Validating Written Accuracy Measures in Second Language Writing

A Dissertation Submitted to the University of Tsukuba
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Linguistics

Hideaki OKA

2022

Abstract of the Dissertation

Validating Written Accuracy Measures in Second Language Writing

By

Hideaki OKA

Writing is an essential activity not only for social purposes (e.g., writing e-mails) but also for academic purposes (e.g., publishing research papers) because writing enables us to better convey our ideas, and to a wider audience, even globally. Writers must think of various aspects of written works such as logicalness and composition. These aspects account for the superior quality of the written texts. However, if their written language was inaccurate, they might lose their credibility. Therefore, the ability to write accurately in English is one of the most important skills in the international community because it enables the writers to convey ideas effectively and precisely to both native speakers and those who learn English as a second language.

English teachers require to develop students' writing skills (e.g., clear compositions) in classrooms (Weigle, 2002) and the ability to write accurately. Various teaching methods aid in developing writing accuracy of students such as written corrective feedback (e.g., Van Beuningen et al., 2012). Some previous studies have reported that though the method of corrective feedback is effective in developing writing accuracy, its effects would vary depending on the skills of each individual student (e.g., Hyland & Hyland, 2019). Therefore, teachers need to accurately evaluate the accuracy of each student. Furthermore, by identifying the developmental patterns of learners' language development and writing proficiency, we can consider tailoring instructions to

learners at a certain proficiency level. In fact, several studies have been conducted in recent years (e.g., Barrot & Adgeppa, 2021). It is clear that accurate measurement is very important in capturing developmental patterns. Therefore, assessing accuracy in learners' writing performance is fundamental for capturing the accuracy development and choosing appropriate instructions that will subsequently help teachers choose the most suitable instructional strategies (e.g., corrective feedback) to improve their students' writing.

Recently, L2 writing studies have evaluated writing performance based on the framework of complexity, accuracy, and fluency (CAF). *Accuracy* is defined as "the ability to produce target-like and error-free language" (Housen et al., 2012, p. 2). Furthermore, accuracy, complexity, and fluency have been used as performance descriptors for writing assessment and as indicators of writing proficiency (Housen & Kuiken, 2009).

While accuracy can be measured by accuracy indices based on writing performance (e.g., the number of error-free clauses per the total number of clauses: EFCR), these indices do not consider the gravity of errors, which is the degree to which they inhibit a reader's comprehension. Wigglesworth and Foster (2008) criticized this limitation and proposed an alternative accuracy index: the weighted clause ratio (WCR). WCR scores are awarded to each clause based on a rating scale that consists of four levels (e.g., Lv.1 and Lv.2) showing the degree to which errors in the clauses compromise a reader's comprehension. By using the WCR, English as a foreign languahe (EFL) teachers can not only assess accuracy in detail but also focus more on clauses that heavily hinder reader comprehension. For researchers, WCR assessment not only provides a finer measure of a learner's writing accuracy but also provides a detailed picture of their developmental changes.

While the WCR has been used in the second language writing studies (e.g., Polio & Shea, 2014), the WCR accuracy assessments and the use remain subject to issues. Only a few recent studies (e.g., Foster & Wigglesworth, 2016; Wigglesworth & Foster, 2008) have proposed the WCR, so its validity is still scarce. Although some studies conducted a validation study of the accuracy assessment of the WCR (e.g., Evans et al., 2014), the studies applied the traditional validity frameworks and focused only on the correlation between the WCR and other writing accuracy measures (e.g., EFCR). It would be difficult to ascertain whether these studies provide enough evidence for confirming the validity of the WCR.

Moreover, studies about accuracy development focusing on Japanese English learners have not been conducted. While the studies focus on specific linguistic errors such as subject-verb agreement (e.g., Abe, 2017), these studies do not report on the writing accuracy of Japanese EFL students. The WCR would be able to provide not only the development patterns of writing accuracy but also offer small changes in the writing accuracy in detail as the WCR can divide all clauses into four levels.

Therefore, this dissertation has two main studies: a validation study for the WCR and the accuracy development study using the WCR. By showing the validity of the WCR, researchers can use the WCR and get reliable results when they conduct a development study using the corpus. In addition, the dissertation can provide English teachers with suggestions about the instructions for writing accuracy. By capturing specific changes across proficiency levels, English teachers can effectively distribute their time and would be able to develop students' accuracy efficiently.

As for the validation study of the WCR, the main research question posed in the dissertation is: Is the WCR valid for examining the writing accuracy development of Japanese EFL learners using the learner corpus? To answer this question, the present study set the sub-research questions in each of the studies (from Study 1 to Study 5). Furthermore, the dissertation focused on the accuracy development using the WCR and set the main research question: How does the WCR change as the levels in English proficiency change?

In this dissertation, the validation study for the WCR was conducted first (Chapters 3, 4, and 5). The data in the dissertation was derived from The International Corpus Network of Asian Learners of English (ICNALE) corpus made by Ishikawa (2013), and 100 Japanese EFL learners were randomly selected.

In Study 1, the reliability of the writing accuracy measures (e.g., WCR) are examined. This study accounts for the confirmation of the evaluation and generalization inference. The present study used Cronbachs'α to examine the consistency of the evaluations. In addition, based on the generalizability theory (G-study and D-study), an investigation was done to determine which measurement errors (e.g., rater effects) would influence the score variances in the writing accuracy measures scores. The analysis showed that both Cronbach's α value of the WCR and the traditional accuracy measures were high. Furthermore, the G-study showed that the G coefficient of the WCR was also high (G = .91), although some measurement errors affected the score variances. Based on the results, the scoring and generalization inference were confirmed positively.

In Study 2, the present study used an exploratory factor analysis (EFA) and examined a factor that the WCR would reflect in order to confirm the explanation inference. The present study also explored the relationships between the WCR and textual features (e.g., the number of clauses). The EFA showed that the WCR reflected the same construct that the traditional writing accuracy measures reflected. Moreover, the correlation between the accuracy and complexity constructs was low (r = .08), and the WCR did not highly correlate with textural features (e.g., clauses). These results suggested that the explanation inference was confirmed positively.

