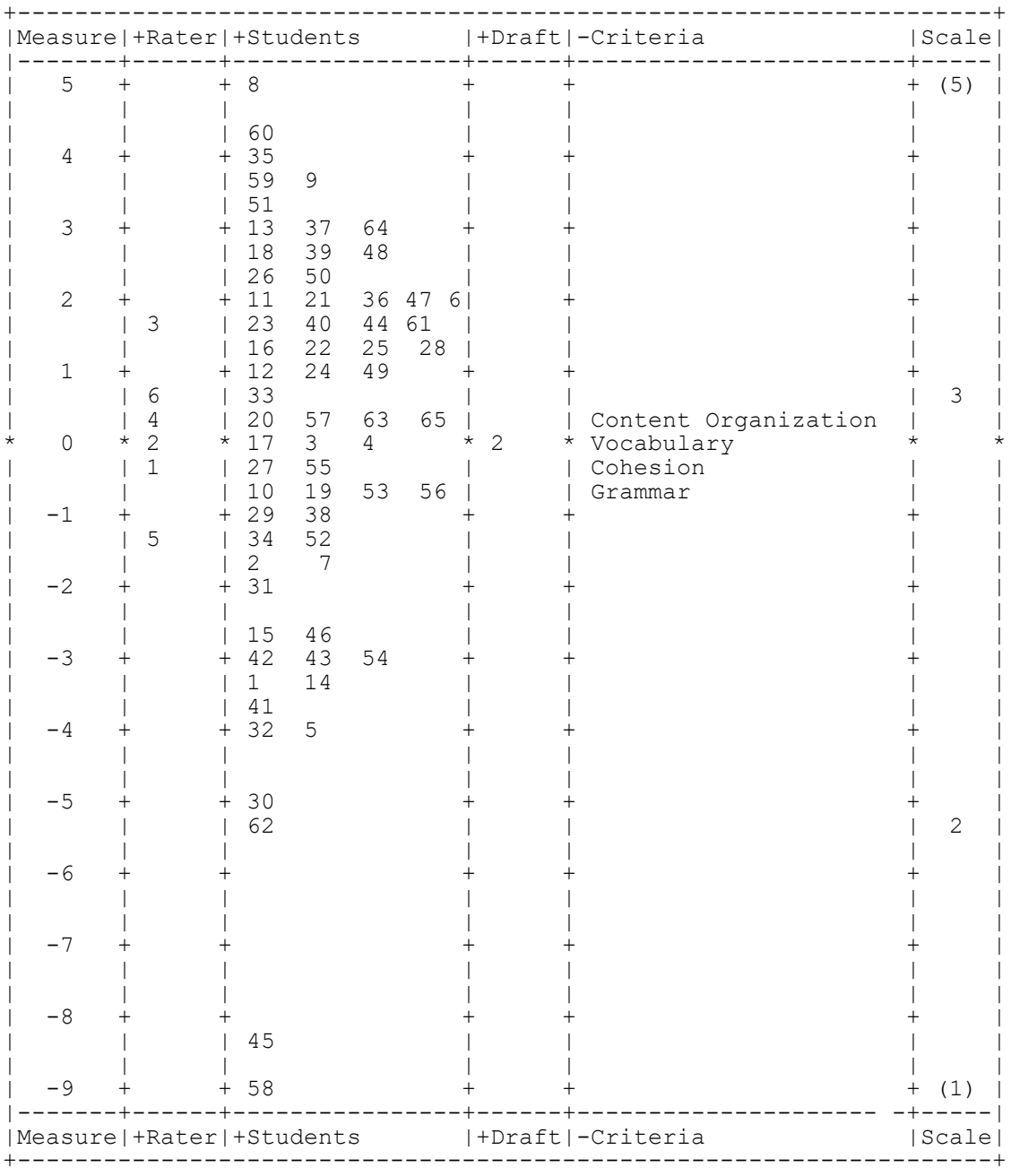
Table 21. *Category Statistics for Task 2 Draft 1*

Score	Count (%)	Average measure	Outfit MNSQ	Rasch-Andrich threshold (*SE*)	Rasch-Thurstone threshold
1	90 (10)	-5.88	1.00	—	Low
2	479 (55)	-2.32	1.00	-5.81 (0.15)	-5.82
3	286 (33)	1.32	0.90	0.04 (0.11)	0.03
4	15 (2)	3.91	1.30	5.77 (0.31)	5.76

FACETS Analyses for Task 2 Draft 2

Figure 5 shows the Wright map for Task 2 Draft 2. As shown in the Figure, the draft was anchored to 0.00 logits. Students' scores were generally spread out toward the top of the Figure. On Draft 2, organization was the most difficult criterion again, while grammar was the easiest.

Figure 5. *Wright Map for Task 2 Draft 2*



The rater measurement report is displayed in Table 22. The infit MNSQ and the outfit MNSQ statistics ranged from 0.74 to 1.36, and from 0.67 to 1.41, respectively; thus, all raters showed acceptable fit to the Rasch model. The Rasch separation statistic was 4.61, and the Rasch reliability estimate was excellent at .96.

Table 22. *Rater Measurement Report for Task 2 Draft 2*

Rater	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt. measure correlation
3	3.16	1.63	0.21	1.36	2.60	1.41	2.50	.48
6	2.99	0.54	0.19	0.96	-0.30	0.89	-0.70	.81
4	2.97	0.35	0.22	1.17	1.20	1.28	1.60	.69
2	2.91	-0.05	0.23	0.90	-0.60	0.86	-0.70	.79
1	2.88	-0.22	0.13	0.94	-0.70	0.93	-0.60	.78
5	2.65	-1.35	0.25	0.74	-1.50	0.67	-1.50	.85

Note. RMSE = .21; *SD* = .96; Separation = 4.61; Strata = 6.48; Reliability = .96; Chi-square = 100.40; *df* = 5; $p < .01$;

The Rasch descriptive statistics for the participants for Task 2 Draft 2 are displayed in Table 23. The infit MNSQ and the outfit MNSQ statistics ranged from 0.02 to 2.47 and 0.02 to 2.26, respectively. Students 8, 17, and 30 underfit the Rasch model. The Rasch separation statistic was 4.28, and the Rasch reliability estimate was .95.

Table 23. *Student Measurement Report for Task 2 Draft 2*

Student	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
8	3.96	4.94	0.72	2.35	1.80	2.32	1.70	0.37
60	3.81	4.17	0.68	0.60	-0.80	0.57	-0.80	0.44
35	3.74	3.89	0.53	0.53	-1.80	0.49	-1.80	0.78
59	3.71	3.73	0.65	1.14	0.40	1.17	0.50	-0.32
9	3.68	3.61	0.63	0.94	0.00	0.95	0.00	0.11
51	3.60	3.32	0.63	0.66	-1.10	0.64	-1.10	0.68
13	3.53	3.06	0.53	1.73	2.10	1.93	2.30	-0.57
37	3.53	3.05	0.53	1.40	1.30	1.34	1.00	0.35
64	3.50	2.92	0.63	1.78	2.20	1.73	2.10	0.12
18	3.46	2.78	0.53	0.96	0.00	0.97	0.00	0.32
39	3.45	2.76	0.54	0.80	-0.50	0.77	-0.50	0.43
48	3.43	2.69	0.55	0.91	-0.10	0.86	-0.20	0.33
50	3.36	2.38	0.56	0.63	-1.10	0.53	-1.20	0.58
26	3.35	2.36	0.53	0.57	-1.60	0.53	-1.70	0.75
36	3.30	2.13	0.58	1.70	1.60	1.76	1.50	0.56
61	3.29	2.09	0.67	0.96	0.00	0.92	0.99	0.12
21	3.28	2.07	0.54	1.29	0.90	1.31	0.90	-0.03
47	3.28	2.05	0.58	0.98	0.00	0.95	0.00	0.71
6	3.25	1.94	0.71	0.72	-0.40	0.65	-0.50	0.38
11	3.24	1.90	0.56	1.01	0.10	1.01	0.10	0.26

Table 23 (continues)

Table 23 (continued)

Student	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
23	3.22	1.77	0.55	0.55	-1.50	0.49	-1.50	0.74
44	3.20	1.71	0.60	0.61	-0.80	0.54	-0.90	0.28
40	3.19	1.64	0.51	0.74	-0.60	0.77	-0.40	0.09
22	3.16	1.46	0.57	1.96	2.10	1.77	1.60	-0.12
25	3.16	1.46	0.57	1.96	2.10	1.77	1.60	-0.12
28	3.16	1.46	0.57	0.41	-2.00	0.34	-2.00	0.83
16	3.12	1.24	0.59	1.29	0.70	1.40	0.80	0.89
24	3.10	1.12	0.59	0.78	-0.40	0.66	-0.60	0.33
49	3.07	0.94	0.63	0.85	-0.10	0.87	0.00	0.10
12	3.06	0.88	0.61	0.91	0.00	0.97	0.10	-0.11
33	3.02	0.60	0.63	0.92	0.00	0.93	0.00	0.49
20	3.00	0.44	0.55	1.23	0.60	1.22	0.50	0.43
57	3.00	0.43	0.81	0.05	-2.40	0.04	-2.40	0.00
63	3.00	0.43	0.81	0.05	-2.40	0.04	-2.40	0.00
65	3.00	0.43	0.81	0.05	-2.40	0.04	-2.40	0.00
17	2.95	0.11	0.62	2.22	2.00	2.14	1.80	0.24
3	2.95	0.07	0.79	0.68	-0.30	0.75	-0.10	-0.36
4	2.95	0.07	0.79	1.68	1.00	1.74	1.10	0.30
55	2.91	-0.21	0.78	0.37	-1.10	0.28	-1.30	0.62
27	2.87	-0.40	0.62	0.90	0.00	0.90	0.00	0.44
10	2.83	-0.64	0.60	0.68	-0.70	0.64	-0.70	0.17
19	2.83	-0.64	0.60	0.58	-1.00	0.50	-1.10	0.31
53	2.81	-0.77	0.72	0.76	0.30	0.79	0.20	0.24
56	2.81	-0.77	0.72	2.47	2.30	2.66	2.30	0.56
38	2.74	-1.11	0.55	1.61	2.00	1.62	1.80	0.20
29	2.73	-1.13	0.58	0.72	-0.70	0.75	-0.40	0.30
52	2.71	-1.25	0.67	0.62	-1.10	0.56	-1.10	0.76
34	2.67	-1.40	0.54	0.67	-1.40	0.63	-1.50	0.70
2	2.57.	-1.80	0.64	1.04	0.20	1.05	0.30	0.03
7	2.57	-1.80	0.64	1.37	1.60	1.38	1.70	-0.71
31	2.52	-2.00	0.55	1.02	0.20	0.96	0.00	0.30
15	2.39	-2.53	0.56	1.25	0.90	1.27	0.90	0.58
46	2.34	-2.74	0.61	0.98	0.00	0.80	-0.30	0.35
54	2.28	-2.98	0.70	0.82	-0.30	0.73	-0.60	0.53
42	2.25	-3.14	0.65	1.11	0.30	1.26	0.60	0.06
43	2.25	-3.14	0.65	0.87	-0.20	0.69	0.40	0.40
14	2.24	-3.18	0.59	1.07	0.30	1.18	0.50	0.28
1	2.24	--3.20	0.78	0.85	-0.10	0.73	-0.30	0.45
41	2.17	-3.61	0.72	0.90	0.00	0.72	-0.20	0.24
5	2.13	-3.94	0.97	0.99	0.20	1.48	0.70	-0.59
32	2.11	-4.12	0.84	0.60	-0.40	0.43	-0.60	0.29
30	2.04	-4.93	0.70	2.12	1.70	2.37	1.70	0.45
62	2.00	-5.49	1.20	0.02	-1.50	0.02	-1.50	0.00
45	1.66	-8.27	0.56	0.95	-0.10	0.91	-0.20	0.44
58	1.50	-8.92	0.65	0.79	-1.10	0.77	-1.10	0.62

Note. RMSE = .65; *SD* = 2.81; Separation = 4.28; Strata = 6.05; Reliability = .95; Chi-square = 1296.60; *df* = 64; $p < .01$.

The Rasch statistics for Task 2 Draft 2 are shown in Table 24. The infit MNSQ and the outfit MNSQ statistics were 1.00 and .99, respectively.

Table 24. *Draft Measurement Report for Task 2 Draft 2*

Draft	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt. measure correlation
2	2.94	0.00	0.08	1.00	0.10	0.99	-0.20	.78

The Rasch statistics for the five rating criteria for Task 2 Draft 2 are displayed in Table 25. Content was the most difficult criterion and grammar the easiest to score highly on. The infit MNSQ and the outfit MNSQ statistics ranged from 0.92 to 1.08 and 0.83 to 1.17, respectively. All of the criteria met the 0.50–1.50 infit MNSQ criterion. The Rasch separation statistic was 2.52 and the Rasch reliability estimate was good at .86.

Table 25. *Criteria Measurement Report for Task 2 Draft 2*

Criterion	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt. measure correlation
Content	3.00	0.44	0.17	1.05	0.50	1.00	0.00	.74
Org	2.99	0.38	0.17	1.04	0.40	1.02	0.20	.75
Vocab	2.95	0.12	0.17	0.92	-0.70	0.90	-0.70	.82
Cohesion	2.90	-0.27	0.17	0.93	-0.50	0.83	-1.20	.78
Grammar	2.83	-0.67	0.18	1.08	0.70	1.17	1.10	.78

Note. RMSE = .17; *SD* = .43; Separation = 2.52; Strata = 3.69; Reliability = .86; Chi-square = 28.90; *df* = 4; $p < .01$. Vocab = vocabulary; Org = organization.

The rating scale statistics for Task 2 Draft 2 are displayed in Table 26. Fit to the Rasch model was excellent and each point on the scale was well separated from the other points, as shown in both threshold columns.

Table 26. *Category Statistics for Task 2 Draft 2*

Score	Count (%)	Average measure	Outfit MNSQ	Rasch-Andrich threshold (*SE*)	Rasch-Thurstone threshold
1	14 (2)	-8.56	1.10	—	Low
2	213 (24)	-3.24	1.00	-8.76 (0.37)	-8.76
3	479 (55)	0.80	1.00	-1.92 (0.12)	-1.92
4	161 (19)	2.94	1.00	3.10 (0.11)	3.08
5	3 (0)	4.27	0.90	7.58 (0.59)	7.58

FACETS Analyses for Task 2 Draft 3

Figure 6 shows the Wright map for Task 2 Draft 3. As shown in the Table, the draft was anchored to 0.00 logits. Compared to Task 1 Draft 1, the students' scores were spread out. Organization was the most difficult criterion, while cohesion was the easiest. Organization continued to be difficult to score highly on.

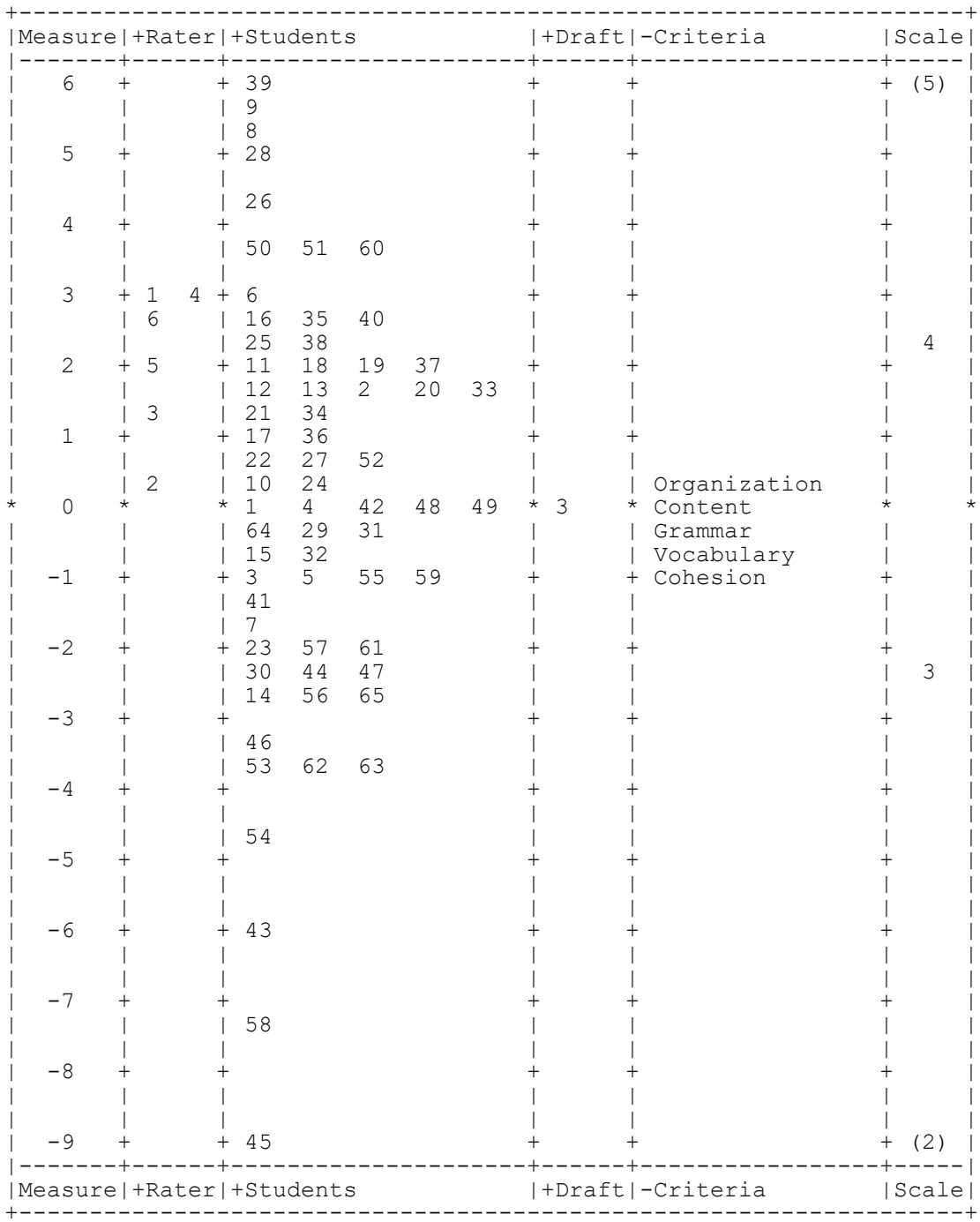
The Rasch statistics for the six raters are shown in Table 27. The raters met the 0.50–1.50 fit criterion. Rater 1 was the most severe rater, while Rater 2 was the most lenient. The infit MNSQ and the outfit MNSQ statistics ranged from 0.63 to 1.46 and 0.59 to 1.42, respectively. The point measure correlations ranged from .56 to .85. The Rasch separation statistic was 5.68, and Rasch reliability estimate was .97.

Table 27. *Rater Measurement Report for Task 2 Draft 3*

Rater	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt. measure correlation
1	4.15	3.15	0.12	1.00	0.00	1.01	0.10	.79
4	4.13	3.04	0.23	1.46	3.20	1.42	1.80	.56
6	4.09	2.85	0.19	0.91	-0.60	0.90	-0.70	.85
5	3.95	2.16	0.21	0.63	-3.20	0.59	-3.20	.87
3	3.77	1.30	0.21	1.18	1.10	1.20	1.10	.70
2	3.49	0.22	0.22	0.84	-1.10	0.88	-0.60	.72

Note. RMSE = .20; *SD* = 1.14; Separation = 5.68; Strata = 7.90; Reliability = .97; Chi-square = 175.90; *df* = 5; $p < .01$.

Figure 6. *Wright Map for Task 2 Draft 3*



The Rasch descriptive statistics for the participants for Task 2 Draft 3 are shown in Table 28. The infit MNSQ and the outfit MNSQ statistics ranged from 0.02 to 2.47,

and 0.02 to 2.66, respectively. The Rasch separation statistic was 4.16, and the Rasch reliability estimate was excellent at .95.

Table 28. *Student Measurement Report for Task 2 Draft 3*

Student	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
39	4.96	5.70	1.85	Maximum	—	—	—	0.00
9	4.97	5.83	1.11	1.11	0.30	0.71	0.50	0.26
8	4.95	5.44	1.12	0.89	0.00	0.42	0.20	0.43
28	4.93	5.06	1.07	0.94	0.20	0.45	0.00	0.36
26	4.85	4.18	0.81	1.430	0.90	2.11	1.30	-0.26
50	4.78	3.72	0.78	1.09	0.30	1.14	0.40	-0.01
60	4.77	3.65	1.07	1.08	0.30	1.26	0.50	-0.25
51	4.75	3.53	1.08	1.06	0.30	1.13	0.40	-0.12
6	4.64	3.02	0.75	0.92	0.00	1.23	0.50	0.55
35	4.59	2.80	0.60	0.96	0.00	0.90	-0.10	0.31
16	4.55	2.64	0.56	0.89	-0.10	0.91	-0.10	0.48
40	4.54	2.60	0.50	0.78	-0.90	0.74	-0.90	0.61
25	4.44	2.24	0.55	0.66	-1.20	0.61	-1.30	0.70
38	4.44	2.22	0.54	1.85	3.00	2.18	3.70	-0.24
37	4.40	2.08	0.56	0.66	-1.60	0.65	-1.70	0.81
11	4.38	2.02	0.56	0.94	0.00	0.91	0.00	0.31
18	4.38	2.02	0.56	0.69	-0.70	0.68	0.70	0.51
19	4.38	2.02	0.56	0.41	-1.80	0.39	-1.80	0.73
20	4.32	1.77	0.47	1.90	2.50	1.88	2.30	-0.08
13	4.31	1.71	0.55	0.60	-1.10	-.56	-1.10	0.48
33	4.30	1.67	0.52	0.78	-1.0	0.76	-1.00	0.53
34	4.30	1.67	0.52	0.89	-0.40	0.88	-0.40	0.34
12	4.27	1.57	0.57	1.12	0.40	1.12	0.40	0.65
2	4.27	1.54	0.66	2.22	2.40	2.23	2.20	-0.10
21	4.23	1.38	0.53	0.53	-1.80	0.49	-1.80	0.68
36	4.17	1.13	0.52	1.13	0.50	1.08	0.30	0.58
17	4.17	1.11	0.54	1.20	0.60	1.15	0.40	0.93
22	4.10	0.81	0.54	2.19	2.70	2.27	2.60	0.67
27	4.10	0.81	0.54	1.06	0.20	0.98	0.00	0.38
52	4.10	0.79	0.64	0.67	-.90	0.64	-0.90	0.64
10	4.00	0.26	0.53	0.88	-0.20	0.89	-0.10	0.72
24	3.99	0.23	0.55	1.75	1.60	1.84	1.70	0.43
42	3.95	0.01	0.57	0.24	-2.30	0.21	-2.40	0.45
48	3.95	0.01	0.57	0.38	-1.60	0.39	-1.90	0.18
49	3.95	0.01	0.57	0.32	-1.80	0.31	-1.90	0.18
1	3.93	-0.10	0.65	0.74	-0.50	0.75	0.40	0.73
4	3.93	-0.10	0.65	0.55	-1.20	0.51	-1.20	0.53
64	3.93	-0.10	0.70	1.27	0.60	1.28	0.60	0.42
29	3.90	-0.25	0.67	1.96	1.70	1.99	1.60	0.88
31	3.87	-0.38	0.59	1.61	1.20	1.60	1.10	-0.23
15	3.82	-0.59	0.54	0.41	-2.20	0.37	-2.20	0.75

Table 28 (continues)

Table 28 (continued)

Student	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
32	3.80	-0.71	0.57	1.52	1.10	1.55	1.10	0.23
3	3.73	-1.01	0.70	0.44	-1.10	0.42	-1.20	0.79
5	3.73	-1.01	0.70	0.47	-1.10	0.47	-1.00	0.76
55	3.71	-1.10	0.70	0.54	-0.70	0.56	-0.60	-0.30
59	3.71	-1.10	0.70	0.54	-0.70	0.56	-0.60	-0.30
41	3.66	-1.27	0.54	0.70	-0.70	0.67	-0.70	0.19
7	3.60	1.50	0.71	0.14	-2.40	0.14	-2.30	0.97
23	3.50	-1.87	0.53	2.16	2.70	2.23	2.70	-0.27
57	3.47	-1.99	0.64	0.79	-0.70	0.67	-0.80	0.54
61	3.47	-1.99	0.64	0.83	-0.30	0.80	-0.40	0.20
44	3.37	-2.36	0.51	0.80	-0.70	0.79	-0.80	0.45
47	3.37	-2.36	0.51	0.92	-0.20	0.91	-0.20	0.23
30	3.36	-2.40	0.46	0.67	-1.40	0.62	-1.50	0.64
56	3.28	-2.76	0.62	0.74	-0.90	0.75	-0.90	0.62
65	3.28	-2.76	0.62	1.00	0.10	1.00	0.00	-0.02
14	3.27	-2.80	0.59	2.39	2.60	2.79	2.60	0.04
46	3.14	-3.46	0.56	0.91	-0.10	0.85	-0.20	0.16
53	3.12	-3.57	0.66	1.03	0.20	1.03	0.20	-0.18
62	3.12	-3.57	0.66	1.06	0.20	1.07	0.30	-0.26
63	3.12	-3.57	0.66	1.68	1.60	1.83	1.80	0.33
54	2.97	-4.60	0.79	0.63	-0.40	0.67	-0.30	-0.17
43	2.76	-5.84	0.62	0.68	-0.60	0.63	-0.60	0.17
58	2.38	-7.41	0.64	0.92	-0.20	0.90	-0.30	0.28
45	2.11	-0.90	0.66	0.86	-0.20	0.74	-0.40	0.48

Note. RMSE = .69; *SD* = 2.88; Separation = 4.16; Strata = 5.88; Reliability = .95; Chi-square = 1160.60; *df* = 64; $p < .00$.

The Rasch statistics for Task 2 Draft 3 are shown in Table 29. The infit MNSQ and the outfit MNSQ statistics were .99 and 1.00, respectively.

Table 29. *Draft Measurement Report for Task 2 Draft 3*

Draft	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
3	3.95	0.00	0.08	0.99	0.00	1.00	0.00	.80

The Rasch statistics for the five rating criteria for Task 2 Draft 3 are displayed in Table 30. Organization was the most difficult criterion to score highly on, while cohesion was the easiest. Each criterion showed acceptable fit to the Rasch model. The infit MNSQ and the outfit MNSQ statistics ranged from 0.79 to 1.18 and 0.77 to 1.19,

respectively. The Rasch separation statistic was 2.18, and the Rasch reliability estimate was good at .83.

Table 30. *Criteria Measurement Report for Task 2 Draft 3*

Criterion	Fair average	Rasch measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt. measure correlation
Org	4.04	0.50	0.17	1.18	1.60	1.19	1.10	0.77
Content	4.00	0.29	0.17	0.79	-2.00	0.77	-1.50	0.81
Grammar	3.94	-0.06	0.17	1.01	0.10	0.98	0.00	0.81
Vocab	3.91	-0.21	0.17	1.11	1.00	1.19	1.30	0.76
Cohesion	3.84	-0.52	0.17	0.88	-1.10	0.86	-1.00	0.83

Note. RMSE = 17; *SD* = .37; Separation = 2.18; Strata = 3.23; Reliability = .83; Chi-square = 22.80; *df* = 4; *p* < .01; Chi-square = 3.40; *df* = 3; *p* < 00. Org = Organization; Vocab = Vocabulary.

The Task 2 Draft 3 rating scale statistics are displayed in Table 31. Each step displayed excellent fit to the Rasch model, and they were adequately separated, as shown in the two threshold columns.

Table 31. *Category Statistics for Task 2 Draft 3*

Score	Count (%)	Average measure	Outfit MNSQ	Rasch-Andrich threshold (*SE*)	Rasch-Thurstone threshold
2	21 (3)	-5.05	1.00	—	Low
3	177 (21)	-0.29	1.10	-4.80 (0.33)	-4.80
4	418 (50)	2.47	1.00	0.26 (0.12)	0.24
5	218 (26)	5.28	0.90	4.54 (0.11)	4.53

CHAPTER 5

RESULTS

In this chapter, the results of each hypothesis are discussed. Scores for three lexical and two cohesion indices were calculated using TALLED 1.41 (Zenker & Kyle, 2021), TAALES 2.2 (Kyle, Crossley, & Berger, 2018), and TAACO 2.04 (Crossley, Kyle, & Dascalu, 2019). Content words were used for lexical indices. Overall writing quality was calculated using FACETS. Table 32 shows the final list of dependent measure indices after checking multicollinearity. In order to have more detailed characteristics, lexical indices, cohesion indices, and overall writing quality were used.

Table 32. *Summary of the Final Dependent Measures With Index Names*

Metric	Index	Tools	Index name
Lexical diversity (a type-token ratio)	The moving-average type-token ratio for content words (50-word window) content words	TALLED 1.41	MATTR
Frequency	COCA academic frequency logarithm for content words	TAALES 2.2	COCAFrq
Lexical sophistication	COCA academic trigram association strength (low-frequency words)	TAALES 2.2	Trigrams
Local cohesion	All connective	TAACO 2.04	All connectives
Global cohesion	Adjacent lexical overlap between paragraphs (Content words)	TAACO 2.04	AdjacentP
Writing Quality	Overall writing quality scores (Rubrics)	Rubrics	Overall writing quality

Measures

The Moving-Average Type-Token Ratio (MATTR)

The measures of textual lexical diversity (MTLD) and the moving-average type-token ratio (MATTR) (Covington & McFall, 2010) are the most stable indices and perform well across in all text lengths in lexical diversity indices (Zenker & Kyle, 2021). MATTR is particularly stable across all text lengths (Zenker & Kyle, 2021). It calculates the moving average for all segments of a given text length. For example, in the case of 50-token texts, tokens 1-50, 2-51, 3-52 and so on of the type-token ratio for content words are calculated, and the type-token-ratio measurements are then averaged to yield the MATTR value.

COCA Academic Frequency (COCAFrq)

Kyle et al. (2018) suggested assessing the use of content words—nouns, verbs, adjectives, and adverbs—to track lexical development. The mean frequency for content words were analyzed using TAALES 2.2. (Kyle et al., 2018). This index is used because as learners become proficient, they are likely to use less frequent, or advanced words.

COCA Academic Trigrams (Trigrams)

Kyle et al. (2018) suggested the use of N-grams, which refer to multiword sequences, such as two-word sequences and three-word sequences, to track L2 writing and lexical development. N-grams were measured using the second version of the Tool for Analysis of Lexical Sophistication 2.0 (Kyle et al., 2018), which automatically calculates traditional and newer indices of lexical sophistication. It covers lexical features related to word frequency, word range, n-gram frequency, academic language, and psycholinguistic word properties. The second version includes the breadth and depth of

the available indices by expanding the word frequency, word range, and n-gram frequency indices, and by adding indices related to word recognition norms. Based on the study by Garner et al. (2020), trigram associations were analyzed using TAALES (Kyle & Crossley, 2015). The academic subsection of COCA was used.

All Connectives

Based on the study by Crossley et al. (2016), connectives were measured. Local cohesion indices included connectives (e.g., *because*, *so*, *although*, *whereas*, *moreover*, *and*, *or*, *first, until*, *also*, *moreover*, *however*, *but*). Global cohesion indices lexical overlap between paragraphs. These indices were analyzed using TAACO 2.0.4 (Crossley et al., 2016).

Adjacent Overlap Between Paragraphs (AdjacentP)

Adjacent overlap between paragraphs for global cohesion were measured. These indices were analyzed using TAACO 2.0.4. (Crossley et al., 2016).

Overall Writing Quality

To assess overall writing quality, an analytic rubric consisting of content, organization, vocabulary, cohesion, and grammar was used and total scores were calculated (see Appendix I). The Rasch measures derived from the FACETS analyses served as overall writing quality.

Results of Analyses for Each Hypothesis

Hypothesis 1: Target Vocabulary, Target Cohesion, and Overall Writing Quality

The first hypothesis predicted that the explicit instruction group would score significantly higher than the comparison group on the following dependent variables: (a) overall writing quality, (b) the number of target word the participants used from each

model essay in each task, and (c) the number of target cohesive devices that the participants used from each model essay. Overall writing quality was computed using FACETS 3.80.0 (Linacre, 2017) based on scoring provided by the six raters.

The Total Number of Words Written

The total number of words written were computed. In both tasks, participants were asked to write about 150 words on the first draft, 250 words on the second draft, and 300 words on the third draft. As shown in Table 33, the participants achieved the length target for all three drafts, although the participants in the explicit instruction group consistently wrote slightly longer drafts than those in the comparison group.

Table 33. *The Total Number of Words Written in Task 1*

	Draft 1		Draft 2		Draft3	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
M	172.62	164.42	253.41	241.29	317.38	295.32
95% CI U	182.82	172.70	238.41	257.70	338.31	312.49
95% CI L	164.42	156.14	236.42	224.88	296.45	278.16
SD	29.24	22.56	43.01	44.74	59.99	46.80
Skewness	1.27	1.17	0.24	1.91	0.71	1.32
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	2.01	2.09	-0.95	4.03	-0.52	1.55
SEK	4.05	0.82	0.79	0.82	0.79	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis. Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

Table 34 shows that the participants in both groups again met the length targets for the three drafts. In this case, the two groups wrote similar amounts.

Table 34. *The Total Number of Words Written in Task 2*

	Draft 1		Draft 2		Draft 3	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
M	175.21	177.06	239.79	246.03	292.35	290.39
95% CI U	184.50	186.73	257.11	258.28	307.25	312.28
95% CI L	165.91	167.40	222.48	233.79	277.46	268.50
SD	26.63	26.36	49.62	33.38	42.69	59.68
Skewness	0.87	0.85	1.20	0.77	0.91	4.01
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	0.20	0.31	1.85	0.34	0.61	19.05
SEK	0.40	0.82	0.79	0.82	0.79	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis. Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

Target Word Use

Table 35 shows the descriptive statistics for target word use in Task 1. Skewness was almost within -.5 to .5 range. Kurtosis ranged from -1.26 to .01. The explicit instruction group increasingly used target words.

Table 35. *Descriptive Statistics for Target Word Use in Task 1*

	Draft 1		Draft 2		Draft 3	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
M	1.59	1.10	5.18	1.39	6.91	1.77
SE	1.85	0.11	0.43	0.17	0.44	0.20
95% CI U	1.96	1.32	0.61	1.74	0.78	2.17
95% CI L	1.21	0.88	0.43	1.04	6.01	1.38
SD	1.08	0.60	2.50	0.96	2.59	1.09
Skewness	0.61	-0.02	0.03	0.59	0.04	-0.35
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	-0.18	0.01	-0.75	0.61	-1.26	-1.14
SEK	0.11	0.82	0.17	0.82	0.79	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis. Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

Table 36 shows the descriptive statistics for target word use in Task 2. Skewness was moderate, and kurtosis ranged from -.77 to 3.15. The explicit instruction group increasingly used target words, but the number of target words used was fewer than that used in Task 1.

Table 36. *Descriptive Statistics for Target Word Use in Task 2*

	Draft 1		Draft 2		Draft 3	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
M	0.94	0.26	1.68	0.32	2.50	0.55
SE	0.16	0.09	0.25	0.10	0.25	0.12
95% CI U	1.27	0.45	2.18	0.52	3.00	0.80
95% CI L	0.61	0.07	1.17	0.12	2.00	0.30
SD	0.95	0.51	1.45	0.54	1.44	0.68
Skewness	0.57	1.92	0.54	1.46	0.90	0.86
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	-0.77	3.15	-0.64	1.36	0.22	-0.33
SEK	0.79	0.82	0.79	0.82	0.79	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis.; Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

Target Cohesion Use

Table 37 shows the descriptive statistics for target cohesion use in Task 1. Skewness ranged from -.40 to .42, and kurtosis ranged from -6.30 to 1.27. The explicit instruction group increasingly used target cohesion.

Table 37. De*scriptive Statistics for Target Cohesion Use in Task 1*

	Draft 1		Draft 2		Draft 3	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
M	1.44	1.06	3.85	1.32	4.41	1.74
SE	1.53	0.10	0.28	-0.16	0.24	0.20
95% CI U	1.75	1.27	4.43	1.66	4.89	2.15
95% CI L	1.13	0.85	3.28	0.99	3.93	1.33
SD	0.90	0.57	1.65	0.91	1.37	1.13
Skewness	0.32	0.02	-5.60	0.71	0.84	-0.35
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	-0.55	0.34	-6.30	1.27	-0.92	-1.23
SEK	0.79	0.82	0.79	0.82	0.79	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis; Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

Table 38 shows the descriptive statistics for target cohesion use in Task 2. Target cohesion use was positively skewed. Kurtosis was below 3.00, except for the first and second draft for the comparison group.

Table 38. *Descriptive Statistics for Target Cohesion Use in Task 2*

	Draft 1		Draft 2		Draft 3	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
M	0.35	0.03	0.68	0.06	1.03	0.26
SE	0.09	0.03	0.17	0.45	0.17	0.08
95% CI U	0.54	0.10	0.33	-0.03	1.38	0.42
95% CI L	0.16	-0.03	1.03	0.16	0.68	0.09
SD	0.54	0.18	1.00	0.25	1.00	0.45
Skewness	1.23	5.57	1.67	3.73	0.52	1.16
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	0.64	31.00	2.64	12.72	-0.86	-0.70
SEK	0.79	0.82	0.40	0.82	0.79	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis; Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

Univariate Normal Assumptions and Descriptive Statistics

All of the data were interval in the present study. The independent variable consisted of two independent groups—explicit instruction and the comparison groups. Each participant was a member of only one of the two groups. The sample size for each group was more than 30, and the two groups were about equal. The assumption of a univariate normal distribution was checked for overall writing quality. Descriptive statistics, skewness, and kurtosis for Task 1 and Task 2 were examined.

Overall Writing Quality

Table 39 shows the descriptive statistics for overall writing quality. Skewness ranged from -1.09 to .92, and kurtosis ranged from -.26 to 1.20.

Table 39. *Descriptive Statistics for Overall Writing Quality in Task 1*

	Draft 1		Draft 2		Draft 3	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
ME	9.60	8.83	15.35	12.06	19.83	16.19
SE	0.40	0.43	0.38	0.48	0.39	0.45
95% CI U	10.41	9.70	16.11	13.04	20.60	17.10
95% CI L	8.78	7.96	14.58	11.06	19.04	15.27
SD	2.33	2.37	2.20	2.70	2.25	2.49
Skewness	0.92	0.40	0.01	-0.12	-1.09	-0.09
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	1.20	0.48	-0.26	0.64	0.32	0.77
SEK	0.78	0.82	0.79	0.42	0.79	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis; Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

Table 40 shows the descriptive statistics for overall writing quality in Task 2. Skewness and kurtosis were acceptable as they were between -1.96 to 1.96.

Table 40. *Descriptive Statistics for Overall Writing Quality in Task 2*

	Draft 1		Draft 2		Draft 3	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
M	11.67	11.15	14.75	14.71	20.91	18.98
SE	0.31	0.50	0.37	0.54	0.38	0.66
95% CI U	12.30	12.18	15.50	15.81	21.70	20.32
95% CI L	11.04	10.12	14.00	13.62	20.13	17.63
SD	1.80	2.81	2.15	2.98	2.23	3.66
Skewness	0.05	0.17	-0.02	-0.77	-0.68	-0.02
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	-0.83	-0.35	0.12	-0.01	-0.28	-0.52
SEK	0.79	0.82	0.79	0.82	0.79	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis; Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

In sum, overall writing quality in both Task 1 and Task 2 were normally distributed. Box plots were examined, and no outliers were found.

Next, Table 41 shows the descriptive statistics for the difference scores of overall writing quality in Task 1. The skewness for the difference score of overall writing quality ranged from -.53 to .91, and kurtosis was between -0.70 to 1.79. Both were acceptable as they were less than the 1.96 criterion.

Table 41. *Descriptive Statistics* for *Overall Writing Quality Difference Scores in Task 1*

	Draft 2 – Draft 1		Draft 3 – Draft 2		Draft 3 - Draft 1	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
M	5.75	3.22	4.48	4.13	10.23	7.36
SE	0.38	5.28	0.37	0.41	0.43	0.49
95% CI U	6.52	4.20	0.52	4.97	11.11	8.35
95% CI L	4.98	2.14	3.72	3.29	9.36	6.36
SD	2.22	2.94	2.17	2.30	2.51	2.71
Skewness	0.91	-0.53	-0.32	0.32	-0.44	-0.21
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	1.79	1.00	0.13	-0.70	-0.50	1.18
SEK	0.79	0.82	0.79	0.82	0.79	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis; Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

Table 42 shows the descriptive scores for the difference scores of overall writing quality in Task 2. Skewness and kurtosis were between -1.96 to 1.96, but Draft 3 for the explicit instruction group had a kurtosis statistic of 2.53. In sum, the data for Task 1 and Task 2 were normally distributed. Box plots were examined, and no outliers were identified.

Table 42. *Descriptive Statistics for Overall Writing Quality Difference Scores in Task 2*

	Draft 2 – Draft 1		Draft 3 – Draft 2		Draft 3 - Draft 1	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
M	3.07	3.56	6.17	4.26	9.25	7.83
SE	2.79	0.42	0.28	0.44	0.35	0.55
95% CI U	3.62	4.42	6.67	5.17	9.97	8.95
95% CI L	2.53	2.70	5.60	3.36	8.53	6.70
SD	1.57	2.34	1.62	2.48	2.06	3.07
Skewness	0.64	1.01	-0.74	0.18	-0.25	-0.05
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	0.45	1.30	0.52	-1.22	2.53	-1.20
SEK	0.79	-0.82	0.79	0.82	0.55	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis; Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

Initial Group Difference Check for Overall Writing Quality

In order to examine whether there were initial differences between the explicit instruction group and the comparison group, one independent samples *t*-test was conducted on the first draft of Task 1. The independent variable was group (two levels: the explicit instruction group and the comparison group). The dependent variable was overall writing quality. The test was not significant, $t(63) = 1.31$, $p = .19$, Cohen's $d = 0.33$.

Main Analysis of Overall Writing Quality for Hypothesis 1

One mixed-design MANOVA was conducted to evaluate the effect of instruction on the first, second, and third drafts of overall writing quality. The independent variable was Group (two levels: explicit instruction and comparison groups). The dependent variable was overall writing quality of two tasks: the comparison and contrast essay and the problem-solution essay. The within-subject factor was Draft with three levels, the first, second, and third drafts.

Box's test was not significant, $F(21, 14325.62) = 1.54$, $p = .05$. Multivariate tests of between-subjects effects were significantly different, $\Lambda = .02$, $F(2, 62) = 2102.61$, $p < .001$, $\eta^2 = .33$. Multivariate tests of within-subjects effects for Draft were significant, $\Lambda = .03$, $F(4, 60) = 427.78$, $p < .001$, $\eta^2 = .97$. The multivariate test of within-subjects effects for Draft and Group was significant, $\Lambda = .62$, $F(4, 60) = 9.26$, $p < .001$, $\eta^2 = .38$.

The univariate test for Draft showed that overall writing quality for Task 1 was significant, $F(2, 126) = 406.85$, $p < .001$, $\eta^2 = .87$. Pairwise comparisons showed that the first and second drafts, first and third drafts, and second and third drafts were significantly different at $p < .001$. Estimated marginal means showed the third draft ($M = 18.00$, $SE = .29$) was the highest, followed by the second draft ($M = 13.70$, $SE = .30$), and the first draft ($M = 9.21$, $SE = .29$) was the lowest. The explicit instruction group ($M = 14.92$, $SE = .33$) scored higher than the comparison group ($M = 12.36$, $SE = .34$) (See Figure 7).

Univariate tests for Draft showed that overall writing quality for Task 2 was significant, $F(2, 126) = 483.76$, $p < .001$, $\eta^2 = .89$. Pairwise comparisons showed that the first and second drafts, the first and third drafts, and the second and third drafts were

significantly different at $p < .001$. Estimated marginal means showed that the third draft ($M = 19.95$, $SE = .37$) was the highest, followed by the second draft ($M = 14.73$, $SE = .32$), and the first draft ($M = 11.41$, $SE = .29$) was the lowest. The explicit instruction group ($M = 15.78$, $SE = .40$) scored higher than the comparison group ($M = 14.95$, $SE = .42$). Moreover, overall writing quality for Task 2 ($M = 15.36$, $SE = .29$) was higher than that of Task 1 ($M = 13.64$, $SE = .24$).

Univariate tests for Draft and Group showed that overall writing quality for Task 1 was significant, $F(2, 126) = 12.96$, $p < .001$, $\eta^2 = .17$. An examination of graphs showed that the distance between the two groups became larger on the second and third drafts than on the first draft (See Figure 7). This result might have caused the interaction effects, indicating that the improvement for the explicit instruction group was large.

Univariate tests for Draft and Group showed that overall writing quality for Task 2 was significant, $F(2, 126) = 6.40$, $p < .05/2 = .03$, $\eta^2 = .09$. An examination of the graphs based on estimated marginal means and means showed that the distance between the two groups on the third drafts became larger compared with the first and second drafts, indicating that the improvement of the explicit instruction group became larger (See Figure 8).

Levene's test of equality of error variances was not significant except for overall writing quality for the third draft of Task 2, $F(1, 63) = 7.58$, $p < .01$. Tests of between subjects main effects for Group showed that overall writing quality in Task 1 was significant, $F(1, 63) = 29.24$, $p < .001$, $\eta^2 = .32$. Estimated marginal means showed that the explicit instruction group ($M = 14.92$, $SE = .33$) scored higher than the comparison group ($M = 12.36$, $SE = .34$). Tests of between subjects main effects for Group showed

that overall writing quality for Task 2 was not significant, $F(1, 63) = 2.06$, $p = .16$. Estimated marginal means showed that the explicit instruction group ($M = 15.78$, $SE = .40$) and the comparison group ($M = 14.95$, $SE = .42$) performed similarly on overall writing quality in Task 2.