In Study 3, the correlations between the WCR and the English proficiency levels (i.e., Common European Framework of Reference for Languages: CEFR levels) set in the ICNALE corpus were examined to confirm the Extrapolation inference. To do so, the present study used a correlation analysis, which can examine the non-linearity of the relationships. The correlation analysis showed that the WCR correlated more strongly with the English proficiency levels than with the traditional accuracy measures. Therefore, the extrapolation inference was confirmed positively.

Furthermore, to confirm the utilization inference, the present study compared the amount of information that the writing accuracy measures can provide. Firstly, the descriptive statistics and non-parametric statistics (Kruskal-Wallis test: K-W test) were used to show the amount of information provided by the writing accuracy measures. Through the three studies, the present study aimed to confirm the validation of the WCR used for the accuracy development study in the corpus. These analyses showed that the WCR was capable of providing much information about the degree of accuracy in and among groups. These results suggested that the utilization inference was confirmed positively.

In Study 4, the accuracy development and the relationships among the CAF measures have been examined. The present study used the WCR and EFCR, and compared the differences in the development patterns. It also examined the manner in which the complexity and fluency measures also changed as the WCR score changed. The KW test showed that the WCR captured the small changes in the writing accuracy made by Japanese EFL learners better than the EFCR. In addition, the results showed that the scores of complexity and fluency measures significantly increased in the intervals where the WCR score did not increase. Therefore, the writing accuracy measured by the

WCR improved as the English proficiency levels increased, although other dimensions were likely to have affected the development.

In Study 5, the study focused on the changes in the number of clauses included in the WCR rating scale. Although the detailed explanation was written in Chapter 2, the WCR divided the clauses into four types according to the influence of linguistic errors (e.g., Lv.1). The KW test was used and showed which categories of the clauses changed among the CEFR levels. The analysis revealed that some clause types were significantly different. In particular, all clause types between A2 and B2+ were significantly different.

In conclusion, the dissertation demonstrated the importance of the WCR in investigating writing accuracy development. In addition, the dissertation provided suggestions for English classes by assessing the development patterns of writing accuracy. While there are some limitations, this study expands the measurement theory of writing accuracy in the CAF framework, and provides new insights into accuracy development in Japanese EFL learners.

Acknowledgments

I would like to express my gratitude to all those who kindly supported me. Without their help, I would not have been able to complete this doctoral dissertation.

First, I would like to express my gratitude to Professor Akiyo Hirai, my academic supervisor. She guided me in pursuing my research and showed me how enjoyable it can be. She taught me the necessary knowledge and skills, including how to write papers and conduct research. Additionally, she gave me constructive feedback on my statistical analysis and led me to better interpret the results. Her guidance was instrumental in helping me complete this dissertation.

Second, I would like to express my gratitude to Professor Hirosada Iwasaki at the University of Tsukuba. He provided numerous important suggestions regarding the content of the study such as the curriculum guidelines, writing research, and corpus research, which were crucial in deepening the academic position and significance of this paper.

Third, I would like to express my gratitude to Professor Yuichi Ono at the University of Tsukuba. He provided many useful suggestions for this paper from the theoretical and philosophical perspectives. He pointed out things I had missed in previous studies and helped me improve my interpretation of the results.

Next, I would like to express my gratitude to Professor Hidetoshi Saito, my external supervisor, at Ibaraki University. He provided diverse useful suggestions, from the lack of prior research to the conclusions. His opinions helped me improve the paper. I deeply wish to thank him.

Moreover, I would like to thank members of my seminar, Ms. Angelina Kovalyova, Mr. Kohei Funaki, Mr. Toshihide O'ki, Mr. Yoshishige Suda, Mrs. Yukimi Hayahune, and

Mr. Yusuke Kubo. They were very enthusiastic in their consideration of my research and offered many valuable opinions.

I would also like to say thank you to my precious colleagues, Mr. Hiroki Maeda, Ms. Tomoko Ogiso, and Mr. Yamato Sasaki. Their opinions were crucial in deepening the content of my thesis. They also supported me mentally throughout the completion of my dissertation.

Finally, I would like to thank my family for giving me an environment enabling me to study and for supporting me in my doctoral research. They kindly supported me whenever my research was not going well. I am convinced that they are the last piece to complete this doctoral thesis. I could not thank them enough, and I hope to return the favor in the future.

Hideaki OKA

Table of Contents

Section Page

Abstract.......... i
Acknowledgment.......... vii
Table of Contents.......... ix
List of Tables.......... xiii
List of Figures.......... xvi
List of Abbreviation.......... xviii
Publications.......... xix

Chapter 1 Introduction

1.1 Background of the Current Research.......... 1
1.2 Organization of the Dissertation.......... 3

Chapter 2 Literature Review

2.1 Writing Models and Skills.......... 6
2.1.1 The Necessity of Second Language Writing.......... 6
2.1.2 Theories of the Writing Process.......... 7
2.1.3 Theoretical Models About Writing Abilities.......... 10
2.1.4 Assessing Written Abilities in L2 Writing.......... 11
2.2 Complexity, Accuracy, and Fluency (CAF).......... 13
2.2.1 The History and Definition of CAF.......... 13
2.2.2 Importance of Accuracy and Difficulty of Developing Accuracy in Writing.......... 16
2.2.3 The Measurement of CAF.......... 19
2.2.4 The New Accuracy Measure: Weighed Clause Ratio.......... 26
2.2.5 The Discussion About Written Accuracy Measurement.......... 29
2.3 Corpus.......... 32
2.3.1 Definition and History.......... 32
2.3.2 The Limitation of Developmental Studies of CAF and Measurement Issues of WCR.......... 37
2.4 Theory of Language Testing.......... 39
2.4.1 Construct and Measurement in L2 Writing.......... 39
2.4.2 The Introduction of Reliability and Validity in Language Testing.......... 42
2.4.3 Overview of Reliability and its Approaches.......... 43
2.4.4 Overview of Validity and its Approaches.......... 49