Figure 7. *Overall Writing Quality for Task 1*

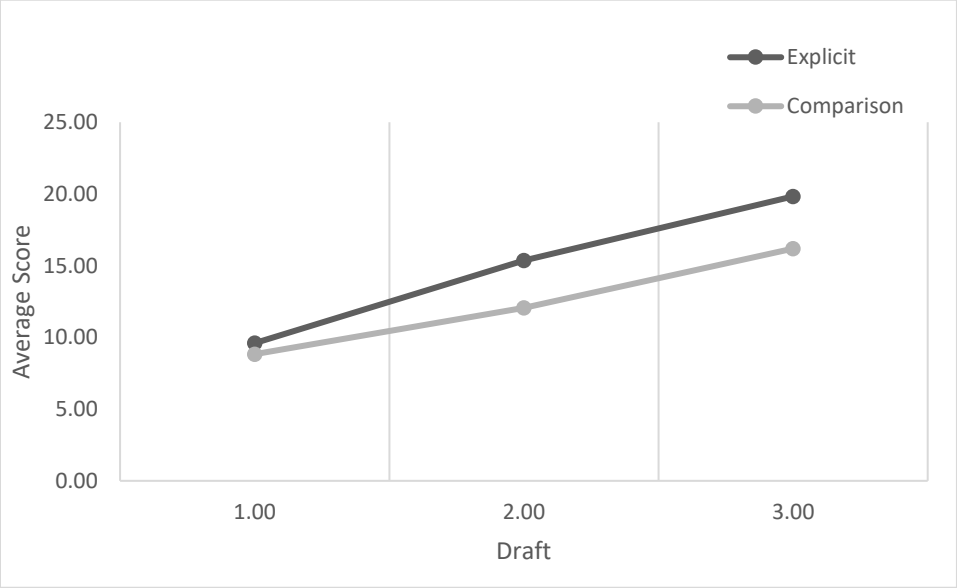
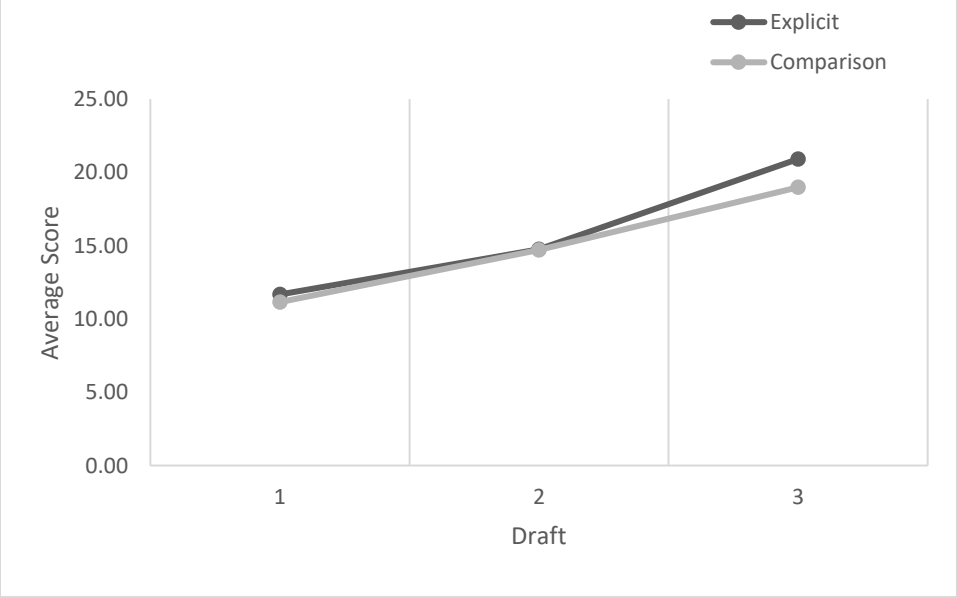


Figure 8. *Overall Writing Quality for Task 2*



Summary of the Results for Hypothesis 1

In summary, significant Draft differences were found. Learners' drafts were significantly different in both groups for both tasks. The third drafts were scored the highest, followed by the second drafts, and then the first drafts were the lowest. The explicit instruction group scored significantly higher than the comparison group on overall writing quality on both tasks. Scores for Task 2 were higher than those of Task 1. The graphs show that the differences between the two groups became larger (See Figures 7 and 8); thus, there were intervention effects.

Hypothesis 2: The Development of Lexical Complexity

The second hypothesis stated that the explicit instruction group would show greater improvements compared with the comparison group on the following dependent variables: (a) MATTR, (b) COCAFrq, and (c) Trigrams. For each analysis, the assumptions of the MANOVA and the normality of each measure were examined.

Univariate Normal Assumptions and Descriptive Statistics

The independent variable consisted of two independent groups—the explicit instruction group and the comparison group. Each participant was a member of only one of the two groups. Sample sizes were more than 30, and two groups were approximately equal in size. Tables 43 to 48 show the descriptive statistics for each variable. Table 43 shows the descriptive statistics for MATTR. Skewness and kurtosis were met the -1.96 to 1.96 criterion.

Table 43. *Descriptive Statistics for MATTR in Task 1*

	Draft 1		Draft 2		Draft 3	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
M	0.65	0.65	0.73	0.67	0.74	0.70
SE	0.01	0.02	0.11	0.01	0.00	0.01
95% CI U	0.67	0.68	0.76	0.73	0.76	0.72
95% CI L	0.62	0.62	0.71	0.67	0.76	0.68
SD	0.07	0.09	0.07	0.08	0.06	0.06
Skewness	-0.17	-0.02	0.19	-0.29	0.26	0.39
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	-0.36	0.46	-1.09	-0.41	-0.53	0.15
SEK	0.79	0.82	0.79	0.82	0.79	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis; Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

Table 44 shows the descriptive statistics for COCAFrq. Skewness and kurtosis met the -1.96 to 1.96 criterion except for Draft 2 for the explicit instruction group.

Table 44. *Descriptive Statistics for COCAFrq in Task 1*

	Draft 1		Draft 2		Draft 3	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
ME	2.31	2.33	2.26	2.32	2.22	2.32
SE	0.22	0.02	0.03	0.02	0.02	0.02
95% CI U	2.35	2.37	2.31	2.39	2.27	2.36
95% CI L	2.26	2.29	2.20	2.31	2.17	2.28
SD	0.13	0.10	0.16	0.09	0.14	0.11
Skewness	-0.39	-0.05	-0.56	-0.81	-0.57	-0.19
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	-0.39	-0.07	2.35	1.13	-0.27	0.78
SEK	0.79	0.82	0.02	0.82	0.79	0.42

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis; Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

Table 45 shows the descriptive statistics for Trigrams. The two groups scored similarly on the three drafts. Skewness and kurtosis met the -1.96 to 1.96 criterion. Box plots were examined and no outliers were found. Hence, the lexical indices in Task 1 were normally distributed.

Table 45. *Descriptive Statistics for Trigrams in Task 1*

	Draft 1		Draft 2		Draft 3	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
ME	8.44	8.44	8.64	8.38	8.86	8.38
SE	0.12	0.08	0.09	0.06	0.61	0.06
95% CI U	8.68	8.60	8.82	8.50	8.79	8.51
95% CI L	8.21	8.27	8.47	8.25	8.54	8.25
SD	0.68	0.45	0.50	0.34	0.35	0.34
Skewness	-0.24	0.32	0.24	-0.15	-0.02	-0.64
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	-0.53	-0.20	-0.05	-0.81	0.68	-0.14
SEK	0.79	0.82	0.79	0.82	0.79	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis; Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

Table 46 shows the descriptive statistics for MATTR in Task 2. The two groups scored similarly on the three drafts. The skewness and kurtosis statistics met the -1.96 to 1.96 criterion.

Table 46. *Descriptive Statistics for MATTR in Task 2*

	Draft 1		Draft 2		Draft 3	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
M	0.72	0.71	0.73	0.73	0.74	0.74
SE	0.01	0.02	0.01	0.02	0.01	0.02
95% CI U	0.74	0.75	0.75	0.76	0.76	0.77
95% CI L	0.70	0.67	0.71	0.69	0.72	0.71
SD	0.07	0.11	0.05	0.10	0.06	0.08
Skewness	-0.53	-0.35	-0.27	-0.87	-0.04	-1.35
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	-0.44	-0.50	-1.00	-0.08	-0.30	1.37
SEK	0.79	0.82	0.79	0.82	0.79	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis; Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

Table 47 shows the descriptive statistics for COCAFrq. The two groups scored similarly on the three drafts.

Table 47. *Descriptive Statistics for COCAFrq in Task 2*

	Draft 1		Draft 2		Draft 3	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
M	2.18	2.21	2.18	2.22	2.15	2.22
SE	0.18	0.02	0.02	0.02	0.02	0.02
95% CI U	2.22	2.26	2.22	2.26	2.19	2.26
95% CI L	2.14	2.16	2.14	2.18	2.12	2.18
SD	0.11	0.13	0.12	0.11	0.10	0.11
Skewness	0.31	0.43	0.18	0.42	0.33	0.25
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	-0.88	-0.07	-0.98	0.14	-0.71	-0.20
SEK	0.79	0.82	0.79	0.82	0.79	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis; Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

Table 48 shows the descriptive statistics for Trigrams. The skewness and kurtosis statistics met the -1.96 to 1.96 criterion. Box plots were examined and no outliers were identified. Hence, the lexical indices in Task 2 were normally distributed.

Table 48. *Descriptive Statistics for Trigrams in Task 2*

	Draft 1		Draft 2		Draft 3	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
M	8.41	8.46	8.46	8.48	8.44	8.54
SE	0.07	0.10	0.07	0.08	0.05	0.07
95% CI U	8.56	8.66	8.60	8.66	8.55	8.67
95% CI L	8.26	8.26	8.33	8.48	8.34	8.41
SD	0.43	0.54	0.39	0.48	0.30	0.36
Skewness	-0.01	-0.21	-0.56	-0.19	0.44	0.30
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	0.95	1.72	0.79	-0.41	-0.26	-1.12
SEK	0.79	0.82	0.79	0.82	0.79	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis; Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

Tables 49 to 54 show the descriptive statistics for the difference scores. Table 49 shows the descriptive statistics for MATTR in Task 1 based on difference scores. The means for the two groups were similar for all three drafts.

Table 49. *Descriptive Statistics for MATTR Difference Scores in Task 1*

	Draft 2–Draft 1		Draft 3–Draft 2		Draft 3–Draft 1	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
M	0.09	0.05	0.01	0.00	0.09	0.05
SE	0.01	0.01	0.01	0.01	0.01	0.01
95% CI U	0.12	0.07	0.02	0.18	0.12	0.08
95% CI L	0.06	0.03	-0.01	-0.02	0.06	0.02
SD	0.08	0.06	0.03	0.05	0.07	0.07
Skewness	0.13	0.27	0.59	0.26	0.02	-0.18
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	-0.54	-0.92	-0.17	2.47	-0.83	1.09
SEK	0.79	0.82	0.78	0.82	0.79	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis; Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

Table 50 shows the descriptive statistics for COCAFrq in Task 1 based on the difference scores.

Table 50. *Descriptive Statistics for COCAFrq Difference Scores in Task 1*

	Draft 2–Draft 1		Draft 3–Draft 2		Draft 3–Draft 1	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
M	-0.05	0.02	-0.04	-0.03	-0.09	-0.01
SE	0.03	0.02	0.01	0.01	0.03	0.02
95% CI U	0.01	0.05	-0.02	-0.01	-0.04	0.03
95% CI L	-0.10	-0.01	-0.06	-0.05	-0.14	-0.05
SD	0.16	0.01	0.06	0.05	0.15	0.10
Skewness	-0.08	-0.29	-0.57	0.43	0.08	-0.53
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	0.02	-0.79	0.69	1.60	1.26	0.40
SEK	0.02	0.82	0.79	0.82	0.79	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis. Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

Table 51 shows the descriptive statistics for Trigrams in Task 1 based on difference scores. The means for the explicit instruction group were considerably higher

than those for the comparison group. The skewness and kurtosis statistics met the -1.96 to 1.96 criterion.

Table 51. *Descriptive Statistics for Trigrams Difference Scores in Task 1*

	Draft 2–Draft 1		Draft 3–Draft 2		Draft 3–Draft 1	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
M	0.29	0.06	0.23	0.00	0.22	-0.06
SE	0.13	0.07	0.14	0.05	0.12	0.08
95% CI U	0.47	0.08	-0.10	0.11	0.46	0.11
95% CI L	-0.08	-1.98	0.34	-0.10	-0.02	-0.23
SD	0.79	0.38	-0.09	0.28	0.68	0.46
Skewness	-0.25	0.08	-0.09	0.79	-0.03	0.83
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	-0.06	-1.03	-0.80	1.93	-0.96	1.74
SEK	0.07	0.82	0.79	0.82	0.79	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis; Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

The results of the analyses for the lexical indices for Task 2 are displayed in the following Tables. Table 52 shows the descriptive statistics for MATTR for Task 2. The means for both groups were very similar for all three drafts. The skewness and kurtosis statistics met the -1.96 to 1.96 criterion except for Draft 3 for the explicit instruction group.

Table 52. *Descriptive Statistics for MATTR Difference Scores* in Task 2

	Draft 2–Draft 1		Draft 3–Draft 2		Draft 3–Draft 1	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
M	0.01	0.02	0.01	-0.01	0.02	0.03
SE	0.01	0.01	0.01	0.01	0.01	1.00
95% CI U	0.03	0.03	0.02	0.02	0.04	0.05
95% CI L	-0.01	0.00	0.00	0.00	0.00	0.03
SD	0.05	0.05	0.03	0.03	0.06	0.06
Skewness	0.35	-0.33	1.44	0.70	0.57	-0.21
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	0.50	1.81	4.00	0.71	-0.02	-0.16
SEK	0.79	0.82	0.79	0.82	0.79	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis. Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

Table 53 shows the descriptive statistics for COCAFrq based on difference scores in Task 1. The means for both groups were very similar for all three drafts. The skewness and kurtosis statistics met the -1.96 to 1.96 criterion.

Table 53. *Descriptive Statistics for COCAFrq Difference Scores in Task 2*

	Draft 2–Draft 1		Draft 3–Draft 2		Draft 3–Draft 1	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
M	0.00	0.01	0.03	0.00	-0.02	0.01
SE	0.01	0.01	0.01	0.01	0.01	0.01
95% CI U	0.03	0.04	-0.11	0.01	0.01	0.03
95% CI L	-0.02	-0.01	-0.04	-0.02	-0.05	-0.01
SD	0.07	0.06	0.04	0.03	0.08	0.06
Skewness	-0.22	0.78	-0.02	-0.35	0.22	0.53
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	0.06	0.59	-0.03	0.67	-0.28	0.13
SEK	0.79	0.82	0.79	0.82	0.79	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis; Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

Table 54 shows the descriptive statistics for Trigrams based on difference scores in Task 2. The means for both groups were similar for all three drafts. The skewness and

kurtosis statistics met the -1.96 to 1.96 criterion except for Draft 2 to Draft 3 for both groups.

Table 54. *Descriptive Statistics for Trigrams Difference Scores in Task 2*

	Draft 2–Draft 1		Draft 3–Draft 2		Draft 3–Draft 1	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
M	0.05	0.02	-0.02	0.06	0.03	0.08
SE	0.05	0.06	0.05	0.06	0.06	0.09
95% CI U	0.15	0.15	0.09	0.18	0.16	0.27
95% CI L	-0.05	-0.10	-0.13	-0.06	-0.10	-0.10
SD	0.30	0.33	0.31	0.33	0.37	0.51
Skewness	0.03	-0.92	-0.51	1.51	0.02	-0.06
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	1.24	2.31	5.02	5.33	0.28	4.73
SEK	0.79	0.82	0.79	0.82	0.09	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis; Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

Main Analysis for Hypothesis 2 for Task 1

Two mixed-design MANOVAs were conducted to evaluate whether there were any differences between the participants in the explicit instruction group and the participants in the comparison group in the number of changes in their scores on MATTR, COCAFrq, and Trigram over time. Task 1 (a comparison-contrast essay) and Task 2 (a problem-solution essay) were analyzed separately. First, a mixed design MANOVA with one within-subjects' factor and one between factor was conducted for Task 1. The independent variable was Group (two levels: explicit instruction and comparison groups). The dependent variables were MATTR, COCAFrq, and Trigrams in Task 1. The within-subject factors were Draft with three levels: the first, second, and third drafts.

The sample sizes were equal between two groups; therefore, the assumptions were considered to be met.

Box's test of quality of covariance matrices was significant, $F(45, 12801.86) = 1.85$, $p < .001$, so Pillai's Trace tests are reported.

Multivariate tests of between-subjects effects Pillai's Trace tests showed that the main effect of Group was significant, $F(3, 61) = 3.06$, $p < .05$, $\eta^2 = .13$. Within-subjects effects for multivariate Pillai's Trace showed that the main effect of Draft was significant, $F(6, 58) = 16.72$, $p < .001$, $\eta^2 = .63$. Within-subject effects for multivariate Pillai's Trace showed that the interaction between Draft and Group was not significant, $F(6, 58) = 1.97$, $p = .08$.

The results of the univariate tests for Draft showed that MATTR was significant, Epsilon Huynh-Feldt (ε̂), $F(1.60, 100.87) = 49.91$, $p < .001$, $\eta^2 = .44$. Tests of within-subjects contrasts showed that the first and second drafts were significantly different, $F(1, 63) = 59.31$, $p < .001$, $\eta^2 = .49$, but the second and third drafts did not differ significantly, $F(1, 63) = .04$, $p = .85$. Estimated marginal means showed that the third drafts ($M = .72$, $SE = .01$) and the second drafts ($M = .72$, $SE = .01$) scored similarly, and they scored higher than the first drafts ($M = .65$, $SE = .01$). Next, COCAFrq changed significantly over time, Epsilon Huynh-Feldt (ε̂), $F(1.30, 82.06) = 6.55$, $p < .01$, $\eta^2 = .09$. Tests of within-subjects contrasts showed that the first and second drafts were not significantly different, $F(1, 63) = .74$, $p = .39$, but the second and third drafts did differ significantly, $F(1, 63) = 26.23$, $p < .001$, $\eta^2 = .29$. Learners produced fewer low-frequency words between the second and third drafts. Estimated marginal means showed that the first draft ($M = 2.32$, $SE = .02$) was higher than the second draft ($M = 2.31$, $SE = .02$), followed by

the third draft (M = 2.27, SE = .02). Trigrams were not significantly different, Epsilon Huynh-Feldt (ε̂), F(1.45, 91.24) = .90, p = .38. Estimated marginal means showed that the first draft (M = 8.44, SE = .07), second draft (M = .8.51, SE = .05), and third draft (M = 8.52, SE = 8.52) means were similar.

Levene's tests of the equality of error variances were not significant except for the second draft of COCAFrq in Task 1. The results of between-subjects effects for MATTR showed no significant difference, F(1, 63) = 2.65, p = .11. Estimated marginal means showed that the explicit instruction group (M = .71, SE = .01) scored higher than the comparison group (M = .68, SE = .01). Between-subjects effects for COCAFrq was significant, F(1, 63) = 7.66, p < .05. Estimated marginal means showed that the comparison group (M = 2.33, SE = .02) scored slightly higher than the explicit instruction group (M = 2.26, SE = .02). The results of between-subjects effects for Trigrams were significant, F(1, 63) = 4.63, p < .05. Estimated marginal means showed that the explicit instruction group (M = 8.58, SE = .06) scored higher than the comparison group (M = 8.40, SE = .06).

In summary, in Task 1, the results of the multivariate tests showed that Draft and Group were significantly different and the interaction between Group and Draft was not significant. Within-subjects multivariate effects for Draft were found between the first and second drafts of MATTR and between the second and third drafts of COCAFrq. No significant within-subject Draft effects was found for Trigrams (See Figures 9, 10, and 11). Group differences were not found for MATTR but were found for COCAFrq, and Trigrams.

Figure 9. *MATTR for Task 1*

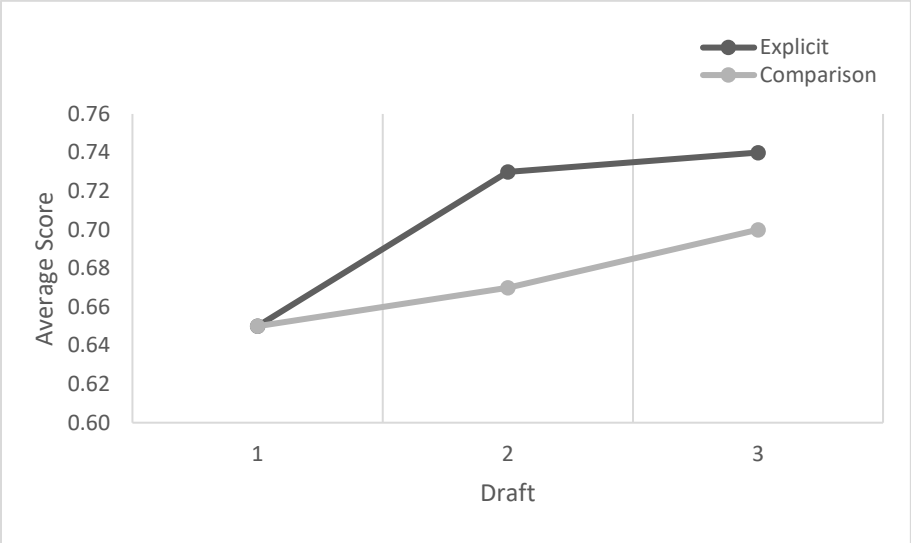


Figure 10. *COCAFrq for Task 1*

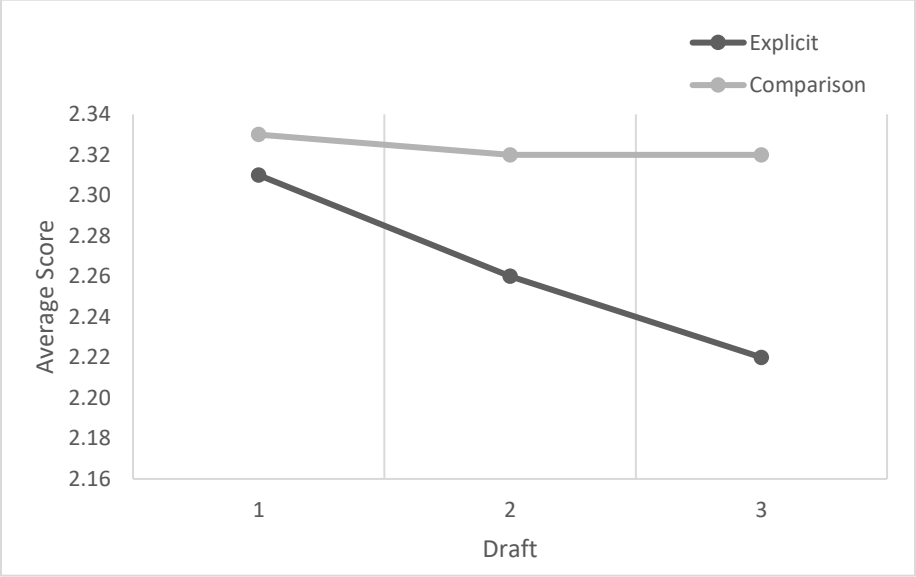
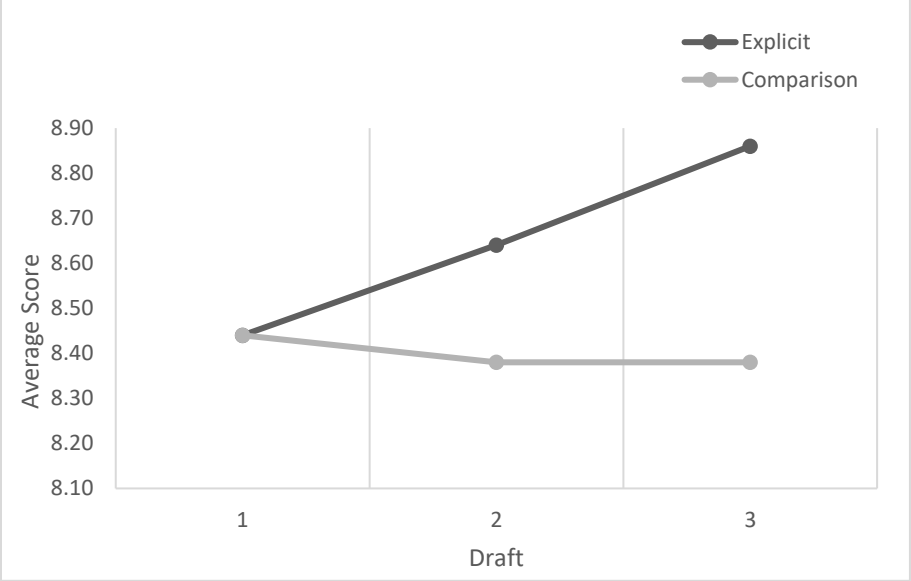


Figure 11. *Trigrams for Task 1*



Main Analysis for Hypothesis 2 for Task 2

One mixed-design MANOVAs with two within-subjects' factors and one between-subjects factor were conducted to determine whether there were any differences among the first, second, and third drafts of the lexical indices between the two groups for Task 2 (a problem-solution essay). The independent variable was Group (two levels: the explicit instruction group and the comparison group). The dependent variables were MATTR, COCAFrq, and Trigrams in Task 2. The within-subjects factor was Draft with three levels: the first, second, and third drafts.

Box's test of quality of covariance matrices was not significant, $F(45, 12801.86) = 1.19$, $p = .18$. Between-subjects effects for multivariate tests showed that Group was not significantly different, $\Lambda = .94$, $F(3, 61) = 1.20$, $p = .06$. Multivariate tests of within-subjects effects showed that Draft was significant, $\Lambda = .72$, $F(6, 58) = 3.71$, $p < .01$, $\eta^2 = .32$. Within-subject effects for multivariate showed that the interaction between Draft and Group was not significant, $\Lambda = .89$, $F(6, 58) = 1.26$, $p = .12$.

Univariate tests for Draft showed that MATTR was significant, $F(2, 126) = 6.81$, $p < .05/3 = .02$, $\eta^2 = .10$. Estimated marginal means showed that the third draft ($M = .74$, $SE = .01$) was higher than the second draft ($M = .73$, $SE = .01$), followed by the first draft ($M = .72$, $SE = .01$). Next, univariate tests showed that COCAFrq was not significant, $F(2, 126) = 2.03$, $p = .14$, nor were Trigrams, $F(2, 126) = .81$, $p = .45$.

Levene's tests of equality of error variances were not significant except for the first and second drafts of MATTR. Between-subjects effects showed that MATTR was not significantly different between groups, $F(1, 63) = .06$, $p = .11$. Estimated marginal means showed that the explicit instruction group ($M = .73$, $SE = .01$) and the comparison group ($M = .73$, $SE = .01$) scored similarly. Between-subjects effects showed that COCAFrq was not significant, $F(1, 63) = 2.01$, $p = .09$. Estimated marginal means showed that the explicit instruction group ($M = .2.17$, $SE = .01$) scored lower than the comparison group ($M = 2.22$, $SE = .02$). Between-subjects effects for Trigrams were not significantly different, $F(1, 63) = .35$, $p = .56$. Estimated marginal means showed that the explicit instruction group ($M = 8.44$, $SE = .06$) scored slightly lower than the comparison group ($M = 8.50$, $SE = .07$).

In summary, in Task 2, the multivariate tests showed that there were no group differences, but there were significant within-subject Draft differences. Among the lexical indices, MATTR was significantly different over time. No significant differences were found for COCAFrq and Trigrams (See Figure 12, 13, and 14). Within-subjects Draft and Group were not significantly different.

Figure 12. *MATTR for Task 2*

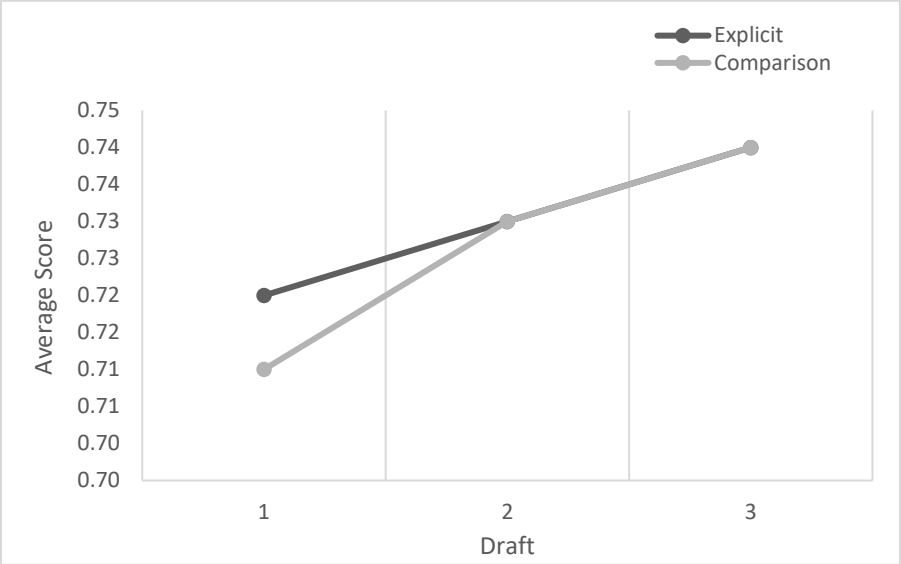


Figure 13. *COCAFrq for Task 2*

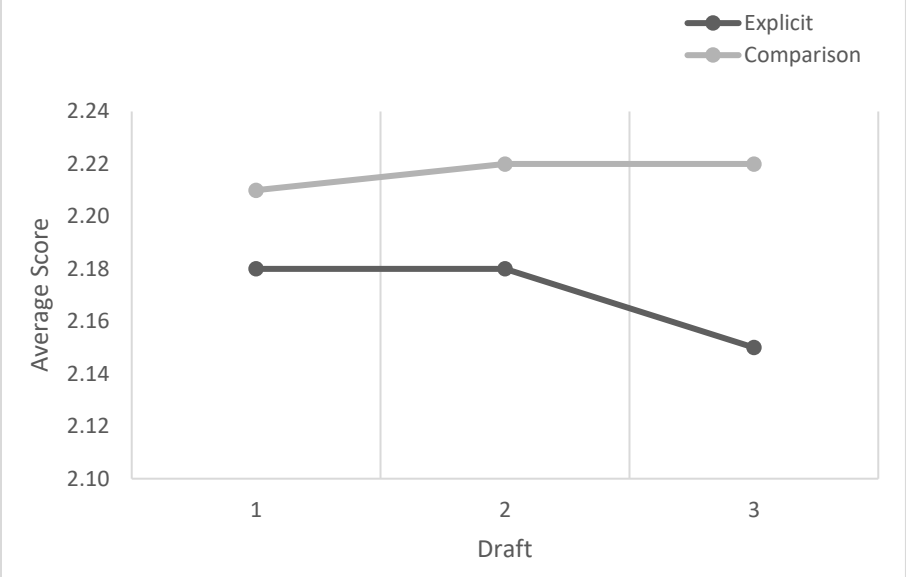
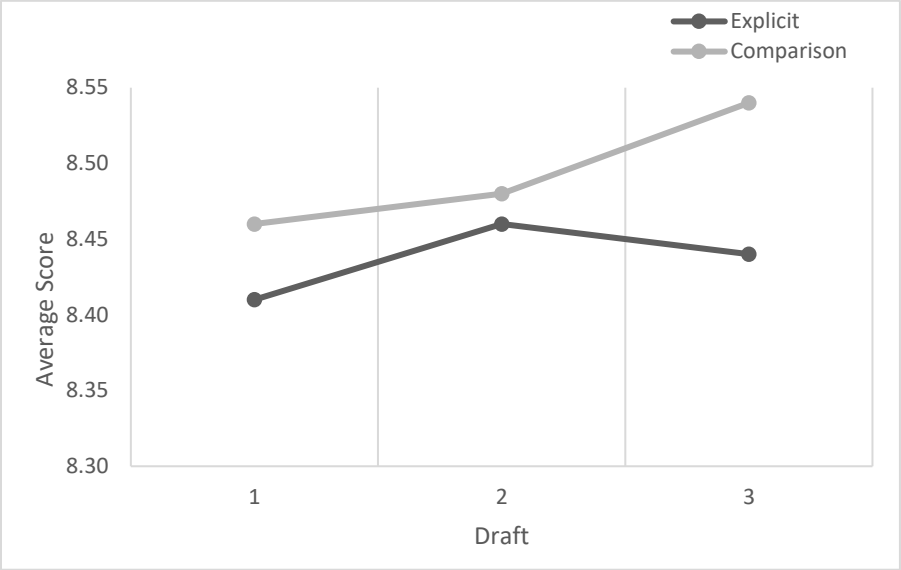


Figure 14. *Trigrams for Task 2*



Summary of the Results for Hypothesis 2

In summary, group differences were found for Task 1, but not for Task 2. Draft was significantly different in both Task 1 and 2. Within-subjects multivariate effects for Draft were found for MATTR in both tasks. A significant Draft effect was found for COCAFrq in Task 1, but not in Task 2. A significant Draft effect was not found for Trigrams. More Draft effects were found in Task 1 than Task 2. As for between-subjects multivariate tests, significant group differences were not found for MATTR in both tasks, but differences were found for COCAFrq and Trigrams in Task 1. The comparison group scored higher than the explicit instruction group for COCAFrq, and the explicit group scored higher than the comparison group for Trigrams in Task 1. No group differences were found in Task 2. More effects were found in Task 1 than Task 2.

Hypothesis 3: The Development of Cohesion

Hypothesis 3 predicted that there would be significant differences in the cohesion measures given different types of instruction over the semester and that the explicit instruction group would exceed that of the comparison group. The dependent variables were local and global cohesion. Local cohesion was measured using All connectives and global cohesion was measured using AdjacentP.

Univariate Normal Assumptions and Descriptive Statistics

The independent variable consisted of two independent groups, the explicit instruction group and the comparison group. Each participant was a member of one of the two groups, so the independence of observations was met. The sample size was more than 30 for each group and both groups were approximately equal in size. Table 55 shows the descriptive statistics for All connectives. The means were extremely similar for the two groups. The skewness and kurtosis statistics met the -1.96 to 1.96 criterion.

Table 55. *Descriptive Statistics for All Connectives in Task 1*

	Draft 1		Draft 2		Draft 3	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
M	0.07	0.07	0.08	0.07	0.08	0.07
SE	0.00	0.00	0.00	0.00	0.00	0.00
95% CI U	0.07	0.08	0.08	0.06	0.08	0.08
95% CI L	0.06	0.06	0.07	0.06	0.07	0.07
SD	0.02	0.03	0.01	0.02	0.01	0.02
Skewness	0.62	0.78	-0.19	0.17	0.34	0.18
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	-0.37	1.01	0.01	0.00	-0.34	-0.63
SEK	0.79	0.82	0.79	0.82	0.79	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis; Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

Table 56 shows the descriptive statistics for AdjacentP. The explicit instruction group scored consistently higher than the comparison group. The skewness and kurtosis statistics met the -1.96 to 1.96 criterion.

Table 56. *Descriptive Statistics for AdjacentP in Task 1*

	Draft 1		Draft 2		Draft 3	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
M	4.80	3.20	6.81	5.90	8.44	6.50
SE	0.50	0.54	0.26	0.33	0.31	0.30
95% CI U	5.81	4.30	7.35	6.59	9.07	7.10
95% CI L	3.78	2.10	6.28	5.22	7.82	5.90
SD	2.91	3.00	1.53	1.86	1.79	1.64
Skewness	-0.04	0.30	-0.51	-0.10	0.46	0.66
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	0.07	-1.25	-0.30	-0.61	1.25	0.77
SEK	0.79	0.82	0.79	0.82	0.79	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis; Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

Table 57 shows the descriptive statistics for All connectives for Task 2. The means were extremely similar for the two groups.

Table 57. *Descriptive Statistics for All Connectives in Task 2*

	Draft 1		Draft 2		Draft 3	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
ME	0.04	0.05	0.04	0.05	0.04	0.05
SE	0.00	0.00	0.00	0.00	0.00	0.00
95% CI U	0.04	0.06	0.04	0.06	0.05	0.05
95% CI L	0.03	0.04	0.03	0.04	0.04	0.04
SD	0.02	0.02	0.01	0.02	0.02	0.02
Skewness	0.14	0.17	-0.13	0.15	-0.03	0.14
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	-0.50	-0.15	-1.05	-0.15	-0.25	-0.11
SEK	0.79	0.82	0.79	0.82	0.79	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis; Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

Table 58 shows the descriptive statistics for AdjacentP. Box plots were examined for Task 1 and Task 2 and no outliers were found. Both tasks were normally distributed.

Table 58. *Descriptive Statistics for AdjacentP in Task 2*

	Draft 1		Draft 2		Draft 3	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
ME	5.80	0.48	6.24	5.50	7.90	7.20
SE	0.31	0.51	0.33	0.49	0.33	0.48
95% CI U	5.70	5.79	6.91	6.50	8.58	8.18
95% CI L	4.46	3.71	5.57	4.50	7.22	6.23
SD	1.79	2.85	1.93	2.73	1.94	2.66
Skewness	0.73	0.07	1.25	-0.21	0.05	0.05
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	0.29	-0.66	3.28	-0.29	-0.14	-0.80
SEK	0.79	0.82	0.79	0.82	0.79	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; SES = Standard error of skewness; SEK = Standard error of kurtosis; Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

Table 59 shows the descriptive statistics for All connectives in Task 1 based on difference scores. Skewness and kurtosis were between -1.96 to 1.96.

Table 59. *Descriptive Statistics for All Connectives Difference Scores in Task 1*

	Draft 2–Draft 1		Draft 3–Draft 2		Draft 3–Draft 1	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
M	0.01	0.00	0.00	0.00	0.01	0.00
SE	0.00	0.00	0.00	0.00	0.00	0.00
95% CI U	0.02	0.10	0.00	0.00	0.02	0.01
95% CI L	0.00	0.00	0.00	0.00	0.00	0.00
SD	0.02	0.02	0.01	0.01	0.02	0.02
Skewness	-0.52	0.12	-0.04	-0.42	-0.44	0.04
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	-0.43	-0.68	-0.73	-0.05	-1.03	-0.95
SEK	0.79	0.82	0.79	0.82	0.79	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis; Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

Table 60 shows the descriptive statistics for AdjacentP in Task 1 based on difference scores. The means for the two groups fluctuated considerably but ultimately increased on Draft 3.

Table 60. *Descriptive Statistics for AdjacentP Difference Scores in Task 1*

	Draft 2–Draft 1		Draft 3–Draft 2		Draft 3–Draft 1	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
M	0.20	2.69	1.63	0.60	3.64	3.30
SE	0.56	0.56	0.30	0.26	0.52	4.80
95% CI U	3.15	3.83	2.23	1.13	4.71	4.27
95% CI L	0.89	1.55	1.03	0.08	2.58	2.32
SD	3.24	3.11	1.73	1.43	3.05	2.66
Skewness	-0.27	0.41	0.75	-0.98	-0.45	0.02
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	0.04	-0.40	0.79	2.18	0.23	-0.71
SEK	0.79	0.82	0.79	0.82	0.79	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis; Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

Table 61 shows the descriptive statistics for All connectives in Task 2 based on difference scores.

Table 61. *Descriptive Statistics for All Connectives Difference Scores in Task 2*

	Draft 2–Draft 1		Draft 3–Draft 2		Draft 3–Draft 1	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
M	0.00	0.00	0.01	-0.78	0.01	0.00
SE	0.00	0.00	0.00	0.00	0.00	0.00
95% CI U	0.00	0.00	0.01	0.00	0.01	0.00
95% CI L	0.00	-0.01	0.00	0.00	0.00	-0.01
SD	0.01	0.01	0.01	0.01	0.02	0.01
Skewness	0.06	-2.79	0.37	-0.34	0.15	-0.60
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	0.41	11.27	-0.78	-0.12	-1.01	0.07
SEK	0.79	0.82	0.79	0.82	0.79	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis; Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

Table 62 shows the descriptive statistics for AdjacentP in Task 2 based on difference scores. Skewness and kurtosis were between -1.96 to 1.96.

Table 62. *Descriptive Statistics for AdjacentP Difference Scores in Task 2*

	Draft 2–Draft 1		Draft 3–Draft 2		Draft 3–Draft 1	
	Explicit	Comparison	Explicit	Comparison	Explicit	Comparison
M	1.16	0.75	1.66	1.70	2.82	0.25
SE	0.37	0.33	0.25	0.37	0.40	0.46
95% CI U	1.91	1.42	2.16	2.45	3.62	3.39
95% CI L	0.41	0.01	1.15	0.95	2.02	1.51
SD	2.14	1.83	1.46	2.05	2.31	2.57
Skewness	-0.13	-0.39	-0.13	0.70	-0.14	0.03
SES	0.40	0.42	0.40	0.42	0.40	0.42
Kurtosis	1.53	1.32	0.01	0.67	0.37	-0.87
SEK	0.79	0.82	0.79	0.82	0.79	0.82

Note. CI U = Confidence interval upper limit; CI L = Confidence interval lower limit; *SES* = Standard error of skewness; *SEK* = Standard error of kurtosis; Explicit = The explicit instruction group (*n* = 34); Comparison = The comparison group (*n* = 31).

Main Analysis for Hypothesis 3 for Task 1

Two mixed-design MANOVAs were conducted to determine whether any differences existed among the first, second, and third drafts for the cohesion indices between the two groups for Task 1 (a comparison-contrast essay) and Task 2 (a problem-solution essay). First, a mixed-design MANOVA with one within-subjects' factor was conducted to evaluate the development of cohesion in Task 1. The independent variable was Group (two levels: the explicit instruction group and the comparison group). The dependent variables were All connectives as a measure of local cohesion and AdjacentP as a measure of global cohesion. The within-subject factors were Draft with three levels: the first, second, and third drafts.

Box's test of the equality of covariance matrices was not significant, $F(21,14325.62) = 1.09$, $p = .35$. Between-subjects effects for multivariate tests showed

that Group was significantly different, $\Lambda = .77$, $F(2, 62) = 9.08$, $p < .001$, $\eta^2 = .23$. Within-subjects effects for multivariate tests showed that Draft was significant, $\Lambda = .32$, $F(4, 60) = 31.87$, $p < .001$, $\eta^2 = .68$. Within-subject effects for multivariate tests showed that the interaction between Draft and Group was not significant, $\Lambda = .87$, $F(4, 60) = 2.24$, $p = .08$.

Univariate tests for Draft showed that All connectives were significant, $F(2, 126) = 6.29$, $p < .003$ (.05/2), $\eta^2 = .09$. Tests of within-subjects contrasts showed that the first and second draft of All connectives differed significantly, $F(1, 63) = 8.02$, $p < .001$, $\eta^2 = .11$. The second and third drafts of All connectives were not significantly different, $F(1, 63) = .02$, $p = .90$ (See Figures 15). Estimated marginal means showed that the first draft ($M = .07$, $SE = .00$) was slightly lower than the second draft ($M = .07$, $SE = .00$) and the third draft ($M = .07$, $SE = .00$). AdjacentP was significant over time, $F(2, 126) = 58.61$, $p < .001$, $\eta^2 = .48$. Tests of within-subjects contrasts showed that the first and second drafts differed significantly, $F(1, 63) = 35.65$, $p < .001$, $\eta^2 = .36$, as did the second and third drafts, $F(1, 63) = 31.84$, $p < .001$, $\eta^2 = .34$ (See Figures 16). Tests of within-subjects multivariate effects showed that Time and Group for All connectives were not significant, $F(2, 126) = 1.46$, $p = .24$, as was Draft and Group for AdjacentP, $F(2, 126) = 1.27$, $p = .28$.

Levene's tests of equality of error variances were not significant. Tests of between-subjects effects for All connectives were not significant, $F(1, 63) = 1.48$, $p = .23$. Tests of between-subjects effects for AdjacentP was significant, $F(1, 63) = 14.08$, $p < .001$, $\eta^2 = .18$. the univariate test of the contrasts for AdjacentP was significant, $F(1, 63) = 13.08$, $p < .001$, $\eta^2 = .23$. Estimated marginal means showed that the explicit

instruction group (M = 6.68, SE = .27) scored higher than the comparison group (M = 5.20, SE = .29).

In summary, in Task 1, there were significant group differences and Draft differences. No interaction between Group and Draft was found. Within-subjects Draft effects showed that All connectives changed significantly from the first to the second drafts, but not from the second to the third drafts. Within-subjects Draft effects were found for AdjacentP from the first to second drafts, as well as from the second to third drafts. Learners produced connectives on the second draft, but afterwards they did not use them. On the other hand, AdjacentP was significantly different over the three drafts. There were significant group differences in AdjacentP as the explicit instruction group scored higher than the comparison group. There were no significant differences for All connectives.