2.4.5 Approaches to Validation 57
2.4.6 Measuring Accuracy Based on Argument-Based Validation 63
2.4.7 The Interpretive Arguments in the Validation Study 66
2.5 The Purposes of the Present Dissertation 73

Chapter 3

Study 1: Investigating the Reliability of Accuracy Assessment With Accuracy Measures

3.1 Evaluation and Generalization Inferences 76
3.1.1 Purposes and Research Questions 76
3.1.2 Method 78
3.1.2.1 Participants 78
3.1.2.2 Data Collection 79
3.1.2.3 Traditional Written Accuracy Measures 80
3.1.2.4 Raters for the WCR 80
3.1.2.5 Revising the WCR Rating Scale and Training Raters 81
3.1.2.6 Scoring 82
3.1.2.7 Data Analysis 83
3.1.3 Results 84
3.1.3.1 Descriptive Statistics 84
3.1.3.2 Inter-Rater Reliability 87
3.1.3.3 G-Study: Estimating Variance Components 89
3.1.3.4 D-Study: Simulating the G Coefficients 101
3.1.4 Discussion 101
3.1.5. Conclusion 108

Chapter 4

Study 2: Investigating the Factors Which Accuracy Measures Would Reflect

4.1 Explanation Inference 109
4.1.1 Purposes and Research Questions 109
4.1.2 Method 111
4.1.2.1 Participants 111
4.1.2.2 Written Accuracy Measures 111
4.1.2.3 Complexity Measures 112
4.1.2.4 Textual Features 114
4.1.2.5 Scoring 114
4.1.2.6 Data Analysis 115

4.1.3 Results 118
4.1.3.1 Descriptive Statistics 118
4.1.3.2 Results in EFA 119
4.1.3.3 Results in the Correlation Analysis 123
4.1.3 Discussion... 127
4.1.4 Conclusion of Study 2 131

Chapter 5
Study 3: Investigating the Relationships Between WCR and CEFR and Interpreting the Usefulness of the WCR

5.1 The Extrapolation and Utilization Inferences.. 134
5.1.1 Purposes and Research Questions 134
5.1.2 Method.. 137
5.1.2.1 Participants 137
5.1.2.2 Written Accuracy Measures.. 137
5.1.2.3 Complexity Measures 139
5.1.2.4 Scoring.. 139
5.1.2.5 Data Analysis 140
5.1.3 Results 141
5.1.4 Discussion... 151
5.1.5 Conclusion 155
5.1.5.1 The Summary of Study 3 156
5.1.5.2 The Summary of Validation Study 157

Chapter 6
Study 4: Investigating the Development of Written Accuracy

6.1 The Relationships Among CAF Measures 161
6.1.1 Purposes and Research Questions 161
6.1.2 Method.. 162
6.1.2.1 Participants 162
6.1.2.2 Written Accuracy Measures.. 162
6.1.2.3 Complexity Measures 163
6.1.2.4 Fluency Measures 164
6.1.2.5 Scoring.. 165
6.1.2.6 Data Analysis 166
6.1.3 Results 166
6.1.4 Discussion... 187

6.1.5 Conclusion 191

Chapter 7

Study 5: Examining the Development Patterns of Categories in the WCR Rating Scale

7.1 Overview of Study 5 193
7.1.1 The Summary of Study 4 193
7.1.2 Purposes and Research Questions 193
7.2.1 Method 194
7.2.1.1 Participants 194
7.2.1.2 WCR As a Written Accuracy Measure 194
7.2.1.3 Scoring 195
7.2.1.4 Data Analysis 196
7.2.3 Results 196
7.2.4 Discussion 204
7.2.5 Conclusion of Study 5 206

Chapter 8 General Discussion

8.1 Overview of Findings and General Discussion 208

Chapter 9 Conclusion

9.1 Summary of Findings in the Present Study 228
9.2 Implications for Researchers and Teachers 229
9.3 Limitations 231

References 233

Appendix 248
The Rating Scale of the WCR Developed by the Present Study 248

List of Tables

Table Page

Chapter 2

2.1 Rating Scale of a Weighted Clause Ratio 27
2.2 Example of Evaluation 47
2.3 Summary of Error Sources 48
2.4 Interpretive Argument in the Validation Study 71
2.5 Summary of the Dissertation 74

Chapter 3

3.1 Description of Traditional Written Accuracy Measures 80
3.2 Factors and Interpretations of Variance Components 84
3.3 Descriptive Statistics for All Accuracy Measures in PTJ 85
3.4 Descriptive Statistics for All Accuracy Measures in SMK 86
3.5 Inter-Rater Reliability and Adjusted Cronbach's α in PTJ 87
3.6 Inter-Rater Reliability and Adjusted Cronbach's α in SMK 88
3.7 Estimated Variance Components and Percentage of Variabilities in EFC Score 89
3.8 Estimated Variance Components and Percentage of Variabilities in EFT Score 90
3.9 Estimated Variance Components and Percentage of Variabilities in EFCR Score 91
3.10 Estimated Variance Components and Percentage of Variabilities in EFTR Score 93
3.11 Estimated Variance Components and Percentage of Variabilities in EFSR Score 94
3.12 Estimated Variance Components and Percentage of Variabilities in EFT/S Score 95
3.13 Estimated Variance Components and Percentage of Variabilities in EFT/W Score 96
3.14 Estimated Variance Components and Percentage of Variabilities in EFC/T Score 97
3.15 Estimated Variance Components and Percentage of Variabilities in WEFC/WC Score 99
3.16 Estimated Variance Components and Percentage of Variabilities

in WCR Score 100
3.17 G Coefficient Values for Each Written Accuracy Measure 101

Chapter 4

4.1 Rating Scale of a Weighted Clause Ratio 112
4.2 Description of Written Accuracy Measures.. 112
4.3 Description of Written Complexity Measures 113
4.4 Summary of Kaiser-Meyer-Olkin Measure 116
4.5 Descriptive Statistics for Accuracy and Complexity Measures............................ 118
4.6 Summary of Factor Loading With Accuracy and Complexity Measures in EFA.. 122
4.7 Descriptive Statistics of Words in Essays... 123
4.8 Descriptive Statistics of Clauses in Essays... 124
4.9 Descriptive Statistics of T-Units in Essays 124
4.10 Descriptive Statistics of Sentences in Essays 125
4.11 Correlation Values... 127