Figure 15. *All Connectives in Task 1*

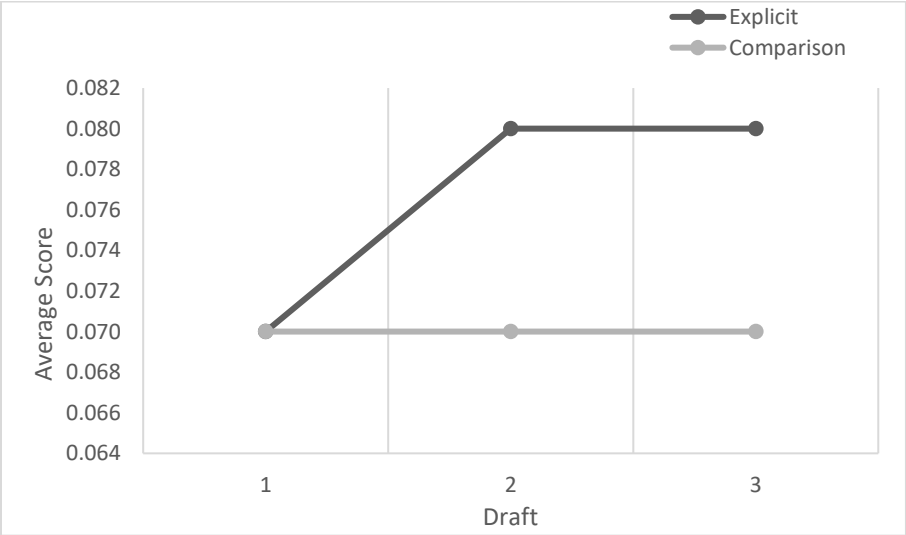
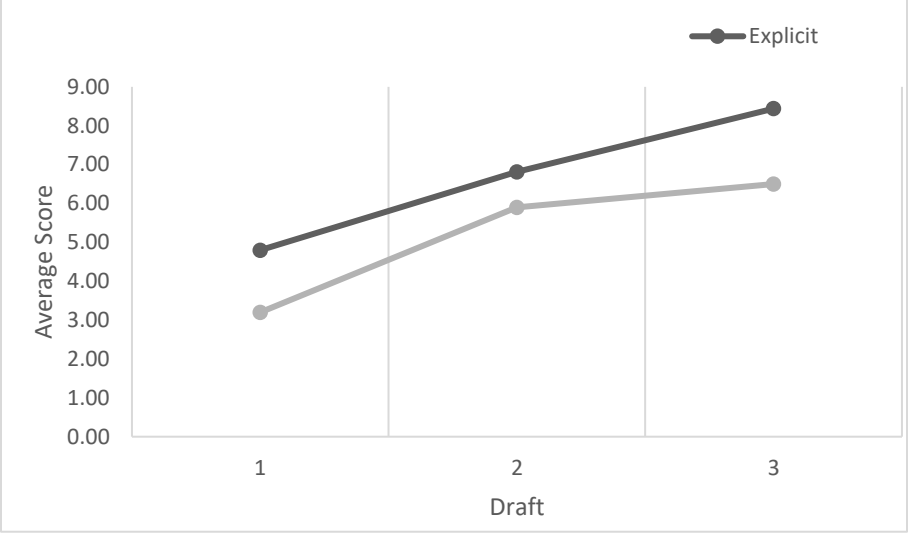


Figure 16. *AdjacentP for Task 1*



Main Analysis for Hypothesis 3 for Task 2

For Task 2, a mixed-design MANOVA was conducted to evaluate whether any differences existed among the first, second, and third drafts for cohesion between the two groups for Task 2 (a problem-solution essay). The independent variable was Group (two levels: the explicit instruction group and the comparison group). The dependent variable was All connectives as a measure of local cohesion and AdjacentP as a measure of global cohesion. The within-subject factor was Draft with three levels: the first, second, and third drafts.

Box's test of quality of covariance matrices was significant, $F(21,14325.62) = 1.64$, $p = .03$. Between-subjects effects for multivariate Pillai's Trace tests showed that Group was not significant, $F(2, 62) = 2.95$, $p = .06$. Within-subjects effects for multivariate Pillai's Trace showed that Draft was significant, $F(4, 60) = 21.86$, $p < .001$, $\eta^2 = .59$. Within-subject effects for multivariate Pillai's Trace showed that the interaction between Draft and Group was also significant, $F(4, 60) = 2.86$, $p < .05$, $\eta^2 = .16$.

Univariate tests for Draft showed that All connectives was not significant, Epsilon Huynh-Feldt (ε̂), $F(1.68, 105.92) = 1.81$, $p = .18$ (See Figure 17). AdjacentP was significantly different, Epsilon Huynh-Feldt (ε̂), $F(1.82, 114.76) = 53.11$, $p < .001$, $\eta^2 = .46$ (See Figure 18). Tests of within-subjects contrasts showed that the first and second drafts differed significantly, $F(1, 63) = 14.81$, $p < .001$, $\eta^2 = .19$. The second and third drafts also differed significantly, $F(1, 63) = 58.62$, $p < .001$, $\eta^2 = .48$.

Univariate tests for Draft and Group showed that All connectives were significant, $F(1.68, 105.92) = 6.44$, $p < .004$. There were larger distances between the two groups for the first and second drafts compared to the third draft. This result might have caused the interaction. Univariate tests for Draft and Group showed that AdjacentP was not significant, $F(1.82, 114.76) = .38$, $p = .67$.

Levene's tests of equality of error variances were not significant. Tests of between-subjects effects for All connectives was not significant, $F(1, 63) = 4.17$, $p = .05$. Tests of between-subjects effects for AdjacentP were not significant, $F(1, 63) = 1.41$, $p = .24$.

Figure 17. *All Connectives for Task 2*

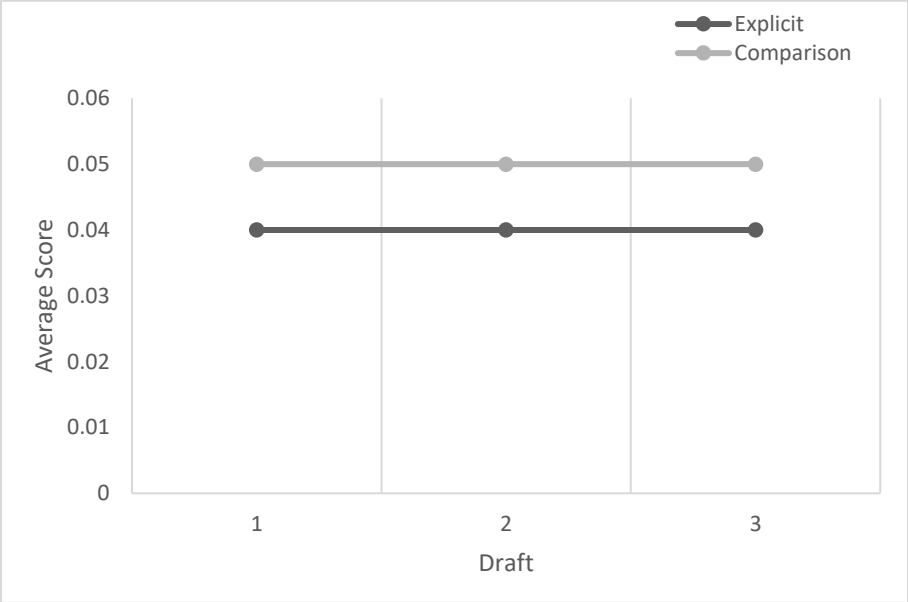
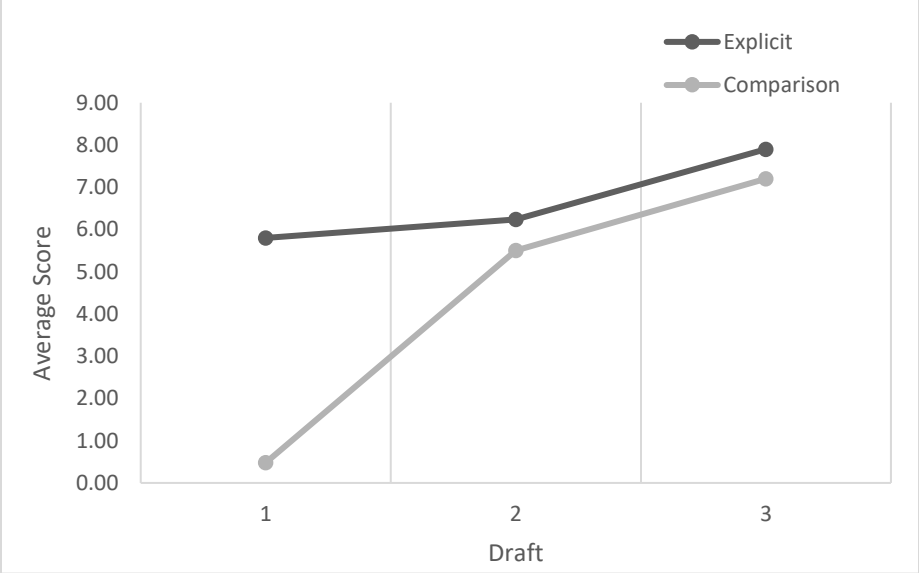


Figure 18. *AdjacentP for Task 2*



In summary, in Task 2, no group differences were found for both All connectives and AdjacentP, but there were Draft differences. Significant differences were not found for All connectives, but were found for AdjacentP. Significant interactions between Group and Draft were found for All connectives, but not for AdjacentP.

Summary of the Results for Hypothesis 3

In summary, there were significant group differences for AdjacentP for Task 1, but not for Task 2. There were significant Draft differences for All connectives in Task 1, but not in Task 2. Significant Draft differences were found for AdjacentP in both tasks. The interaction between Group and Time was not found for Task 1, but was found for Task 2.

Hypothesis 4: The Relationship Between Lexical Complexity and Overall Writing Quality

The fourth hypothesis predicted significant relationships between lexical complexity and overall writing quality for first-year Japanese university EFL students and that the explicit instruction group would exceed that of the comparison group. The third draft was used for the analyses. It was hypothesized that the lexical indices would be correlated with overall writing quality. Pearson correlation coefficients were used to measure the strength of the relationships between the third drafts of overall writing quality and the three lexical indices. The skewness and kurtosis statistics met the criterion (See Tables 43–48). Both groups were approximately equal in size and the dependent variable scores were independently collected, so the assumptions of the analyses were met.

Main Analysis for Hypothesis 4 for Task 1

The results are presented in Table 72 for Task 1. MATTR (M = .72, SE = .06) was significantly and positively correlated with overall writing quality (M = 18.98, SE = 2.98) at r = .40, p < .008 (.05/6). COCAFrq (M = 8.53, SE = .38) was significantly and

negatively correlated with overall writing quality at r = -.37, p < .008. Trigrams were not significantly correlated with overall writing quality at r = .16, p = .20.

The results indicated that when the participants used a variety of words, their essays were rated higher (See Table 63), and when they used low-frequency words, their essays were rated lower. The use of low-frequency words was not appropriate, or their use was judged unnatural by the raters. The use of Trigrams was not related to overall writing quality.

Table 63. *Pearson Correlations Between Lexical Complexity and Overall Writing Quality in Task 1*

	1	2	3	4
1. Overall WQ	—			
2. MATTR	.40	—		
3. COCAFrq	-.37	-.34*	—	
4. Trigrams	.16	.30	-.45*	—

* Correlation is significant at .008 (0.05/6) (2-tailed).

Note. Overall WQ = Overall writing quality.

Main Analysis for Hypothesis 4 for Task 2

The results presented in Table 64 show that none of the correlations were significant at p < .008 (.06/6). MATTR (M = .74, SE = .07) was correlated with overall writing quality at r = .03, p = .80, COCAFrq (M = 2.18, SE = .11) was correlated with overall writing quality at r = -.24, p = .05, and Trigrams (M = 8.49, SE = .33) were correlated with overall writing quality at r = .09, p = .49.

Table 64. *Pearson Correlations Between Lexical Complexity and Overall Writing Quality in Task 2*

	1	2	3	4
1. Overall WQ	—			
2. MATTR	.03	—		
3. COCAFrq	-.24	-.10	—	
4. Trigrams	.09	.01	-.10	—

Note. Overall WQ = Overall writing quality.

Hypothesis 5: The Relationship Between Cohesion and Overall Writing Quality

The fifth hypothesis predicted a significant relationship between cohesion and Overall Writing Quality for first-year university EFL students and that the explicit instruction group would exceed that of the comparison group. It was hypothesized that both the local and global cohesion indices, All connectives and AdjacentP, would correlate significantly with overall writing quality. Pearson correlation coefficients were used to measure the strength of the relationships among the third drafts of All connectives, Adjacent P, and overall writing quality. The skewness and kurtosis statistics met the criterion (See Tables 55–58), and both groups were approximately equal in size. Dependent variable scores were independently collected, so the assumptions of the analysis were met.

Main Analysis for Hypothesis 5 for Task 1

The results presented in Table 65 show the correlations between cohesion and overall writing quality for Task 1. All connectives (M = .07, SE = .01) were not significantly correlated with overall writing quality, r = .06, p = .62. AdjacentP (M = 7.52, SE = 1.97) was also not significantly correlated with overall writing quality (M = 18.09, SE = 2.98), r = .32, p = .01.

Table 65. *Pearson Correlations Between Cohesion and Overall Writing Quality in Task 1*

	1	2	3
1. Overall WQ	—		
2. All connectives	.06	—	
3. AdjacentP	.32	-.04	—

Note. Overall WQ = Overall writing quality.

Main Analysis for Hypothesis 5 for Task 2

Table 66 shows the correlations between All connectives (M = .04, SE = .02), AdjacentP (M = 7.57, SE = 2.32), and overall writing quality (M = 19.99, SE = 3.13) for Task 2. All connectives were not significantly correlated with overall writing quality, r = .13, p = .30. AdjacentP was significantly and positively correlated with overall writing quality at r = .38, p < .017. The correlation was moderately strong. All connectives were not significantly correlated with overall writing quality.

Table 66. *Pearson Correlations Between the Cohesion Measures and Overall Writing Quality in Task 2*

	1	2	3
1. Overall WQ	—		
2. All connectives	.13	—	
3. AdjacentP	.38*	.27	—

* Correlation is significant at the .017 (.05/3) level (2-tailed).
Note. Overall WQ = Overall writing quality.

Summary of the Results for Hypothesis 5

The results partially supported Hypothesis 5. All connectives were not significantly correlated with overall writing quality in both tasks, but AdjacentP was positively correlated with overall writing quality in Task 2. The fact that AdjacentP was

not significantly correlated indicated that the participants learned in Task 1 and improved on AdjacentP in Task 2.

CHAPTER 6

DISCUSSION

In this chapter, the results of the five hypotheses are summarized and interpreted. I also present research implications and pedagogical implications regarding L2 writing performances at the end of the chapter. Data from the interviews with the six students are included to better explain some of the quantitative findings.

The Effects of Explicit Instruction on Target Vocabulary, Target Cohesion, and Overall Writing Quality

Hypothesis 1 concerned the effect of explicit instruction and feedback on target word, target cohesion use, and overall writing quality. Hypothesis 1 was supported. Target word and cohesion use from the model essays were analyzed for the explicit instruction group only. The descriptive statistics showed that although the total number of target words and cohesion use was limited, the explicit instruction group increasingly used them in their drafts. The interviewees in the explicit instruction group mentioned that they tried to use the target words and cohesive devices as well as a variety of words, so they became more aware of vocabulary and cohesion in their writing than the participants in the comparison group. This finding was partly shown by the fact that the interviewees in the comparison group rarely mentioned vocabulary. The results demonstrated that explicit instruction helped direct the learners' attention to linguistic form, such as vocabulary and cohesion, and thereby raised their awareness of those forms. Drawing their attention to target words and cohesion helped them become aware of and use a wider range of vocabulary and cohesive devices productively. However, the

interviewees and other participants also mentioned that they could not use all the target vocabulary at times because some of the vocabulary did not fit the content of their essays even though common vocabulary was selected for the present study. Although the target words and cohesion phrases were part of the model essays, acquiring them was not the purpose of this study.

The results of a mixed-design MANOVA on overall writing quality showed significant group differences. The explicit instruction group scored significantly higher than the comparison group on both Task 1 and Task 2, indicating that there were positive effects of explicit instruction and feedback on overall writing quality. Effect sizes were large at $\eta 2$ = .87 and .89. The results also showed that overall writing quality improved significantly over time. Although the explicit instruction group improved more, the comparison group also displayed improvement over the three drafts. The third draft scored the highest, followed by the second draft, and then the first draft. Although the explicit instruction group was more aware of vocabulary and cohesion, interviewees in both groups mentioned that they also focused on how ideas were structured in their essays. The tasks were to write essays, so the participants paid attention to the overall quality of their essays in addition to the target linguistic forms. Among improvements over drafts, larger differences between the two groups were found in the second draft in part because the second draft was where the explicit instruction group increased their overall writing quality scores the most. Compared to this result, the comparison group tended to score similarly on the first and second drafts. Hence, explicit instruction sped up the learning of the explicit instruction group participants.

The results showed that the overall writing quality score for Task 2 was significantly higher than that for Task 1. This finding indicated that some learning had occurred during Task 1, so the participants could focus more on writing in Task 2. It might also be that the problem-solution essay was more complex than the comparison-contrast essay, thus leading the participants to scoring higher on Task 2. In addition, because of a practice effect, the comparison group also improved because they received general writing instruction as a whole class. The results of the present study are in line with previous studies that found that writing quality measures often improve even in short-term studies (Bulte & Housen, 2014; Crossley et al. 2016; Larsen-Freeman, 2006; Storch, 2009). This course was the participants' first time writing an essay, so improvements were likely; however, the importance of revising drafts needs to be emphasized, as there was considerable progress made from the first to the third draft.

Providing model essays helped the participants with their writing, vocabulary, and cohesion. As Hanaoka and Izumi (2012) and Hanaoka (2007) found, model essays are beneficial. McCarthy and Carter (1997) stated that written input is a key source of positive evidence for language learning. Exposure to model essays benefits learners as they allow them to learn different writing genres. Beginner writers often do not have enough schema about English essays, so it is helpful for them to see how words are used in context, how ideas are organized, and how cohesion is achieved in the target language. Understanding the similarities and differences in different genres is helpful for students as it lessens their burden when learning to write in a foreign language.

In addition, embedding target vocabulary and cohesion in the model essays helps the learners understand how to use those words in original contexts. Using words from

readings is a useful activity because it allows learners to recognize and recycle words they have studied and use vocabulary productively in their essays. Hence, providing input concerning target forms through model essays is beneficial and presenting them frequently and repeatedly is beneficial for learners. Interviewees from the explicit instruction group commented that it was useful to know how the words were used in the model essays and they began to recognize the frequency of certain words they used in their essays. They also noticed that they could use words they had already learned in their essays. Not only did they notice how target words were used, but they also learned how ideas were structured, and how examples are used to support topic sentences.

The results of the present study show that explicit instruction strongly affects target word use, cohesion, and overall writing quality. The present study contributes to the understanding of the role of explicit instruction in L2 writing because the number of studies on this issue is limited (Hennebry et al., 2017; Nation, 2005; Norris & Ortega, 2000). These results support the claims made by Hinkel (2002), Nation, (2013), Read (2000), and Schmitt (2000) that explicit teaching is effective for the acquisition of L2 vocabulary. The results are also in line with studies by Lee (2003) and Lee and Munice (2006). As Young-Davy (2014) stated, explicit instruction speeds up learning processes and promotes noticing.

In the present study, general feedback was provided on content, overall writing quality, target vocabulary and cohesion, organization, grammar, and punctuation. The present study supports previous studies that have shown that learners improve their writing after receiving feedback (Bitchener, 2008; Bitchener & Knoch, 2010; Ferris &

Roberts, 2001; Sheen et al., 2009). Feedback and subsequent revisions promote literacy development (Leki, 2006; Storch & Tapper, 2009).

The Development of Lexical Complexity

Hypothesis 2 stated that the lexical indices would improve over the three drafts, and that the explicit instruction group would improve more than the comparison group. Three lexical indices—lexical diversity, frequency, and trigrams—were used in the present study to examine the characteristics of vocabulary development based on previous work in which automated computational tools were used (Crossley & McNamara, 2011, 2014; Crossley et al., 2018; Kim et al., 2018; Zenker & Kyle, 2021).

The results of the present study showed that some lexical indices showed improvement, while others did not. Hence, this hypothesis was partially supported. More improvements were found for Task 1 than for Task 2. In Task 1, MATTR and COCAFrq improved significantly, but Trigrams did not. The learners significantly increased their use of a variety of words and low-frequency words as they increased the length of their essays. In Task 2, only MATTR improved over the three drafts. MATTR is a stable measure across text length, so the result reflected actual vocabulary use. The use of low-frequency words improved in Task 1 only.

In terms of group differences on each lexical index, there was a significant difference for COCAFrq. The comparison group scored higher on COCAFrq, but the human raters evaluated the comparison group essays lower on vocabulary use in Task 1, possibly because the learners' choice of words might not have been appropriate. In Task 2, no group differences were found for the three lexical indices.

The present study supports previous studies in which lexical diversity measures were used. Those studies have shown that L2 writers' vocabulary improves over time (Búlte & Housen, 2014; Crossley et al., 2016; Frear & Bitchener, 2015; Mazgutova & Kormos, 2015). In the present study the lexical diversity measure MATTR was used because Zenker and Kyle (2021) found that MATTR 50 is the most stable index for estimating lexical diversity. Moreover, in a preliminary multicollinearity check, MTLD displayed multicollinearity with MATTR50. The results of the present study show that MATTR developed over the three drafts in both tasks, in part due to the explicit instruction provided to the experimental group. Yousefi and Biria (2018) showed that the effect size of the overall effectiveness of explicit instruction on L2 vocabulary was medium to large (d = .80), there was a small effect size (d = .53) for advanced and elementary levels (d = .54), and a large effect size (d = .95) for intermediate L2 learners. As previous researchers have noted, lexical diversity is a useful measure of L2 writing development (Búlte & Housen, 2014; Crossley & McNamara, 2014; Crossley et al., 2018; Zenker & Kyle, 2021). The learners in the present study were low proficiency, so it might have been easier for them to show improvement.

In the present study, COCAFrq improved on Task 1, but not on Task 2. One interpretation for this result is that the participants learned from writing the Task 1 essay, so they wrote more ideas in the first draft of Task 2 and thereby did not make major changes in their later drafts. This possibility might be one reason why the learners' vocabularies did not differ significantly over multiple drafts in Task 2. Muncie (2002) also reported that learners' vocabularies were not significantly different over multiple drafts, but there was a tendency for lower-frequency words to increase over successive

drafts. A few interviewees in the explicit instruction group mentioned that they learned to use easier words they knew in their essays instead of more difficult words. As a result, their scores for COCAFrq decreased over time. The comparison group participants used more low-frequency words, but their word choices were often unnatural. This issue is possibly one reason why the human raters evaluated the essays of the comparison group lower than that of the explicit instruction group. The use of low-frequency words differs depending on the learners' proficiency level and the genre. In the present study, the academic section of COCA was used, which was likely too technical considering the proficiency level of participants in the present study. However, the use of word frequency as a measure can uncover important characteristics of L2 writers.

Some lexical indices might not improve in the short-term. For instance, the results of the present study showed that Trigrams did not improve during the academic semester. As Guo et al. (2020) and Kim et al. (2018) stated, bigram and trigram association strength are important factors in human judgments of L2 writing proficiency; however, their participants had relatively high English proficiency. Bestgen and Granger (2014) showed that more proficient L2 learners tend to use more strongly associated word combinations. The productive use of commonly co-occurring word combinations requires higher proficiency and repeated exposure to target language vocabulary. In a study with secondary school learners, De Wilde (2023) found that trigrams accounted for a small part of the variance in a regression model; however, the underuse of trigrams suggests that learners have low proficiency, which was the case for the learners in the present study. Researchers have suggested that even advanced proficiency L2 writers are likely to underuse strongly associated collocations (Durrant & Schmitt, 2009); the same can be

said for three-word combinations. Instead of commonly occurring three-word sequences, as Eguchi and Kyle (2023) stated, second language learners use word combinations that differ from those commonly used by native English speakers. At least one year might be needed to develop the ability to use more frequent word combinations, as the study by Kim et al. (2018) showed. The results of the present study suggest that although the gains were small, n-grams can discriminate learners at different levels of proficiency. Hence, learners need more exposure to word combinations.