Chapter 5

5.1 Rating Scale of a Weighted Clause Ratio 138
5.2 Description of Written Accuracy Measures.. 138
5.3 Description of Written Complexity Measures 139
5.4 Descriptive Statistics of the Number of Words 142
5.5 Descriptive Statistics for Measures 143
5.6 Correlations Between Measures and CEFR Levels.. 145
5.7 Descriptive Statistics of Written Accuracy Measures... 146
5.8 Summary of the Results of the K-W Test in the EFTR 148
5.9 Summary of the Results of the K-W Test in the EFCR.. 149
5.10 Summary of the Results of the K-W Test in the EFT/W 150
5.11 Summary of the Results of the K-W Test in the WCR 151
5.12 Interpretive Argument in the Validation Study and the Results 158

Chapter 6

6.1 Rating Scale of a Weighted Clause Ratio 163
6.2 Description of Written Complexity Measures 164
6.3 Description of the Fluency Measures 164
6.4 Descriptive Statistics of the Number of Words 167
6.5 Descriptive Statistics for Measures 167
6.6 Descriptive Statistics for Measures in Each CEFR Level 169

6.7 Summary of the Results of K-W Test in WCR 172
6.8 Summary of the Results of K-W Test in EFCR 173
6.9 Summary of the Results of K-W Test in VP/T 175
6.10 Summary of the Results of K-W Test in DC/C 176
6.11 Summary of the Results of K-W Test in C/S 178
6.12 Summary of the Results of K-W Test in MLC 179
6.13 Summary of the Results of K-W Test in MLT 181
6.14 Summary of the Results of K-W Test in MLS 182
6.15 Summary of the Results of K-W Test in W/Tx 184
6.16 Summary of the Results of K-W Test in C/Tx 185
6.17 Summary of the Results of K-W Test in T/Tx 187

Chapter 7

7.1 Rating Scale of a Weighted Clause Ratio 195
7.2 Descriptive Statistics of the Number of Words 197
7.3 Descriptive Statistics for the WCR Score in Each CEFR Level 197
7.4 Descriptive Statistics of WCR 199
7.5 Summary of the Results of the K-W Test for the Accurate Clauses 201
7.6 Summary of the Results of K-W Test for the Lv.1 Clauses 202
7.7 Summary of the Results of K-W Test for the Lv.2 Clauses 203
7.8 Summary of the Results of K-W Test for the Lv.3 Clauses 204

List of Figures

Figure Page

Chapter 2

2.1 U-Shaped Development 17
2.2 Summary of Accuracy Measures 22
2.3 Dart Analogy in Reliability and Validity 42
2.4 Relationships Between Constructs and Variables 52
2.5 Three Inferences in Kane et al. (1999) 58
2.6 Structure of an Argument 59
2.7 Framework of Argument-Based Approach in Chapelle et al. (2008) 61

Chapter 3

3.1 Summary of Evaluation Inference 76
3.2 Summary of Generalization Inference 77

Chapter 4

4.1 Summary of Explanation Inference 109
4.2 Scree Plot 120
4.3 Parallel Analysis 121
4.4 Relationships Among Measures 126

Chapter 5

5.1 Summary of Extrapolation Inference 135
5.2 Summary of Utilization Inference 136
5.3 Graph of Correlations 144
5.4 Changes in EFTR Score 147
5.5 Changes in EFCR Score 147
5.6 Changes in EFT/W Score 147
5.7 Changes in WCR Score 147

Chapter 6

6.1 Changes in WCR Score 171
6.2 Changes in EFCR Score 172
6.3 Changes in VP/T Score 174
6.4 Changes in DC/C Score 176

6.5 Changes in C/S Score 177
6.6 Changes in MLC Score.. 179
6.7 Changes in MLT Score 180
6.8 Changes in MLS Score 182
6.9 Changes in W/Tx Score 183
6.10 Changes in C/Tx Score 185
6.11 Changes in T/Tx Score 186

Chapter 7

7.1 Changes in the A2 Level Group 200
7.2 Changes in the B1_1 Level Group 200
7.3 Changes in the B1_2 Level Group 200
7.4 Changes in the B2+ Level Group 200

List of Abbreviation

Abbreviation	Meaning
CAF	Complexity, accuracy, and fluency
CEFR	Common European Framework of Reference for Languages
C/S	Clauses per total of all sentence
C/Tx	Clauses per text
DC/C	Dependent clauses per total of all clause
EFC	Error-free clauses
EFL	English as a foreign language
EFT	Error-free T-units
EFCR	Error-free clauses per total of all clause
EFC/T	Error-free clauses per total of all T-unit
EFSR	Error-free sentences per total of all sentence
EFTR	Error-free T-units per total of all T-unit
EFT/S	Error-free T-units per total of all sentence
EFT/W	Error-free T-units per total of all word
MLC	Mean length of clause
MLS	Mean length of sentence
MLT	Mean length of T-unit
T/Tx	T-units per text
VP/T	Verb phrases per total of all T-unit
WCR	Weighted clause ratio
WEFC/WC	Words in error-free clauses per total of all word in clauses
W/Tx	Words per text

Publications

The part of the dissertation was reported in the publications listed below. In addition, the part of Study 1 (Chapter 3) was reported in "Reliability and optimal designs for measuring accuracy in L2 writing with a weighted clause ratio" as the published paper. The copyright belongs to the Japan Society of English Language Education.