In sum, lexical diversity is a stable measure and reflects learners' use of vocabulary. As for the use of frequency lists, the spoken subsection of COCA or other less technical word lists might produce different results and be more appropriate for low-proficiency learners.

The Development of Cohesion

Hypothesis 3 stated that local and global cohesion would develop over the three drafts and that the explicit instruction group would perform better than the comparison group where cohesion is concerned. The results of previous cohesion studies have shown mixed results (Connor, 1990; Crossley et al., 2016; Crossley & McNamara, 2012), but in this study, local and global cohesion generally showed growth in both tasks. Hypothesis 3 was mostly supported for local and global cohesion in Task 1 and for global cohesion in Task 2.

The learners improved in their use of All connectives from the first to the second drafts, but they did not improve from the second to the third drafts in Task 1. In Task 2, All connectives did not improve. There were no significant group differences for both tasks, but the explicit instruction group tended to score higher than the comparison group.

Based on my impressions, after the participants wrote the second draft, they tended not to revise their essays greatly; they mainly changed minor errors, such as correcting grammatical errors or adding limited information to meet the growing essay length requirements. This approach to revising might be reflected in the use of connectives from the second to the third drafts. In addition, they were not fully capable of using connective devices, although the concept of connectives was easy for them to grasp. Regular practice and explicit instruction are needed to promote the use of connective devices.

Genres can affect the use of connectives given that the learners used fewer connectives in Task 2 compared to Task 1. Connectives were explicitly used for listing ideas or reasons and contrasting ideas or opinions. In this sense, Task 1 promoted the listing of comparison points and contrasting ideas to connect sentences. In contrast, the problem-solution essay in Task 2 required the use of explanations and examples. Narratives might promote the greater use of more explicit connectives. For example, Tabari and Johnson (2023) analyzed narrative and argumentative essays for the use of connective devices and found that narrative texts rely on connectives, while argumentative essays rely on global repetition.

Global cohesion improved significantly in both Task 1 and Task 2, and the explicit instruction group scored significantly higher than the comparison group. Thus, the present study indicates that lower-proficiency learners can improve their use of cohesion through instruction on global cohesion. Model essays can help learners write better essays as they learn how ideas are presented and connected in English texts. The interviewees in the explicit instruction group mentioned that they learned how ideas are expressed in the model essays. On the other hand, the comparison group wrote drafts

without exposure to model essays and feedback. They depended on their limited knowledge about writing and general writing instruction. More effects were found for global cohesion than for All connectives and Trigrams in the present study. Hence, Hypothesis 3 was mostly supported.

Lexical Complexity and Overall Writing Quality

Hypothesis 4 stated that lexical complexity was related to overall writing quality, and that the explicit instruction group would produce higher quality essays than the comparison group. This hypothesis was supported for both Task 1 and Task 2. The results of the present study supported previous studies that have shown that lexis was positively related to L2 writing proficiency (Crossley & McNamara, 2011; Guo et al., 2013; Laufer & Nation, 1995; Staple & Reppen, 2016). For instance, Kim et al. (2018) found that lexis was significantly and positively correlated with writing quality scores. In particular, MATTR and COCAFrq were correlated strongly with writing quality in Task 1. MATTR was positively correlated with overall writing quality and the effect size was large in Task 1, indicating that if learners used a variety of words, their essays were evaluated higher by raters. Peng et al. (2023) also found that lexical diversity as assessed with MTLD was strongly correlated with writing quality, which was strongly correlated with MATTR in the preliminary analysis of the present study. On the other hand, COCAFrq was negatively correlated, and the effect size was large in Task 1. If the participants used low-frequency words, their essays were evaluated lower by the raters in Task 1. In Task 2, none of the correlations were significant, however. Negative correlations with overall writing quality indicated that the comparison group might not have used low-frequency words appropriately. In addition, in Task 2, lexical diversity was not preferred in the

problem-solution essay. Overall writing quality was assessed with vocabulary, content, organization, grammar, and punctuation, so other features of the essays were correlated with overall writing quality in this genre.

In both tasks, Trigrams were weakly correlated with overall writing quality, a finding that differed from Kim et al. (2018), who stated that n-gram frequencies were positively correlated with proficiency and overall writing quality. This finding might be true for higher proficiency learners who use more native-like expressions, but the results of the present study indicate that this is not the case for low-proficiency learners. They might be primarily focused on using single words and therefore pay little attention to two- and three-word combinations.

In sum, COCAFrq correlated negatively with overall writing quality for both tasks. This result was the opposite of the results reported by Guo et al. (2013) perhaps because the learners in the present study might not have used the low-frequency words appropriately. The use of a wide variety of words as assessed by MATTR, was evaluated lower in Task 2, possibly because the comparison-contrast essay might not have required the use of low-frequency words. In contrast, low-frequency words might be used more frequently in problem-solution essays. Different genres require the different use of vocabulary.

Cohesion and Overall Writing Quality

Hypothesis 5 concerned the extent to which cohesion was related to overall writing quality. Hypothesis 5 was partially supported for Task 1 and Task 2. In both tasks, All Connectives was not significantly correlated with overall writing quality. Hence, the results are inconclusive. The underuse of connectives might be one reason for

this result. Studies with L1 writers have shown a negative correlation between connectives and writing quality, though Crossley and McNamara (2011) showed a positive correlation for L2 writers. Tabari and Johnson (2023) found that no single combination of cohesive devices could predict L2 writing quality. Genre differences likely influence the results, so results in this area must be viewed with genre in mind. In addition, comparing high- and low-rated essays might shed light on the use of connectives and their relation to overall writing quality.

On the other hand, AdjacentP was significantly correlated with overall writing quality for Task 2, and not for Task 1. In Task 1, the participants were still learning about writing; however, they had improved cohesiveness between paragraphs by Task 2. The results for AdjacentP in the present study are in line with the findings reported by Crossley and McNamara (2011) and Crossley et al. (2016). Crossley et al. (2019) found that essays that had greater global cohesion were scored higher. Repeated word use between paragraphs created better writing quality. The use of global cohesion by low-proficiency learners is more easily learned than local cohesion. In addition, the participants were university students, so L1 knowledge of writing might have helped them write reasonably good essays at the global level.

Pedagogical Implications

The effect of explicit instruction, such as the use of model essays and feedback, is beneficial for L2 writing development including vocabulary, cohesion, and overall writing quality. It helps to draw learners' attention to teaching targets, such as vocabulary, cohesion, and overall writing quality, and speeds up learning processes in a complex task of writing. It promotes learners to notice how words and cohesions are

used, as well as how ideas are expressed in English. It is beneficial to teach writing skills explicitly as well as to provide feedback to learners explicitly.

Providing target words in model essays is beneficial. Interviewees from the explicit group mentioned this when they said that they learned the type of vocabulary used in the essays. As Logan (1988) stated, by providing more exposure to written input and by having learners allocate attention to vocabulary and cohesion in the input, the probability that the input becomes intake is increased. Providing model essays also functions as one way to increase the learners' understanding of how to write their essays. In addition, using vocabulary in context also promotes deeper semantic processing (Barcroft, 2015). Learners can expand their lexical network and learn different nuances of the target vocabulary.

As Laufer (1998) claimed, productive vocabulary does not necessarily develop in parallel with receptive vocabulary, so engaging in explicit vocabulary and cohesion tasks is important (Craik & Lockhart, 1972; Laufer & Hulstijn, 2001). As Nation (2013) stated pushed output is useful for vocabulary learning, as it allows learners to use words in a variety of contexts, test hypotheses concerning its use, and engage in metalinguistic reflection. Model essays were found to be useful, so they could practice writing a paragraph or an essay using the target vocabulary from the reading materials that they read. Model essays can help learners use words they already know and new words productively, as they are a source of positive input.

Revising skills are important, so having learners revise their essays over multiple drafts is a key to allowing learners to reassess their use of vocabulary and cohesive devices. Some Japanese learners are not accustomed to revising their essays multiple

times, so the benefits of revising should be emphasized more in some educational contexts. In addition to revising skills, the importance of brainstorming ideas should also be emphasized, as the development of related ideas are connected to the use of vocabulary and cohesion.

If lower-proficiency learners become more fluent in expressing their ideas in English essays, they can focus more on other aspects of writing, such as vocabulary choice, cohesion, and grammar. Along with explicit teaching of vocabulary and cohesion, writing fluency exercises can benefit learners greatly. As a part of fluency exercises, learners can be provided with model essays and practice writing in different genres. Accumulating experience writing in many different contexts is important in developing L2 writing skills.

CHAPTER 7

CONCLUSION

In this final chapter, I present an overall summary of the findings of the study, discuss the limitations of the study, propose possibilities for future research, and offer concluding remarks.

Summary of the Findings

The purpose of this study was to investigate the effects of explicit instruction on lexis, cohesion, and writing quality scores in L2 writing, and how lexis and cohesion are correlated with overall writing quality using automated computational tools. The participants, 65 Japanese university FEL students in two intact reading and writing classes, were divided into an explicit instruction group and a comparison group. The explicit instruction group was provided with a model essay with embedded target vocabulary and cohesion, while the comparison group was not. Both groups completed a comparison-contrast essay and a problem-solution essay, and they wrote three drafts for each essay over one academic semester. Interviews with six students from each class were used to support the results from the quantitative analyses. This study produced six main findings.

The first finding is that overall writing quality improved over time. The essays written by the explicit instruction group were rated higher than those written by the comparison group for both Tasks 1 and 2. The results suggest that explicit instruction and feedback are beneficial for improving overall writing quality as evaluated by human raters. Providing instruction, feedback, and a model essay is useful for supporting

students in their efforts to learn not only how to express ideas in the second language, but also how vocabulary is used in context. First-year Japanese university students do not have experience writing essays in either Japanese or English, so explicitly teaching aspects of essay writing such as brainstorming, outlining, organization, grammar, vocabulary, cohesion, and formatting is helpful in developing writing skills.

The second finding is that explicit instruction is helpful in the development of productive vocabulary and cohesion skills in L2 writing. In both essays, the explicit instruction group used target words and cohesive devices more skillfully as they progressed through the three drafts. As Nation (2013) suggested, explicit instruction is beneficial for improving productive vocabulary learning. Mazgutova and Kormos (2015) noted that feedback is useful for productive vocabulary learning as well as other features of writing. Furthermore, the effects of instruction were observed in terms of cohesion.

The third finding is that neither group demonstrated improvements in Trigrams over three drafts. This result suggests that beginner writers lack the use of trigrams in their writing, so they need to study phrasal verbs, idioms, and two or three words that frequently co-occur if they are to be more proficient in L2 writing. The fact that the explicit instruction group scored higher than the comparison group on COCA academic trigram MI2 indicated that explicit instruction can help learners improve their use of trigrams.

The fourth finding is that both groups showed significant improvements on global cohesion over the three drafts in both tasks. This finding is important because global cohesion is related to L2 proficiency. Local cohesion showed significant improvement on Task 1, but not on Task 2, a finding that might have been affected by genre differences.

This result is similar to that of previous studies that have also shown mixed results concerning cohesion, though there is a tendency for higher-proficiency writers to use more cohesive devices than lower-proficiency writers (Connor, 1990; Witte & Faigley, 1981).

The fifth finding is that the lexical indices were correlated with the writing quality measure, a result that is consistent with previous studies (Crossley et al., 2011; Crossley & McNamara, 2012; Kim et al., 2018; Verspoor et al., 2012). This finding emphasizes the importance of vocabulary instruction in second language writing courses. For instance, learners that used a wider variety of words in their essays were evaluated better by the raters.

The sixth finding is that global cohesion correlated positively with overall writing quality in Task 2. This study showed that global cohesion can be taught and that as learners acquire the ability to use it more effectively, it exerts a positive influence on writing quality.

Limitations of the Study

All drafts were written at home, a factor that might have influenced the results. As a result, time on task was not controlled for and the learners could have consulted dictionaries, translation software, or other sources of information about the English language.

In this study the lexical indices were focused on the participants' use of content words, so function words were not examined. Depending on the proficiency level, the examination of function words can reveal more about learners' levels of writing abilities.

Suggestions for Future Research

Investigating the effects of reading model essays on writing quality can be a potentially profitable research area given that model essays provide an excellent source of evidence about how the foreign language works. For instance, researchers could investigate the use of multiple versus single model essays and the use of input enhancement in the model essays.

Detailed research into the similarities and differences between higher- and lower-rated essays might shed light on how various writing skills such as the use of cohesive devices develop. This type of research could also be conducted using longitudinal designs in which lower-proficiency writers' development is tracked over one or more years. Moreover, essays produced by L2 writers can be compared with texts written by native speakers of English to investigate how L2 writers can move closer to native speakers' norms.

The relationship between cohesive devices and overall writing quality is in the early stage of research (Tabari & Johnson, 2023), so researchers need to investigate this issue further. Cohesion is important given that different languages handle cohesion differently; thus, L1 transfer is bound to occur. The degree to which this type of transfer occurs is largely unknown, so it is an area that requires a large number of studies.

Only two genres, comparison-contrast and problem-solution, were analyzed in this study, so no results are available for other genres. Different genres can require different kinds of vocabulary and cohesive devices. To acquire a more detailed understanding of this area of research, other genres need to be investigated. Moreover,

replication studies are needed to confirm the results with genres that have been investigated previously.

Finally, depending on the research participants' English proficiency, particular indices might or might not capture growth in vocabulary and cohesion. Thus, more research on the efficacy of commonly used indices needs to be conducted with second language learners at different proficiency levels. In addition, replication studies are needed to confirm previous findings.

Concluding Comments

In the present study I investigated the effects of explicit instruction and feedback on learning vocabulary and cohesion in L2 writing development. Explicit instruction and feedback with the provision of model essays had a positive effect on the development of productive vocabulary and cohesion.

Writing is a creative product, so its assessment is a complex phenomenon. Choosing indices that accurately assess the use of lexis and cohesion is a great challenge but one that is crucial to better understanding the components of high-quality essays. By understanding more about the relationships among lexis, cohesion, and writing quality, researchers and educators can better serve the most important part of the teaching and learning equation—the students.

REFERENCES

Barcroft, J. (2002). Semantic and structural elaboration in L2 lexical acquisition. *Language Learning, 52*(2), 323–363. https://doi.org/10.1111/0023-8333.00186

Barcroft, J. (2004). Effects of sentence writing in L2 lexical acquisition. *Second Language Research, 20*(4), 303–334. https://doi.org/10.1191/0267658304sr233oa

Barcroft, J. (2006). Can writing a new word detract from learning it? More negative effects of forced output during vocabulary learning. *Second Language Research, 22*(4), 487–497. https://doi.org/10.1191/0267658306sr276oa

Barcroft, J. (2015). *Lexical input processing and vocabulary learning*. Benjamins.

Bastola, M. N., & Hu, G. (2023). From the other side of the desk: Supervisor's perceptions of supervisory feedback. *Journal of Second Language Writing, 59*, 100965. https://doi.org/10.1016/j.jslw.2023.100965

Benesse Corportation (n.d.). Internet-based GTEC test.

Biber, D., Nekrasova, T., & Horn, B. (2011). *The effectiveness of feedback for L1-English and L2 writing development: A meta-analysis. TOEFL iBT™ Research Report TOEFL iBT-14. ETS.*

Bitchener, J. (2008). Evidence in support of written corrective feedback. *Journal of Second Language Writing*, *17*, 102–118. https;//doi.org/10.1016/j.jslw.2007.11.004

Bitchener, J., & Knoch, U. (2010). Raising the linguistic accuracy level of advanced L2 writers with written corrective feedback. *Journal of Second Language Writing, 19*(4), 207–217. https://doi.org/10.1016/j.jslw.2010.10.002

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental measurement in the human sciences* (2nd ed.). Routledge

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes, in L2 writing complexity. *Journal of Second Language Writing*, *26*, 42–65. https://doi.org/10.1016/j.jslw.2014.09.005

Chen, M., & Cui. Y. (2022). The effects of AWE and peer feedback on cohesion and coherence in continuation writing. *Journal of Second Language Writing, 57*, 100915. https://doi.org/10.1016/j.jslw.2022.100915

Chaudron, C. (1985). Intake: On models and methods for discovering learners' processing of input. *Studies in Second Language Acquisition*, *7*, 1–14. https://doi.org/10.1017/S027226310000512X

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Connor. U. (1990). Linguistic/rhetorical measures for international persuasive student writing. *Research in the Teaching of English*, *24*, 67–87.

Connor-Linton, J., & Polio, C. (2014). Comparing perspectives on L2 writing: Multiple analyses of a common corpus. *Journal of Second Language Writing*, *26*, 1–9. https://doi.org/10.1016/j.jslw.2014.09.002

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, *17*(2), 94–100. https://doi.org/10.1080/09296171003643098

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*, 213–238. https://doi.org/10.2307/3587951

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2011a). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, *29*(2), 243–263. https://doi.org/10.1177/0265532211419331

Crossley, S. A., & McNamara, D. S. (2011). Shared features of L2 writing: Intergroup homogeneity and text classification. *Journal of Second Language Writing*, *20*, 271–295. https://doi.org/10.1016/j.jslw.2011.05.007

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, *35*(2), 115–135. https://doi.org/10.1177/0265532211419331

Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, *26*, 66–79. https://doi.org/10.1016/j.jslw.2014.09.006

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011b). Predicting lexical proficiency in language learners using computational indices. *Language Testing*, *28*(4), 561–580. https://doi-org.libproxy.temple.edu/10.1177/0265532210378031

Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The tool for the automatic analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods, 51*(1), 14–27. https://doi.org/10.3758/s13428-018-1142-4

Crossley, S. A., Kyle, K., & McNamara, D. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing, 32*, 1–16. https://doi.org/10.1016/j.jslw.2016.01003

Council of Europe. The common European framework of reference for languages.

Cumming A., Kantor, R., Baba, K., Erdoosy, U., Eouanzoui, K., & James, M. (2005). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL* (*TOEFL Monograph, No. MS–30*). Educational Testing Service. https://doi.org/10.1002/J.2333-8504.2005.TB01990.X

Daller, H., Van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, *24*(2), 197–222. https://doi.org/10.1093/applin/24.2.197

Davis, M. (2008). The corpus of contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, *25*(4), 2010. https://doi.org/10.1093/llc/fqq018

de Bot, K. (1996). The psycholinguistics of the output hypothesis. *Language Learning, 46*(3), 529–555. https://doi.org/10.1111/j.1467-1770.1996.tb01246.x

De Jong, N. H., Steinel, A. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*, 5–34. https://doi.org/10.1017/S0272263111000489

De Wilde, V. (2023). Lexical characteristics of young L2 English learners' narrative writing at the start of formal instruction. *Journal of Second Language Writing, 58*, 100960. https://doi.org/10.1016/j.jslw.2022.100960

DeKeyser, R. (2007). *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*. Cambridge University Press.

Dimas, H. M. S. (2011). Explicit vocabulary instruction in an English content-area course with university student teachers: When comprehensible input needs to be comprehended. *Gist Education and Learning Research Journal*, *5*, 84–103.

Doró, K., & Päivi Pietilä, P. (2015). Researching vocabulary in L2 writing: Methodological issues and pedagogical implications. In P. Pietilä, K. Doró, & R. Pípalová (Eds.), *Lexical issues in L2 writing* (pp. 11–28). Cambridge Scholars.

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocation. *International Review of Applied Linguistics in Language Teaching*, *47*(2). https://doi.org/10/1515/iral.2009.007

Eckstein, G., & Ferris, D. (2017). Comparing L1 and L2 texts and writers in first0year composition. *TESOL Quarterly*, *52*(1), 137–162. https://doi.org/10.1002.tesq.376

Eguchi, M., & Kyle, K. (2020). Continuing to explore the multidimensional nature of lexical sophistication: The case of oral proficiency interviews. *The Modern Language Journal*, *104*(2), 381–400. https:/doi.org/10.1111.modl.12637

Ellis, R. (2008). Explicit form-focused instruction and second language acquisition. In B. Spolsky & F. M. Hult (Eds.), *The handbook of educational linguistics* (pp. 437–455). Wiley-Blackwell.

Ellis, R., Sheen, Y., Murakami, M., & Takashima, H. (2008). The effects of focused and unfocused written corrective feedback in an English as a foreign language context. *System*, *36*, 353–371. https://doi.org/10.1016/j.system.2008.02.001

Ferris, D. R., & Roberts, B. J. (2001). Error feedback in L2 writing classes: How explicit does it need to be? *Journal of Second Language Writing*, *10*, 161–184. https://doi.org/10.1016/S1060-3743(01)00039-X

Field, A. (2009). *Discovering statistics using SPSS* (2nd ed.). Sage.

Folse, K. (2006). The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly*, *40*(2), 273–293. https://doi.org/10.2307/40264523

Frear, M. W., & Bitchener, J. (2015). The effects of cognitive task complexity on writing complexity. *Journal of Second Language Writing*, *30*, 45–57. https://doi.org/10.1016/j.jslw.2015.08.009

Friginal, E., & Weigle, S. (2014). Exploring multiple profiles of L2 writing using multi-dimensional analysis. *Journal of Second Language Writing*, *26*, 80–95. https://doi.org/10.1016/j.jslw.2014.09.007

Garner, J., Crossley, S., & Kyle, K. (2020). Beginning and intermediate L2 writer's use of N-grams: An association measures study. *International Review of Applied Linguistic in Language Teaching, 58*(1), 51–74. https://doi.org/10.1515/iral-2017-0089

Goo, J., Granea, G., Yilmaz, Y., & Novella, E. (2015). Implicit and explicit instruction in L2 learning: Norris & Ortega (2000) revisited and updated. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages* (pp. 443–483). Benjamins.

Granger, S., Dagneaux, E., & Meunier, F. (2002). *International corpus of learner English*. Université Catholique de Louvain: Centre for English Corpus Linguistics.

Granger, S., & Gestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching, 52*(3). https://doi.org/10.1515/iral2014-001

Green, A. (2005). EAP study recommendations and score gains on the IELTS academic writing test. *Assessing Writing, 10*, 44–60. https://doi.org/10.1016/j.asw.2005.02.002

Graesser, A. C., McNamara, & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40*(5), 223–234. https://doi.org/10.3102/0013189X11413260

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers, 36*, 193–202. https://doi.org/10.3758/BF03195564

Gries, S. T., & Ellis, N. C. (2015). Statistical measures for usage-based linguistics. *Language Learning, 65*(1), 228-255. https://doi.org/10.2307/411255

Guiraud, P. (1960). *Problémes et méthodes de la statistique linguistique* [*problems and methods of linguistic statistics*]. Presses Universitaries de France.

Guo, L., Crossley, S. A., & McNamara, D. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing, 18*, 218–238. https://dx.doi.org/10.1016/j.asw.2013.05.002

Halliday, M., & Hassan, R. (1976). *Cohesion in English*. Longman.

Hanaoka, O. (2007). Output, noticing, and learning: An investigation into the role of spontaneous attention to form in a four-stage writing task. *Language Teaching Research*, *11*(4), 459–479. https://doi.org/10.1177/1362168807089631

Hanaoka, O., & Izumi, S. (2012). Noticing and uptake: addressing pre-articulated covert problems in L2 writing. *Journal of Second Language Writing*, *21*(4), 332–347. https://doi.org/10.1016/j.jslw.2012.09.008

Hennebry, M., Rogers, V., Macaro, E., & Murphy, V. (2017). Direct teaching of vocabulary after listening: Is it worth the effort and what method is best? *The Language Learning Journal*, *45*(3), 282–300. https://doi.org/10.1080/09571736.2013.849751

Herdan, G. (1960). *Type-token mathematics: A textbook of mathematical linguistics*. Mouton.

Hinkel, E. (2002). *Second language writers' text.* Erlbaum.

Hinkel, E. (2015). Language focus: Teaching academic vocabulary, collocations, and pre-fabs. In E. Hinkel (Ed.), *Effective curriculum for teaching L2 writing* (pp. 185–215). Routledge. https://doi.org/10.4324/9780203830154.

Hulstijn, J. H. (2002). Incidental and intentional learning. In C. Doughty, & M. H. Long, (Eds.), *Handbook of second language acquisition* (pp. 349–381). Blackwell.

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. In S. Ishikawa (Ed.), *Learner corpus studies around the world* (pp. 99–118). Kobe University.

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning*, *63*(1), 87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x

Johnson, W. (1994). Studies in language behavior 1: A program of research. *Psychological Monographs: General and Applied*, *56*(2), 1–15. https://doi.org/10.1037/h0093508

Kawamoto, S. (2017). The effect of written corrective feedback on cohesion in Japanese as a foreign language learners' writing [Unpublished master's thesis]. Purdue University.

Kellogg, R. T. (2001). Competition for working memory among writing processes. *American Journal of Psychology*, *114*, 175-192. https://doi.org/10.2307/1423513

Kim, M. Y., Crossley, S. A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal*, *102*(1), 120–141. https://doi.org/10.1111/modl,12447

Knoch, U. (2007). 'Little coherence, considerable strain for readers': A comparison between two rating scales for the assessment of coherence. *Assessing Writing*, *12*(2), 108–128. https://doi.org/10.1016/j.asw.2007.07.002

Knoch, U., Rouhshad, A., Oon, S. P., & Storch, N. (2015). What happens to ESL students' writing after three years of study at an English medium university? *Journal of Second Language Writing*, *28*, 39–52. https://doi.org/10.1016/j.jslw.2015.02.005

Krashen, S. D. (1985). *The input hypothesis: Issues and implications*. Longman.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*, 757–786. https://doi.org/10.1002/tesq.194

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behavioral Research Methods*, *50*, 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of Chinese learners of English. *Applied Linguistics*, *27*, 590–619. https://doi.org/10.1093/applin/aml029

Laufer, B. (1998). The development of passive and active vocabulary in a second language: same or different? *Applied Linguistics*, *19*, 255–271. https://doi.org/10.1093/applin/19.2.255

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, *22*(1), 1–26. http://doi.org/10.1093/applin/22.1.1

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*, 307–322. https://doi.org/10.1093/applin/16.3.307

Lee, I., Luo, N., & Mak, P. (2022). Teachers' attempts at focused written corrective feedback in situ. *Journal of Second Language Writing*, *52*, 100809. https://doi.org/10.1016/j.jslw.2021.100809

Lee, S. H. (2003). ESL learners' vocabulary use in writing and the effects of explicit vocabulary instruction. *System*, *31*, 537–561. https://doi.org/10.1016/j.system.2003.02.004

Lee, S. H., & Muncie, J. (2006). From receptive to productive: Improving ESL learners' use of vocabulary in a post-reading composition task. *TESOL Quarterly*, *40*(2), 295–320. https://doi.org/10.2307/40264524

Leki, I. (2006). ''You cannot ignore'': L2 graduate students' response to discipline-based written feedback. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing* (pp. 266–286). Cambridge University Press.

Linacre, J. M. (2017). A user's guide to WINSTEPS: Rasch-model computer program. MESA.

Linacre, J. M. (2023). Facets computer program for many-facet Rasch measurement, version 3.86.0. Beaverton, Oregon: Winstep.com

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *96*, 492–527. https://doi.org/10.1037/0033-295X.95.4.492

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496.

Maas, H. D. (1972). Über den Zusammenhang zwischen Wortschatzumfang und L¨ange eines Textes [On the relationship between vocabulary and the length of a text]. *Zeitschrift für Literaturwissenschaft und Linguistik*, *2*(8), 49–73.

Marlowe, J. P. (2019). *The effects of sentence-combining* on the longitudinal *development of syntactic complexity in L2 writing*. ［Unpublished doctoral dissertation]. Temple University Japan.

Malvern, D., Chipere, N., Richards, B., & Durán , P. (2004). *Lexical diversity and language development: Quantification and assessment*. Palgrave Macmillan.

Matsuo, M. (2019). Critical reflection, unlearning, and engagement. *Management Learning*, *50*(4), 465–481. https://doi.org/10.1177/1350506719859681

Mazgutova, D., & Kormos, J. (2015). Syntactic and lexical development in an intensive English for academic purpose programme. *Journal of Second Language Writing*, *29*, 3–15. https://doi.org/10.1016/j.jslw.2015.06.004

McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research*, *19*(6), 741-760. https://doi.org/10.1177/1362168814567889

Merkel, W. (2018). Role reversals: A case study of dialogic interactions and feedback on L2 writing. *Journal of Second Language Writing, 39,* 16-28. https://doi.org/10.1016/j.jslw.2017.11.007

Min, H. T. (2006). The effect of trained peer review on EFL students' revision types and writing quality. *Journal of Second Language Writing, 15,* 118–141. https://doi.org/10.1016/j.jslw.2006.01.003

McCarthy, M., & Carter, R. (1997). Written and spoken vocabulary. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition, and pedagogy* (pp. 20–39). Cambridge University Press.

McCarthy, P. M. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual lexical diversity (MTLD)* [Unpublished doctoral dissertation]. University of Memphis.

McCarthy, P. M., & Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. *Language Testing*, *24*(4), 459–488. https://doi.org/10.1177/02065532207080767

McCarthy, P. M., & Jarvis, S. (2010). MTLD, voc-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*(2), 381–392. https://doi.org/10.3758/brm.42.2.381

McKeown, M. G., Deane, P. D., Scott, J. A., Krovetz, R, & Lawless, R. R. (2017). *Vocabulary assessment to support instruction: Building rich word-learning experiences*. Guildford.

Muncie, J. (2002). Process writing and vocabulary development: Comparing lexical frequency profiles across drafts. *System, 30*, 225–235. https://doi.org/10.1016/S0346-251X(02)00006-4

Muranoi, (2007). Output practice in the L2 classroom. In R. DeKeyser (Ed.), *Practice in a second language: Perspectives from applied linguistics and cognitive psychology* (pp. 51–84). Cambridge University Press.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press.

Nation, I. S. P. (2005). Teaching and learning vocabulary. In. E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 581–595). Erlbaum.

Nation, I. S. P. (2013). *Learning vocabulary in another language*. Cambridge University Press.

Nation, I. S. P. (2015). Principles guiding vocabulary learning through extensive reading. *Reading in a Foreign Language*, *27*(1), 136–145.

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9–13.

Nation, I. S. P., & Heatley, A (1994). *Range: A program for the analysis of vocabulary in texts [software]*. http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Cengage Learning.

Norris, J., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*(3), 417–528. https://doi.org/10.1111/0023-8333.00136

O'Reilly, T., & McNamara, D. S. (2007). The impact of science knowledge, reading skill, and reading strategy knowledge on more traditional "high-stakes" measures of high school students' science achievement. *American Educational Research Journal*, *44*(1), 161–196. https://doi.org/10.3102/0002831206298171

Oshima, A. & Hogue, A. (2007). *Introduction to academic writing* (3rd ed.). Pearson Education.

Pajares, F. (2005). Self-efficacy during childhood and adolescence: Implications for teachers and parents. In F. Pajares & T. C. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 339–353). Information Age.

Peng, Y., Sun, J., Quan, J, Wang, Y., Lv, C. (2023). Predicting Chinese EFL learners' human rater writing quality in argumentative writing through multidimentional computational indices of lexical complexity. *Assessing Writing*, *56*, 100722. https://doi.org/10.1016/j.asw.2023.100722

Polio, C., & Shea, M. C. (2014). An investigation into current measures of linguistic accuracy in second language writing research. *Journal of Second Language Writing*, *26*, 10–27. https://doi.org/10.1016/j.jslw.2014.09.003

Prince, P. (1996). Second language vocabulary learning: The role of context versus translations as a function of proficiency. *The Modern Language Journal*, *80*, 478–493. https://doi.org/10.1111/j.1540-4781.1996.tb05468.x

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press.

Rhee, S. C., & Jung, C. K. (2014). Compilation of the Younsei English Learner Corpus (YELC) 2011 and its use for understanding current usage of English by Korean pre-university students. *Korea Contents Association*, *14*, 1019–1029.

Robb, T., Ross, S., & Shortreed, I. (1986). Salience of feedback on error and its effect on EFL writing quality. *TESOL Quarterly*, *20*, 83–95. https://doi.org/10.2307/3586390

Ruiz-Funes, M. (2015). Exploring the potential of second/foreign language writing for language learning: The effects of task factors and learner variables. *Journal of Second Language Writing*, *28*, 1–19. https://doi.org/10.1016/j.jslw.2015.02.001

Salsbury, T., Crossley, S. A., & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. *Second Language Research*, *27*, 343–360.

Seijas, J. M., & Spino, L. (2023). Written "corrective feedback in Spanish as a heritage language: Problematizing the construct of error. *Journal of Second Language Writing*, *60*, 100989. https://doi.org/10.1016/j.jslw.2023.100989

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, *36*(2), 139–152. https://doi.org/10.1080/09571730802389975

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*(2), 129–158. https://doi.org/10.1093/applin/11.2.129

Schmidt, R. W., & Frota (1986). Developing basic conversational ability in a second language. In R. R. Day (Ed.), *Talking to learn: Conversation and second language acquisition* (pp. 237–326). Newbury House.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge University Press.

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research*, *12*(3), 329–363. https//doi.org/10.1177/1362168808089921

Sheen, Y., Wright, D., & Moldawa, A. (2009). Differential effects of focused and unfocused written correction on the accurate use of grammatical forms by adult ESL learners. *System*, *37*(4), 556–569. https://doi.org/10.106/j.system.2009.09.002

Shintani, N., & Ellis, R. (2013). The comparative effect of direct written corrective feedback and metalinguistic explanation on learners' explicit and implicit knowledge of the English indefinite article. *Journal of Second Language Writing*, *22*, 286–306. https://doi.org/10.1016/j.jslw.2013.03.011

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

Sökmen, A. (1992). *Common threads: An interactive vocabulary builder*. Prentice Hall Regents.

Staples, S., & Reppen, R. (2016). Understanding first-year L2 writing: A lexico-grammatical analysis across L1s, genres, and language rating. *Journal of Second Language Writing*, *32*, 17–35. https://doi.org/10.1016/j.jslw.2016.02.002

Storch, N. (2009). The impact of studying in a second language (L2) medium university on the development of L2 writing. *Journal of Second Language Writing*, *18*, 103–118. https://doi.org/10.1016/j.jslw.2009.02.003

Storch, N., & Tapper, J. (2009). The impact of an EAP course on postgraduate writing. *Journal of English for Academic Purposes*, *8*, 207–223. https://doi.org/10.1016/j.jeap.2009.03.001

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235–253). Newbury House.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidelhofer (Eds.), *Principles and practice in applied linguistics: Studies in honor of H.G. Widdowson* (pp. 125–144). Oxford University Press.

Swain, M. (1998). Focus on form through conscious reflection. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 64–81). Cambridge University Press.

Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 471–483). Erlbaum.

Tabari, M. A., & Johnson, M. D. (2023). Exploring new insights into the role of cohesive devices in written academic genres. *Assessing Writing*, *57*, 100749. https://doi.org/10.1016/j.asw.2023.100749

Treffers-Dallers, J., & Milton, J. (2013). Vocabulary size revisited: the link between vocabulary size and academic achievement. *Applied Linguistics Review*, *4*(1), 151–172. https://doi.org/10.1515/applirev-2013-0007

Van Beuningen, C., De Jong, N. H., & Kuiken, F. (2012). Evidence on the effectiveness of comprehensive error correction in second language writing. *Language Learning*, *62*(1), 1–41. https://doi.org/10.1111/j.1467-9922.2011.00674x

Vasu, K., Ling, C. H., & Nimehchisalem, V. (2016). Malaysian tertiary level ESL students' perception toward teacher feedback, peer feedback and self-assessment in their writing. *International Journal of Applied Linguistics and English Literature*, *5*(5), 158–170. https://doi.org/10.7575/aiac.ijalel.v.5n.5p.158

VanPatten, B. (1990). Attending to form and content in the input: An experiment in consciousness. *Studies in Second Language Acquisition*, *12*, 287–301. https://doi.org/10.1017/S0272263100009177

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, *17*(1), 65–83. https://doi.org/10.1191/026553200676636328

Verspoor, M., Schmid, M. S., & Xu, X. (2012). A dynamic usage-based perspective on L2 writing. *Journal of Second Language Writing*, *21*, 239–263. https://doi.org/10.1016/j.jslw.2012.03.007

Waring, R. (1997). A comparison of the receptive and productive vocabulary sizes of some second language learners. *Immaculata*. The occasional papers at Notre Dame Seishin University.

Westbrook, C. (2014). *UnLock reading and writing 3*. Cambridge University Press.

Williams, H. (2012). Cohesion and pragmatic theory in second-language writing instruction. *Language and Linguistic Compass*, *6*(12), 768–776. https://doi.org/10.1111/lnc3.12005

Witte, S. P., & Faigley, L. (1981). Coherence, cohesion and writing quality. *College Composition and Communication*, *32(4)*, 400–414. https://doi.org/10.2307/356602

Wolfe-Quintero, K., Inagaki, S., & Kim, H.Y. (1998). Second language development in writing: Measures of fluency, accuracy & complexity. Technical report #17. *Second Language Teaching & Curriculum Center*, University of Hawaii at Manoa.

Yang, W., Lu, X., & Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing*, *28*, 53–67. https://doi.org/10.1016/j.jslw.2015.02.002

Yeung, S. S. S., Ng, M. I., Qiao, S., & Tsang, A. (2020). Effects of explicit L2 vocabulary instruction on developing kindergarten children's target and general vocabulary and phonological awareness. *Reading and Writing*, *33*, 671–689. https://doi.org/10.1007/s11145-019-09982-3

Young-Davy, B. (2014). Explicit vocabulary instruction. *ORTESOL Journal*, *31*, 26–32.

Yousefi, M. H., & Biria, R. (2018). The effectiveness of L2 vocabulary instruction: A meta-analysis. *Asian Pacific Journal of Second and Foreign Language Education*, *3*(21), 1–19. https://doi.org/10.1186/s40862-018-0062-2

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, *24*, 100505. https://doi.org/10.1016/j.asw.2020.article.100505Z

APPENDICES

APPENDIX A

SYLLABUS (JAPANESE VERSION)

RW Advanced level
【授業要旨または授業概要】

英語コミュニケーションスキルの習得は、将来グローバル化する社会に積極的に参加していくために非常に重要である。このコースでは、実践的な読む力と書く力を伸ばすことを目的とし、様々なジャンルの英語を理解するリーディング力を強化する実習と、目的や読者に応じて自分の見解を表現するライティング力を向上させる実習を組み合わせていく。２つの能力（受信と発信）を結び付けたコミュニケーションをより複雑なレベルで行うことができるよう、批判的思考力を養う活動（グループワークやディスカッションなど）を段階的に導入していく。**目指すレベルは「ヨーロッパ共通言語基準（CEFR）」の B1 レベル、つまり「自立した英語使用者」である。**

リーディング活動では、長めで複雑なエッセイ、説明文、物語文、ネット上のニュース記事の読解練習、スキャンニングやスキミングによる速読練習（1 分間 100-120 語）を行い、知らない単語や表現を文脈から推測する方策（ストラテジー）も身に付ける。ライティング活動では、受信に基づく発信として、リーディング材料から得た情報を基にして自らの見解、意見を含む文章が書けるようにすると同時に、文法力、語彙力を向上させ、教師や仲間のフィードバックを取り入れて修正した複数の原稿を作成することで文章を改善する能力も伸ばす。これらの実習を通して、批判的に考えることを学び、批判的思考力を要するディスカッション、グループワークなどの活動に参加できるようにする。

【学修の到達目標】

1. リーディング (R)

a. 長めで複雑な文章（一般的な説明文、エッセイ、物語など）の主旨や展開が理解でき、あまり辞書に頼らずに詳細もほぼ理解できる（Can-do リスト：R1, R3）。

b. 日常生活で出会うやや複雑な文章（新聞記事、ガイドブック、仕事関係の文章など）について、未知語を文脈から予測しながら必要な情報を読み取ることができ、それに含まれる語彙や複雑な構文（高校卒業レベル）を理解できる（Can-do リスト：R2, R4, R5, R6）。

2. ライティング (W)

a. 自分の好みや意見、物事の比較などについて、論理的な根拠を示しながら序論、本論、結論から成る、ある程度の長さの文章が書ける（Can-do リスト：W1, W2）。

b. 一般的な個人的、公的文章（説明文、手紙、電子メールなど）を、場面や状況に応じた適切な表現や構文、一般的なつなぎ言葉を使って書くことができる（Can-do リスト：W3, W4, W5）。

c. クラスで自分や仲間が書いた文章について客観的に評価してほぼ適切な修正ができる。また定型的な手紙や電子メールに対してほぼ適切な返事が書ける（Can-do リスト：W3, RW1, RW2, RW3）。

3. 批判的思考 (Critical Thinking: CT)

a. リーディングとライティングの活動の中で、分析、比較、判断などの複雑な思考力を用い、批判的に考えることができる。（Can-do リスト: CT）

【授業計画】スケジュール

第 1 回　Orientation
第 2 回　Orientation, Writing diagnostic test, Self-assessment 1
第 3 回　Unit 1
第 4 回　Unit 1
第 5 回　Unit 1
第 6 回　Unit 1
第 7 回　Process Writing 1, Essay structure
第 8 回　Unit 1 Review
第 9 回　Process Writing 2, Essay structure
第 10 回 Process Writing 3: Draft 1, Editing and revision
第 11 回 Unit 2 or Unit 3
第 12 回 Unit 2 or Unit 3
第 13 回 Unit 2 or Unit 3
第 14 回 Process Writing 4, Peer review
第 15 回 Unit 2 or Unit 3 Review
第 16 回 Process Writing 5: Final Draft, Editing and revision
第 17 回 RW Project Worksheet 1, Self-assessment 2
第 18 回 RW Project Worksheet 2
第 19 回 RW Project Worksheet 3
第 20 回 Unit 4
第 21 回 Unit 4
第 22 回 Uni 4
第 23 回 Unit 4 Review
第 24 回 Review and final reading test

第 25 回 Final writing test preparation
第 26 回 Final writing test
第 27 回 Review and feedback
第 28 回 Course review, Self-Assessment 3

【成績評価の基準および方法】

1. 授業課題(30%):　産出スキル（15%）（ライティングタスク等）
　　受容スキル（15%）（教科書ユニット、小テスト等）
2. RW プロジェクト(10%): プロジェクトワークシート
3. プロセスライティング(20%): 初稿（5%）、第２稿（5%）、最終稿（10%）
4. 期末リーディングテスト (15%): 実力テスト
5. 期末ライティングテスト (25%): 実力テスト

【教科書・参考書】 Textbook: *Unlock Reading and Writing 3* (Cambridge University Press)

APPENDIX B

SYLLABUS (ENGLISH VERSION)

Course Summary

RW Advanced level

Acquiring English communication skills enables our students to actively participate in a globalized society. The course develops practical reading and writing skills working towards producing a well-developed piece of writing supported by reading activities that enhance the ability to understand English in various genres with writing activities that allow students to express views appropriate to both the writing purpose and audience. In order for students to connect input and output at a deeper level, critical thinking skills will be developed through a variety of focused activities. The goals of this course are **equivalent to the B1 level of the Common European Framework of Reference (CEFR), namely, to become an "independent user of English"**.

In reading, students will have practice in understanding fairly long, complex essays, expository passages, and narratives that include internet-based news articles, reading them quickly (100-120 words per minute depending on the text) with scanning and skimming practices and reading strategies. In writing, to connect input and output, students will produce written work based on information obtained from reading materials, incorporating detail to support views and opinions, while the acquisition of grammatical and vocabulary skills will also be emphasized. Improvements to writing will be made by revising multiple drafts in response to teacher/peer feedback. Throughout the course, students will learn to think critically, and become more able to engage in discussion, group work, and other practical applications of critical thinking skills.