岡秀亮. (2020). 「L2ライティングにおける正確性指標－指標間の関係性に着目してー」．日本言語テスト学会第23回 (遠隔) 全国大会 (2020年12月).

Oka, H. (2021). Reliability and optimal designs for measuring accuracy in L2 writing with a weighted clause ratio. *ARELE: Annual Review of English Language Education in Japan*, 32, 49–64.

Oka, H. (2021). Reliability of the weighted clause ratio for measuring L2 writing accuracy: A generalizability theory approach. *AAAL: American Association for Applied Linguistics*, Virtual conference (March, 2021).

Chapter 1

Introduction

1.1 Background of the Current Research

Writing is essential not only for society (e.g., writing an e-mail) but also for academic situations (e.g., publishing research papers), because it enables us to convey our ideas to other people all over the world. Writers must consider all aspects of written work, such as logicalness and composition. While these aspects are important for high-quality written texts, such texts would not make sense if the written language were inaccurate. Therefore, the ability to write accurately in English is one of the most important abilities within the international community because it enables writers to convey ideas effectively and precisely to both native speakers and English language learners.

However, as English teachers need to develop students' writing skills (e.g., clear compositions) in their classrooms (Weigle, 2002), the ability to write accurately should also be developed in English classrooms. To do so, there are many types of teaching methods to develop written accuracy, such as written corrective feedback (e.g., Van Beuningen et al., 2012). Although some previous studies have reported that such feedback would be effective in developing written accuracy, the effects would differ according to the individual factors such as English proficiency (e.g., Hyland & Hyland, 2019). Furthermore, by identifying the developmental patterns of learners' language development and writing proficiency, we can consider tailoring instruction to learners at a certain proficiency level. Therefore, teachers need to accurately evaluate each student's ability in terms of accuracy and know the development patterns of written accuracy to decide an appropriate instruction.

Assessing written accuracy in learners' writing performance is therefore fundamental for capturing accuracy development and choosing appropriate instructions. By doing so, teachers can subsequently choose the most suitable instructional strategies (e.g., corrective feedback) to improve their students' writing.

While there are many kinds of evaluation methods (e.g., holistic rubrics) for measuring written accuracy in L2 writing, these have some limitations, such as the vagueness of scores and time (Weigle, 2002). Recently, L2 writing studies have evaluated writing performance based on the framework of complexity, accuracy, and fluency (CAF). *Accuracy* is defined as "the ability to produce target-like and error-free language" (Housen et al., 2012, p. 2). Furthermore, accuracy, including complexity and fluency (i.e., CAF), has been used as a performance descriptor for written assessments and as an indicator of writing proficiency (Housen & Kuiken, 2009).

While accuracy can be measured by measures based on written performance (e.g., the number of error-free clauses per total number of clause: EFCR), these measures do not consider the gravity of errors, which is the degree to which they influence a reader's comprehension. Wigglesworth and Foster (2008) criticized this limitation and proposed an alternative accuracy measure: *weighted clause ratio* (WCR). WCR scores are awarded to each clause based on a rating scale that consists of four levels (e.g., Lv.1 and Lv.2) indicating the degree to which errors in clauses compromise a reader's comprehension. By using the WCR, EFL teachers can not only assess accuracy in detail, but also focus more on clauses that significantly hinder reader comprehension. For researchers, WCR assessment not only provides a finer measure of a learner's written accuracy but also a detailed picture of their developmental changes.

While WCR has been used in second language writing studies (e.g., Polio & Shea, 2014) and language development studies in recent years (e.g., Barrot & Adgeppa, 2021), issues remain in terms of WCR accuracy assessments and the use. Recent studies (e.g., Foster & Wigglesworth, 2016; Wigglesworth & Foster, 2008) have proposed the WCR; investigations on validity remain scarce. It is clear that appropriate measurement is very important for inferring the degree of an ability. Although some studies conducted a validation study of the accuracy assessment of the WCR (e.g., Evans et al., 2014), these applied traditional validity frameworks and focused only on the correlation between the WCR and other written accuracy measures (e.g., EFCR). It would be difficult for these studies to provide sufficient evidence to confirm the validity of WCR.

Moreover, studies on accuracy development focusing on Japanese English learners have not yet been conducted. While research has focused on specific linguistic errors such as subject-verb agreements (e.g., Abe, 2017), how accurately Japanese EFL students can write has not been reported. WCR can provide not only the development patterns of written accuracy but also small changes in written accuracy in detail because it can divide all clauses into four levels.

Therefore, this dissertation has two main studies: a validation study for the WCR, and an accuracy development study using the WCR. By demonstrating the validity of the WCR, researchers can use it and obtain reliable results when they conduct a development study using a corpus. In addition, this dissertation can provide English teachers with suggestions regarding instructions for written accuracy. By capturing specific changes across proficiency levels, English teachers can effectively distribute their time and efficiently develop students' accuracy.

1.2 Organization of the Dissertation

Regarding the validation study of the WCR, this dissertation examines the following main research purpose: Is the WCR valid for examining the written accuracy development of Japanese EFL learners using a learner corpus? To answer the main research question, the present study used the International Corpus Network of Asian Learners of English (ICNALE), developed by Ishikawa (2013), and the set the research questions for each study (from Studies 1 to 5). Furthermore, the dissertation focused on accuracy development using the WCR and set the following main research purpose: How does the WCR change as English proficiency levels increase?

To answer these questions, this dissertation consists of the following eight chapters: Introduction (Chapter 1), Literature Review (Chapter 2), Study 1 (Chapter 3), Study 2 (Chapter 4), Study 3 (Chapter 5), Study 4 (Chapter 6), Study 5 (Chapter 7), General Discussion (Chapter 8), and Conclusion (Chapter 9).

Chapter 2 reviews previous studies related to the present study. For second language writing, the theory of writing models, frameworks of writing skills, and evaluation of writing skills are reviewed. Chapter 2 also reviews the theory and practice of validation (e.g., the construct of validity). The findings and limitations of this study are summarized at the end of this chapter.

A validation study for the WCR was first conducted (Chapters 3, 4, and 5). In Chapter 3, the reliability of written accuracy measures (e.g., WCR) is examined. The evaluation and generalization inference were taken into account. Cronbach's α was used to examine the consistency of the evaluations. In addition, the generalizability theory (G-Study and D-Study) was adopted to investigate which measurement errors (e.g., rater effects) influence the score variances in the written accuracy measure scores.

In Chapter 4, factor analysis was adopted and the factors that the WCR would reflect were examined to confirm the explanation inference. In addition, correlations between WCR and complexity measures were examined. The relationship between WCR and textual features (e.g., number of clauses) are also explored.

In Chapter 5, the correlations between the WCR and English proficiency levels (i.e., Common European Framework of Reference for Languages: CEFR levels) set in the ICNALE corpus are examined to confirm the extrapolation inference. To do so, a correlation analysis, which can examine nonlinear relationships, was adopted. Furthermore, to confirm the utilization inference, the amount of information provided by written accuracy measures was compared. First, descriptive and non-parametric statistics (Kruskal-Wallis test) were used to show the amount of information provided by the written accuracy measures. Through these three studies, the present study confirmed the validation of the WCR used for the accuracy development study in the corpus.

Chapter 6 examines the development of accuracy and the relationships among CAF measures (i.e., complexity, accuracy, and fluency). WCR and EFCR were used and the differences in the development patterns were compared. This study examined how complexity and fluency measures also changed as the WCR score changed.

Chapter 7 focuses on changes in the number of clauses in the WCR rating scale. Although a detailed explanation is provided in Chapter 2, the WCR divides clauses into four types according to the influence of linguistic errors (e.g., Lv.1). The Kruskal-Wallis test was used to determine which categories of clauses changed among the CEFR levels.

Finally, Chapter 8 summarizes and discusses the findings of the present study. Based on the results of the three studies on validation, I concluded that the WCR can be valid in accuracy development studies using the corpus. In addition, according to the results of Studies 4 and 5, I concluded that the accuracy developed as CEFR changed, although the development would be influenced by complexity and fluency. In addition, the present study suggests that the number of each clause decreased as the CEFR levels changed. Finally, suggestions, implications, and limitations are presented.

Chapter 2

Literature Review

2.1 Writing Models and Skills

2.1.1 The Necessity of Second Language Writing

Writing is an essential skill in social and academic activities. The ability to write is necessary for email communication and writing documents for meetings in society. In addition, writing skills are also very important in academic activities. Writing in English is essential for summarizing and disseminating new findings and proposals in papers and presentation slides around the world.

Writing activities are also an area of research studied for many years. As will be discussed in more detail in a later section, the activity of writing is a very complex cognitive process that requires a variety of knowledge and sub-skills. Considering that it is an important skill required by society, elucidating the writing process and the required skills is important to propose appropriate writing instruction and further improve skills.

More effective teaching and assessment methods can be proposed by examining writing skills from a theoretical perspective. This dissertation focuses on the assessment of writing skills (especially written accuracy). To realize the integration of instruction and assessment, it is necessary to explore writing instruction and conduct detailed research on assessment methods.

As mentioned above, writing is a complex cognitive activity that involves many different processes. Various researchers have proposed a wide variety of writing process models. In the next section, we will examine different writing models. In addition, various knowledge and skills are required for writing activities; the cognitive process of writing will be discussed, as well as the skills needed to implement it.

2.1.2 Theories of the Writing Process

Many cognitive processes occur from the moment a learner begins writing to the moment the learner finishes writing. Various writing process models have been proposed to elucidate the process and related factors. There are two main types of cognitive models that have been proposed: global models, which describe the entire writing process, and local models, which represent a specific point of writing activity. For the famous global models, there are three, and were proposed by Flower and Hayes (1981), Hayes (1996), and Kellogg (1996). In contrast, the local models have various types, such as the sentence composition model (Kaufer et al., 1986) and the early writing development model (Hayes, 2011). In this section, the global models are summarized because the section aims to capture the entire writing process in the minds of the learners.

Flower and Hayes' (1981) model is the first model of the writing process. They described the writing process from several points of view: task environment, writer's long-term memory, planning process, translation process, revision process, and monitoring. The task environment consists of two types: the writing assignment and the text produced by the writer. In addition, a writing assignment consists of three components (i.e., topic, target audience, and motivational cues). In planning, it is assumed to generate ideas for writing, setting goals, and organizing them with components (e.g., knowledge of the topic) stored in writers' long-term memory. Then the ideas are translated to be put into written language. Once written, these texts are reviewed and edited. If certain parts should be rewritten, the authors try to correct the part. Finally, these three processes (i.e., the planning, translation, and revision processes) are monitored to assess whether or not text production is progressing.

One of the most important points in Flower and Hayes' model is that writing activity is a recursive process, not a linear process (e.g., Weigle, 2002). Once the writer produces the text, it is evaluated in the reviewing process; the text is then evaluated as part of the reviewing process, and if it is felt that it needs to be revised, it is shown that it needs to be revised to make it better.

Hayes (1996) revised Flower and Hayes's model and proposed the revised writing model. In Hayes's revised model, the task environment and the individual were set. The social environment consisted of the audience (real or imagined) and collaborators in the task environment. On the other hand, the physical environment includes the text written by writers and the writing medium, such as a computer processor or pen and paper. The composing medium was included because the opportunities to write with handwriting and computers increased (Weigle, 2002).