Learning Goals

1. Reading (R)

a. To be able to understand the gist and development of longer, complex texts (expository passages, essays, narratives, etc. about everyday matters), and process important details with occasional dictionary use. (Can-do list: R1, R3)
b. To be able to grasp the necessary information from typical everyday texts (news articles, expository passages, work-related documents, etc.) while using reading strategies to deal with unknown words. (Can-do list: R2, R4, R5, R6)

2. Writing (W)

a. To be able to produce a multiple-paragraph essay with an introduction, body, and a conclusion about one's preferences, including comparing things and giving logical reasons to support their position. (Can-do list: W1, W2)
b. To be able to produce a typical personal/official piece of writing (explanations, letters, e-mails, etc.); through use of proper expressions, sentence structures and typical linking words. (Can-do list: W3, W4, W5)

c. To be able to critically read a paragraph written by oneself or peers, and to make necessary revisions and write a mostly proper response to formulaic memos and e-mails. (Can-do list: W3, RW1, RW2, RW3)

3. Critical Thinking (CT)

a. To be able to employ complex thinking skills such as analyzation, comparison, and judgment. (Can-do list: CT)
b. To think critically in reading and writing activities. (Can-do list: CT)

Class Schedule

Class 1	Orientation
Class 2	Orientation, Writing diagnostic test, Self-assessment 1
Class 3	Unit 1
Class 4	Unit 1
Class 5	Unit 1
Class 6	Unit 1
Class 7	Process Writing 1, Essay structure
Class 8	Unit 1 Review
Class 9	Process Writing 2, Essay structure
Class 10	Process Writing 3: Draft 1, Editing and revision
Class 11	Unit 2 or Unit 3
Class 12	Unit 2 or Unit 3
Class 13	Unit 2 or Unit 3
Class 14	Process Writing 4, Peer review
Class 15	Unit 2 or Unit 3 Review
Class 16	Process Writing 5: Final Draft, Editing and revision
Class 17	RW Project Worksheet 1, Self-assessment 2
Class 18	RW Project Worksheet 2
Class 19	RW Project Worksheet 3
Class 20	Unit 4
Class 21	Unit 4
Class 22	Uni 4
Class 23	Unit 4 Review
Class 24	Review and final reading test
Class 25	Final writing test preparation
Class 26	Final writing test
Class 27	Review and feedback
Class 28	Course review, Self-Assessment 3

Grading Method

1. Coursework (30%): productive skills (15%) e.g., writing tasks; receptive skills (15%) e.g., textbook unit quizzes
2. RW Project (10%): project worksheets
3. Process Writing (20%):
   First draft (5%), Second draft (5%), Final Draft (10%)
4. Final Reading Test (15%) Proficiency test
   Final Writing Test (25%): Proficiency test

Textbook
Unlock Reading and Writing 3 (Cambridge University Press)

APPENDIX C

CONSENT FORM FOR THE ENTIRE STUDY (JAPANESE VERSION)

調査にご協力いただき、ありがとうございます。下記調査概要をお読みいただき、十分にご理解いただいた上で調査協力にご同意いただける場合、署名欄にご署名下さい。

1. 調査の目的:
   この調査は、ライティング（エッセイ作文、複数の段落からなるもの）の上達度を調べるものです。
2. データ収集の方法: 作文、単語の点数。
3. 個人情報の保護:
   調査によって得られたデータは、研究目的に使用し、それ以外の目的では一切使用致しません。また、成績にも一切関係しません。個人名、学校名などには仮名を使用し、協力者の個人情報と匿名性 を保護します。
4. 参加の任意性と撤回の自由:
   調査への参加は協力者の自由意思に基づくものであり、協力者はいつでも同意を撤回できる権利があり、参加を希望しない場合は、データは使用されません。
5. 研究者の連絡先:
   調査に関しご不明の点などございましたら、いつでも下記連絡先までご連絡下さい。

   住所: XXXXX 電話番号: XXXXX Email: XXXXX

私は上記事項を読み、調査概要を理解しました。つきましては、私の自由意思により調査への協力に同意します。

協力者自署　　日付

研究者の自署　　日付

APPENDIX D

CONSENT FORM FOR THE ENTIRE STUDY (ENGLISH VERSION)

Thank you for your cooperation in my research project. If you agree to participate in the research after reading the following statements, please sign your name below.

1. Purpose of research: This research aims to examine writing development in English.
2. Data collection: Writing samples and vocabulary scores.
3. Protection of confidentiality: Data obtained from the participants will be used exclusively for the purpose of this research. This will not affect your grade. All the names including the participants and schools are identified only by pseudonyms, and confidentiality and anonymity of the participants are protected.
4. Voluntary participation and right to withdraw: Participation in this research is voluntary. The participants have the right to withdraw from the research at any time, and the data will not be used in the study if the participants decide not to participate.
5. If you have any questions about the research project, please feel free to contact me at the contact number below at any time.

Address XXXXX Phone XXXXX Email: XXXXX

I have read and understood the statements above. I agree and consent to cooperate voluntarily.

Signature of the participant Date

Signature of the researcher Date

APPENDIX E

CONSENT FORM FOR INTERVIEWEES (JAPANESE VERSION)

調査協力同意書

下記調査概要をお読みいただき、十分にご理
解いただいた上で調査協力にご同意いただける場合、署名欄にご署名下さい。

1. 研究の目的： これは、ライティング（エッセイ作文）の上達度をみます。
2. データ収集の方法： データは、インタビューによって収集し、インタビューは ICレコーダーで録音され ます。
3. 個人情報の保護：調査によって得られたデータは、研究目的に使用し、それ以外の目的では一切使用
   致しません。また、成績にも一切関係しません。個人名、学校名などには仮名を使用し、協力者の秘密と匿名性 を保護します。
4. 参加の任意性と撤回の自由：
   調査への参加は協力者の自由意思に基づくものであり、協力者はいつでも同意を撤回し、データを廃棄させる権利を持っています。
5. 研究者の連絡先：
   調査に関しご不明の点などございましたら、いつでも下記連絡先までご連絡下さい。

   住所: XXXXX 電話番号: XXXXX Email: XXXXX

私は上記事項を読み、調査概要を理解しました。つきましては、私の自由意思により調査へ の協力に同意します。

協力者自署　　日付

研究者の自署　　日付

APPENDIX F

CONSENT FORM FOR INTERVIEWEES

(ENGLISH VERSION)

Thank you for your cooperation in my research project. If you agree to participate in the research after reading the following statements, please sign your name below.

1. Purpose of research: This research aims to examine writing development.
2. Data collection: Data are obtained through interviews, which are recorded using an IC recorder.
3. Protection of confidentiality: Data obtained from the participants will be used exclusively for the purpose of this research. This will not affect your grade. All the names including the participants and schools are identified only by pseudonyms, and confidentiality and anonymity of the participants are protected.
4. Voluntary participation and right to withdraw: Participation in research is voluntary. The participants have the rights to withdraw from the research at any time, and to have the data destroyed.
5. If you have any questions about the research project, please feel free to contact me at the contact number below at any time.

Address XXXXX Phone XXXXX Email: XXXXX

I have read and understood the statements above. I agree and consent to participate in this study.

Signature of the participant Date

Signature of the researcher Date

APPENDIX G

BACKGROUND QUESTIONNAIRE (JAPANESE VERSION)

英語学習についての背景調査

氏名: _____ 日付: _____

年齢: _____ 性別: 男 女 (〇を付けて下さい)

専攻: _____

E-mail アドレス: _____

Section I. これまでのライティング学習について（〇で、囲んでください。）

パラグラフ（段落）で作文を学習したことはありますか？ はい いいえ

複数のパラグラフからなる作文を学習し書いたことはありますか？はい いいえ

Section II: テストスコアについて
以下の各テストについて、もっとも最近の合計スコア（英検は何級を取得したか）およびリーディングセクションのスコア、受験した時期を記入して下さい。

テスト	合計スコア/級 (例: 85)	リーディングセクションのスコア (例: 24)	受験した時期 (例：2018 年 3 月)
TOEFL			
TOEIC			
IELTS			
英検 2			

APPENDIX H

BACKGROUND QUESTIONNAIRE (ENGLISH VERSION)

Your name: _____ Today's Date: _____

Age: _____ Gender: Male Female (Circle one)

Major field of study: _____

Your e-mail address: _____

Section I. Writing experiences in English

Have you studied paragraph writing? Yes No

Have you studied writing multiple paragraphs or essays? Yes No

Section II: Test scores
List your most recent score for any of the following tests.

Test	Total score/Level	Reading section score	When you took the test (Month/Year)
TOEFL			
TOEIC			
IELTS			
EIKEN2			

APPENDIX I

WRITING RUBRICS

	Content	Organization	Vocabulary	Cohesion	Grammar
6	**Extremely well-developed** piece of writing with **specific, engaging details that address the task/topic appropriately and support the main idea very well**. **No irreverent ideas.**	**Excellent** essay/paragraph organization. **Very clearly and appropriate** introduction with a thesis statement, body and conclusion. Purpose of essay **very clearly** stated.	**Excellent and frequent use** of sophisticated words and idiomatic language appropriate to context. **Varied, creative** word choice enhances communication of meaning.	**Excellent** use of connectives, repetitions, pronouns, and synonyms across a text. Each paragraph flows **very** smoothly.	**Accurate** use of **a variety of sentence structures**. **High level of accuracy** in morphology. **No major errors**. Meaning **clearly** communicated.
5	**Fairly well-developed** piece of writing with **some descriptive and expressive** details that **mostly** address the task and support the main idea. **Mostly** relevant ideas but contains a few irrelevances.	**Very good** essay/paragraph organization. **Fairly clear** with appropriate and adequate introduction with a thesis statement, body and conclusion. Purpose **fairly clearly** stated.	**Very good and frequent use** of sophisticated words and idiomatic language appropriate to context. **Generally**, varied, creative word choice enhances communication of meaning.	**Very good** use of connectives, repetitions, pronouns, and synonyms across a text. Each paragraph flows **fairly** smoothly.	**Comlex structures attempted, not always accurately**. **Some errors** with morphology. **Mistakes** do not interfere with meaning. Few fragments.
4	Simple piece of writing with reasonably clear message and **enough basic details** to support the main idea. Ideas are **generally** relevant, and contains **some irrelevances**.	**Good** essay/paragraph organization. **Somewhat c**lear introduction with a thesis statement, body and conclusion. Purpose **somewhat clearly** stated.	**Mostly** high frequency vocabulary and some idiomatic language. **Some inappropriate word choices** and literal translations, but meaning is **understandable**. Some use creative word choice.	**Good** use of connectives, repetitions, pronouns, and synonyms across a text. Each paragraph flows **somewhat** smoothly. **(*First*, *Second*, *Third*, some more are used.)**	**Mostly simple sentences**, but sometimes combined into multiple-clauses sentences. **Some errors** in morphology and word orders. **Mistakes occasionally** interfere with meaning. A few fragments.

	Content	Organization	Vocabulary	Cohesion	Grammar
3	Simple piece of writing with **somewhat unclear** message and **basic details** (3-4 detail sentences) to support the main idea. **Some** ideas **not always** relevant. (Comparison - one side is not written)	**Fair** essay/paragraph organization with **some lack** of a thesis statement, a topic sentence, supporting details, and concluding sentence. Purpose stated **somewhat unclearly** stated. (Comparison - Two are not compared.) **Some lack of genre organization.**	High frequency vocabulary only. **Some** inappropriate word choices and literal translations, and meaning is **somewhat not understandable**. **Occasionally** use creative word choice.	**Fair** use of connectives, repetitions, pronouns, and synonyms across a text. Each paragraph flows **somewhat poorly**. **(*First*, *Second*, and *Third* are used)**	**Mostly simple sentences**. **Frequent errors** with morphology and word orders. **Mistakes sometimes** interfere with meaning. Some sentence fragments.
2	Simple piece of writing, but **lack of details** (less than 3 sentences) to support the main idea. **Ideas not relevant or logically organized.**	Basic paragraph with **many lack of** a thesis statement, a topic sentence, supporting details, and concluding sentence. **Purpose stated unclearly**. (Comparison - Two are not compared.) Many lack of genre organization	**Simple words only. Lack of** variety or creativity with word choice. **Many inappropriate word choices** and literal translations which can confuse intended meaning. **Seldomly** use creative word choice.	**Poor** use of connectives, repetitions, pronouns, and synonyms across a text. Each paragraph flows **poorly**. **(1 or two connectives are used)**	**Mostly simple sentences. Many errors** with morphology and word orders. **Mistakes often** interfere with meaning. Some sentence fragments.
1	**Undeveloped piece** of writing with unclear message, **serious lack** of supporting details. Does not address the task appropriately. Ideas **not** relevant or logically organized.	Basic paragraph with **lack of** a thesis statement, a topic sentence, supporting details, and concluding sentence.	Misuse even of basic vocabulary items; meaning is obscured. **Lack of** variety.	**Very poor** use of connectives, repetitions, pronouns, and synonyms across a text. Each paragraph flows **very poorly**.	**Poor grasp of** simple sentence patterns. **Many errors** with simple sentences. Many sentence fragments. **Meanings obscured.**

APPENDIX J

INTERVIEW PROTOCOL (JAPANESE VERSION)

1. ライティングにおいて難しかったこととは何でしたか？

   （内容、構成、文法、単語、句読点、フォーマット、ジャンルなど）

2. ライティングにおいて、わりと容易にできたことは何でしたか？

   （内容、構成、文法、単語、句読点、フォーマット、ジャンルなど）

3. どのようにして、作文を書く時、単語を選びましたか？

4. 英作文全体で、重点を置いたことは何ですか？

5. 語彙力を増やすさせるために、どのようなことをしましたか？

6. 英語と日本語における作文の書き方の違いに、何か気づきましたか？

APPENDIX K

INTERVIEW PROTOCOL (ENGLISH VERSION)

1. What was difficult and easy when you were learning to write an essay?

   (e.g., content, organization, grammar, vocabulary, punctuation, format, genres)

2. What was easy when you were learning to write an essay?

   (e.g., content, organization, grammar, vocabulary, punctuation, format, genres)

3. How did you choose vocabulary when you write essays?

4. What aspect of writing did you put more efforts when you write an essay?

5. How have you done to improve your vocabulary?

6. What differences did you find any differences in the essay writing structure between English and Japanese essays?

APPENDIX L

TEACHING MATERIALS

Writing Task 1

Brainstorming

List your ideas as you think about the writing topic on comparing living in a big city and a small town.

After you brainstorm, decide a main idea and supporting examples.

Outlining

Please fill in the blanks.

Thesis: _____

Main idea 1: _____

Support 1: _____

Support 2: _____

Support 3: _____

Main idea 2: _____

Support 1: _____

Support 2: _____

Support 3: _____

Main idea 3: _____

Support 1: _____

Support 2: _____

Support 3: _____

Concluding statement: _____

Writing Task 2

Brainstorming

List your ideas as you think about the writing topic on transportation in Tokyo.

After you brainstorm, decide on a main idea and supporting examples.

Writing Task 2

Outlining

Please fill in the blanks.

Thesis: _____

Main idea 1: _____

Support 1: _____

Support 2: _____

Support 3: _____

Main idea 2: _____

Support 1: _____

Support 2: _____

Support 3: _____

Main idea 3: _____

Support 1: _____

Support 2: _____

Support 3: _____

Concluding statement: _____

APPENDIX M

MODEL ESSAYS

Essay Structure

Every essay has three important parts: the introduction, the body and the conclusion. If one part is missing, the essay is not complete. Read the student essay. Then read the information about each part of the essay. Work with a partner. Find a sentence for each feature of the essay (1–9).

Model Essay 1 (Comparison-Contrast Essay) for Writing Task 1

Hawaii and New York

Hawaii and New York are located at opposite sides of U.S.A. Both places are great cities for tourist destinations. It wasn't until last year I went to New York that I found the differences. I feel that Hawaii and New York are different from each other in two main points, including activities and transportation. I prefer Hawaii to New York for next summer vacation.

The first point is that the type of activities is quite different. Hawaii is famous for blue ocean and warm climate, so swimming and hiking are popular. You can have fun with nature. In addition, life moves at a slower pace, so you don't have to rush. On the other hand, New York is known for exciting entertainment, expensive stores, and a business center. You will be busy going to musicals, concerts, or skyscrapers. A life in New York looks busier than Hawaii.

The second point is that the transportation systems of two areas are completely different. Hawaii has only a bus system, but access to beaches is easy. In addition, buses come on time, and they are not crowded. They are comfortable. Bus drivers are friendlier. On the other hand, the Subway in New York is developed, so you can reach your destination faster. However, it looks unsafe and dirty.

In conclusion, I believe that Hawaii and New York are quite different from each other in terms of activities and transportation systems. I like outdoor activities in nature and nice buses, in a relaxing atmosphere, so I prefer to go to Hawaii for summer vacation.

Vocabulary Exercise for Writing Task 1

Please fill in the blanks in English. The first letter of each word and a Japanese translation are provided. Write a word in the parentheses, and multi-words in the brackets.

Hawaii and New York are (l 位置している) at (o 反対の) sides of U.S.A. Both places are great cities for tourist destinations. It wasn't until last year I went to New York that I found the differences. I feel that Hawaii and New York are different from each other in two main points, including (a 活動) and transportation. I (p 好む) Hawaii to New York for next summer vacation.

[T] [f] [p] [１つ目は]is that the type of activities is quite different. Hawaii is (f 有名な) for blue ocean and warm (c 気候), so swimming and hiking are (p 人気の). You can have fun with nature. [I] [a] [さらに] life moves at a slower (p ペース), so you don't have to (r 急ぐ). [O] [t] [o] [h] [一方で], New York is known for exciting entertainment, expensive stores, and a business center. You will be busy going to musicals, concerts, or skyscrapers. A life in New York looks busier than Hawaii.

[T] [s] [p] [2 つ目は]is that the transportation systems of two areas are completely different. Hawaii has only a bus system, [b しかし] access to beaches is easy. [I] [a] [さらに], buses come on time, and they are not (c 混雑している). They are (c 快適な). Bus drivers are friendlier. [O] [t] [o] [h] [一方で], the Subway in New York is (d 発達している), [s そこで] you can reach your (d 目的地) faster. [H しかしながら], it looks unsafe and dirty.

[I] [c] [結論は], I believe that Hawaii and New York are quite different from each other in terms of activities and transportation systems. I like outdoor activities in nature and nice buses, in a relaxing (a 雰囲気), [s そこで] I prefer to go to Hawaii for summer vacation.

Model Essay 2 (Problem-Solution Essay) for Writing Task 2

Traffic Jams

One of the most serious issues facing the majority of large urban areas is traffic jams. The main reason for this is that there are too many private cars on the roads these days and a viable solution is to introduce more park-and-ride schemes.

The number of people owning cars increases year on year, with most families now having more than one car. Most people like the convenience of travelling at the time they want to rather than being restricted to public transport timetables, so they prefer to drive themselves around rather than taking the bus or train. This is despite the fact that they frequently have to sit in long traffic lines as they near the city center.

A solution that is proving successful in many areas is park-and-ride schemes. This is where you park your car for free in a large car park on the outskirts of the city and take a bus for the final part of your journey. The fee you have to pay for the bus trip is usually very small and this public transport system is generally very regular, running every ten minutes or so. A survey carried out in the city of Exeter showed that the rush hour congestion decreased by 10% when the council set up a park-and-ride scheme to the north of the city.

To conclude, the major organ problem of traffic congestion caused by the excessive number of private cars on city roads can be partly alleviated by the introduction of park-and-ride systems on city fringes. I would certainly use one if it was introduced in my area.

Vocabulary Exercises for Writing Task 2

Please fill-in-the blanks in English. The first letter of each word and a Japanese translation are provided. Please write a word in the parentheses, and multi-words in the bracket.

Traffic Jams

[O] [o] [the] [m] [s] [i] [最も深刻な論点の一つに] facing the (m 大多数の) of large (u 都市の). areas is traffic jams. [T] [m] [r] [主な理由は]is for this is that there are too many private cars on the roads these days and a viable solution is to introduce more park-and-ride (s 事業計画).

The number of people owning cars increases year on year, with most families now having more than one car. Most people like the convenience of travelling at the time they want to rather than being (r 限られた) to public transport timetables, [s] [だから]they prefer to drive themselves around rather than taking the bus or train. This is despite the fact that they frequently have to sit in long traffic lines as they near the city center.

A (s 解決策) that is (p 証明する) successful in many areas is park-and-ride schemes. This is where you park your car for free in a large car park on the (o 郊外) of the city and take a bus for the final part of your journey. The (f 料金) you have to pay for the bus trip is usually very small and this (p 公共の). transport system is generally very (r 規則的な), running every ten minutes or so. A survey carried out in the city of Exeter showed that the rush hour (c 混雑) (d 減少する) by 10% when the (c 評議会) set up a park-and-ride scheme to the north of the city.

[T] [c] [結論は], the major organ problem of traffic congestion caused by the (e 過剰な) number of private cars on city roads can be partly (a 軽減される) by the introduction of park-and-ride systems on city fringes. I would certainly use one if it was introduced in my area.

APPENDIX N

REIVISION CHECKLIST (Oshima & Hogue, 2007)

	Essay Checklist	YES	NO
1	My essay is in the correct format (title centered, first line of every paragraph indented, double-spaced.		
2	My essay fits the assignment.		
3	I wrote an introduction.		
4	The last sentence of my introduction is my thesis statement.		
5	In each body paragraph, I discuss a different aspect of my main topic.		
6	Each body paragraph begins with a topic sentence that tells which aspect it will discuss.		
7	I used supporting details in each paragraph.		
8	I wrote "in conclusion."		
9	Each paragraph discusses only one idea, and there are no sentences that are off the topic.		
10	Each paragraph flows smoothly from beginning to end.		
11	I repeat key nouns and pronouns.		
12	I use transition signals to show relationships among ideas.		
13	I use transition signals to link paragraphs.		
14	I sometimes use a difficult word as well as various words.		
15	I checked my punctuation, capitalization, and spelling.		
16	I varied my sentence structure by writing simple, compound, and complex sentences.		
17	I checked my paragraph for fragments, comma splices, and run-ons.		

APPENDIX O

FACETS COMMAND FILE

```
title = YAMAAI FACETS COMMAND
output measures file = prevla.out; the output file
convergence = 0.1                ; size of largest
remaining marginal score residual at convergence
unexpected = 3.0                 ; size of smallest
standardized residual to report
arrange = m
facets = 6                       ; 5 facets
noncenter = 1                    ; participant facet floats
Inter-rater = 1
Missing = *
positive = 1,2,3                 ; greater score => greater
measure
usort = 1,2,3,4,5                ; sort residuals
Model=?B,?B,?B,?B,?B,?B,R5
Vertical = 1A,2A,3A,4A,5AS

Labels=
1,Rater
1-6
*
2,Students
 1-65
*
3,Task,A
 1=,0.0
 2=,0.0
*
4,Draft
 1-3
*
5,Group
 1-2
*
6,Criteria
1 = Content
2 = Organization
3 = Vocabulary
4 = Cohesion
5 = Grammar
*
```

```
data
1,1,1,1,1,1-5,1,2,2,1,1
1,2,1,1,1,1-5,1,1,2,2,2
1,3,1,1,1,1-5,2,1,2,2,2
1,4,1,1,1,1-5,1,1,2,2,2
1,5,1,1,1,1-5,2,2,2,2,2
1,6,1,1,1,1-5,1,2,1,2,1
1,7,1,1,1,1-5,1,2,1,1,1
1,8,1,1,1,1-5,3,3,3,3,4
1,9,1,1,1,1-5,2,2,2,2,2
1,10,1,1,1,1-5,2,2,2,2,2
1,11,1,1,1,1-5,2,3,3,3,3
1,12,1,1,1,1-5,2,2,2,2,2
1,13,1,1,1,1-5,2,2,2,2,3
1,14,1,1,1,1-5,1,2,1,1,1
1,15,1,1,1,1-5,2,2,2,2,2
1,16,1,1,1,1-5,1,2,1,2,1
1,17,1,1,1,1-5,1,1,1,1,1
1,18,1,1,1,1-5,2,2,2,2,2
1,19,1,1,1,1-5,3,3,3,3,3
1,20,1,1,1,1-5,1,2,1,1,1
1,21,1,1,1,1-5,1,1,2,1,1
1,22,1,1,1,1-5,2,1,2,2,2
1,23,1,1,1,1-5,1,2,2,2,1
1,24,1,1,1,1-5,1,2,1,2,1
1,25,1,1,1,1-5,1,1,2,2,1
1,26,1,1,1,1-5,2,2,2,2,2
1,27,1,1,1,1-5,1,1,1,2,2
1,28,1,1,1,1-5,3,2,2,2,2
1,29,1,1,1,1-5,2,2,2,2,2
1,30,1,1,1,1-5,1,2,1,2,1
Data set truncated
```