In the individual part, four components are involved: the social environment, working memory, long-term memory, and cognitive processes. The actual cognitive processes that occurred during the writing activities are included in the cognitive process part. This part consists of three components: text interpretation, reflection, and text production. Text interpretation involves a variety of inputs such as listening and reading to create internal representations. Upon reflection, the internal representations are changed by listening and reading to new internal representations. Text production is a process of creating the internal representations of written texts. During this cognitive process, three components are interrelated in individuals.

While long-term memory was included in the early models, working memory was not introduced until Hayes' model. Working memory in this model was adapted from Baddeley's (1986) model. Although it is not listed in the figure, working memory consists of three components (phonological memory, visual/spatial sketchpad, and semantic memory). In addition, Hayes' model recognized that the social environment component plays an important role in the writing process; goals, dispositions, beliefs, and attitudes are considered as factors that can influence the performance of the writing task.

However, these writing models were developed to explain first language writing (e.g., Michel et al., 2020). Therefore, it would not be suitable to explain second language (L2) writing processes. Although it might be difficult to completely explain the L2 writing processes, Michel et al. (2020) suggested that a writing model proposed by Kellogg (1996) could be informative to L2 writing. Michel et al. explain that Kellogg's writing model would be suitable as it focuses on linguistic encoding processes. The previous studies have shown that L2 learners put much more effort into language production than first language learners and require much more awareness of the representation of concepts (e.g., Kormos, 2012).

Kellogg's (1996) writing model includes three main processes: Formulation, Execution, and Monitoring. Formulation involves planning the content and deriving relevant concepts and ideas from the task, topic text, and writer's long-term memory. In addition, formulation involves translation processes such as selecting relevant vocabulary, syntactic coding, and creating coherence. In this process, the ideas and concepts in the writer's mind are transformed into language. In the execution phase, writers are guided by the means (e.g., handwriting or word processing) by which they write the language. The writer should verify that what he has written is consistent with his intentions. If it turns out that it needs to be corrected, the writer corrects the incorrect expression in an appropriate way.

Thus far, we have reviewed the writing models proposed in previous studies. As the required elements and knowledge vary greatly from model to model, the writing process is still the subject of much research. The next section provides an overview of theoretical proposals for the competencies and knowledge involved in writing activities. As with writing models, various types of skills and abilities have been proposed to explain the construct of writing proficiency.

2.1.3 Theoretical Models About Writing Abilities

It can be said that the assumed abilities and knowledge behind second language writing performance vary. This section will review previous studies that have modeled writing proficiency. According to Yoon and Burton (2021), models of writing competence should be models that relate to general language competence. The first model of writing competence was proposed by Grabe and Kaplan (1996). The model proposed by Grabe and Kaplan is based on the model proposed by Canal and Swain (1980) and Bachman (1990). Grabe and Kaplan's model consists of three main components: linguistic knowledge, discourse knowledge, and sociolinguistic knowledge. Linguistic knowledge is knowledge about language itself. Specifically, it includes spelling, phonological, and lexical knowledge. This includes knowledge of syntactic structure. In addition to knowledge of basic syntactic structures, this includes, for example, choosing syntactic structures according to writing style. Discourse knowledge includes knowledge of sentence and structural coherence, knowledge of the topic of the task, and knowledge of writing style. Finally, sociolinguistic knowledge includes the sociolinguistic use of the written language and the writer's factors (e.g., age, proficiency).

According to Weigle (2002), one model of competence that constitutes writing proficiency is the communicative fluency proposed by Backman and Palmer (1996). These models consist of two main components: language knowledge and strategic competence.

Linguistic knowledge consists of four other components (i.e., grammatical, textual, functional, and sociolinguistic knowledge), each of which includes different elements. Grammatical knowledge consists of lexical and morphological knowledge, syntactic structure knowledge, and phonological knowledge. Textual knowledge includes knowledge of coherence and rhetoric. In addition, functional knowledge consists of knowledge about language (e.g., ideational functions) required to perform various communicative tasks. Finally, sociolinguistic knowledge includes dialects, linguistic diversity, and cultural background.

The most important difference from the model proposed by Grabe and Kaplan (1996) is the introduction of strategic competence, which is defined by Backman and Palmer as "a set of metacognitive components, or strategies, which can be thought of as higher-order executive processes that provide a cognitive management function in language use, as well as in other cognitive activities (p. 70)." It can be said that methodological competence is the general ability to use linguistic knowledge properly to serve communicative purposes effectively (Weigle, 2002).

2.1.4 Assessing Written Abilities in L2 Writing

This section summarizes the methods used to assess writing skills. Writing assessments measure general writing skills, but the main focus is on accuracy, which the dissertation focuses on.

The writing assessments are crucial to understanding learners' development progress and making certain decisions, such as classifying learners according to their proficiency. The importance of this is also evident in the evaluation of accuracy. Weigle (2002) has demonstrated two methods for assessing writing skills: holistic and analytical assessment.

Holistic assessment is a method of assigning a grade for a writing sample with a rubric that describes the assessment perspectives of each grade (Goulden, 1992, 1994). Rubrics usually consist of five levels of evaluation criteria such as a TOEFL writing rubric.

When scoring writing samples, raters decide whether to award a single score for a writing performance based on the overall impression of the writing samples. The biggest advantage of holistic scoring is that it does not take much time. The raters decide only one score for each writing sample; therefore, its practicality is high.

However, holistic assessment has several disadvantages. One is that a single assessment does not contain diagnostic information about an author, such as composition, grammar, and content. From an accuracy perspective, holistic assessment would not provide learners with detailed information about accuracy levels. For example, a holistic assessment in a TOEFL writing section only shows certain errors, such as sentence structure, vocabulary choice, and word forms. Therefore, learners might not be clear on which parts of their essay are inaccurate. Weigle (2002) claimed that the disadvantage would be problematic for L2 learners. Some learners may write English composition structure and content but with low accuracy, while others may have high accuracy in writing but convey poor content. While the holistic evaluation is an easy method for raters, it is not beneficial for L2 learners sometimes. Another drawback is the vagueness of the descriptions of each score (e.g., Weigle, 2002). Raters might not use the descriptors correspondingly to arrive at the ratings. For example, some raters might assign 2 points for a script because the accuracy of the script was low, while other raters would assign the same score for the same script based on content quality. This tendency is reported in some studies, which suggest that experienced raters tend to give different scores compared to novice raters (e.g., Barkaoui, 2010; Cumming, 1990).

Analytical scoring is another method of evaluating writing. Although the viewpoints included in the rubric vary by test, analytical scoring encompasses multiple perspectives, including structure, content, grammatical accuracy, and vocabulary. Raters must read a writing sample and assign a score for each characteristic. Then, the final grade is calculated based on all the scores. The advantage of analytical assessment is that learners receive more detailed information about their written work than in holistic assessment. Therefore, learners can focus on the different aspects of their writing and practice the features where they feel weak.

However, the disadvantage of the analytical evaluation is that it takes longer than the holistic evaluation (e.g., Goulden, 1992, 1994). Raters must judge the quality of scripts based on the descriptors and decide on each characteristic in the rubric. In addition, the rater's impression would influence the analytic evaluation, although the holistic evaluation has this limitation as well (Evans et al., 2014). It might be natural to recognize that all evaluations by human beings would have this problem (e.g., Weigle, 2002). Moreover, not all analytic assessments have the accuracy characteristics that researchers want to focus on. For example, one example of a rubric in Weigle (2002) has a perspective that relates to accuracy (i.e., language use). However, the descriptor handles the accuracy of grammar and syntactic constructions. Therefore, it would be difficult to evaluate only one feature (e.g., type accuracy) in the analytical evaluation of handwriting.

2.2 Complexity, Accuracy, and Fluency (CAF)

2.2.1 The History and Definition of CAF

Three concepts (complexity, accuracy, and fluency: CAF) have emerged since the 1990s and have been used in recent studies (Housen & Kuiken, 2009). The earlier literature indicated that the origins of these concepts were in studies of L2 pedagogy (Housen & Kuiken, 2009). Housen and Kuiken explained two different concepts: Accuracy and Fluency. Brumfit (1984) distinguished between activities related to fluency (i.e., promoting spontaneous oral production) and activities related to accuracy (i.e., focusing on grammatically correct production). Then, Skehan (1989) proposed a complexity construct and established the first L2 model, which includes CAF as the three main dimensions of language proficiency.

Since the 1990s, studies have examined whether the three-factor model was organized. Foster and Skehan (1996) examined the factor structures of CAF using factor analysis. Foster and Skehan's study concluded that three factors were extracted in relation to CAF. In recent years, Koizumi and In'nami (2014) verified the three-factor model using a structural equation model, although this study focused on speaking in the L2. CAF is considered an important construct not only for describing written and oral performance and assessment, but also for measuring progress in language learning (Housen & Kuiken, 2009).

While CAF has been applied in many areas of study, such as writing scores and Second language acquisition (SLA), the definitions of CAF are not consistent across studies and have been much debated (Housen et al., 2012). There are several definitions of *complexity* (e.g., Ellis & Burkhuizen, 2005; Housen et al., 2012; Skehan, 2003; Wolfe-Quintero et al., 1998). Wolfe-Quintero et al. (1998) described that "grammatical and lexical complexity mean that a wide and variety of both basic and sophisticated structures and words are available to the learner" (p. 69, p.101). While the descriptions by Wolfe-Quintero et al. focused on the words, which would be basic and sophisticated structures, Skehan (2003) defined complexity as "the complexity of the underlying interlanguage system developed" (p.8). Recent studies (Housen et al. 2012) reviewed the history of complexity and defined complexity as "the ability to use a wide and varied range of sophisticated structures and vocabulary in the L2" (p. 2).

In addition, the previous studies claimed that complexity could be divided into two sub-dimensions: syntactic and lexical complexity. Syntactic complexity is defined as "the range of forms that surface in language production and the degree of sophistication of such forms" (Ortega, 2003, p. 492). Despite the detailed explanations about the measurement, the syntactic complexity has been measured by three linguistic units: sentences, clauses, and phrases. Regarding linguistic complexity, most studies have assumed that words in learners' speech reflect vocabulary knowledge and are able to test general L2 knowledge (e.g., Jarvis, 2013).

Although there were some definitions in the previous studies (e.g., Housen et al., 2012; Wolfe-Quintero et al., 1998), it can be said that the words in the definitions would be similar and coherent among the previous studies. For example, Wolfe-Quintero et al. (1998) defined accuracy as "The ability to be free from errors while using language to communicate" (p. 33). Housen et al. (2021) claimed that *accuracy* is defined as "the ability to produce target-like and error-free language" (Housen et al., 2012, p. 2), which the present study used.

Between accuracy and complexity, there is some evidence to suggest that these constructs are related to learners' current level of interlanguage knowledge. The knowledge would include partly declarative and procedural knowledge (Housen & Kuiken, 2009). While accuracy would reflect "the conformity of second language to target language norm" (Wolfe-Quintero et al., 1998, p.4), complexity would be viewed as "the scope of expanding or restructured second language knowledge" (Wolfe-Quintero et al., 1998, p.4). Therefore, Housen and Kuiken (2009) suggested that complexity and accuracy related to L2 knowledge representation.

Finally, *fluency* also has a variety of definitions (e.g., Ellis & Barkhuizen, 2005; Lennon,1990; Housen et al., 2012). For example, Ellis and Barkhuizen (2005) defined fluency as "the production of language in real time without undue pausing or hesitation" (p, 139). In recent studies, Housen et al. (2012) defined fluency "the ability to produce the L2 with native-like rapidity, pausing, hesitation, or reformulation" (p. 2). Both definitions assume that they correspond to the behavior of native English speakers. Some scholars claim that fluency is a multidimensional construct and can be subdivided into subdimensions such as speed and pausing fluency.

2.2.2 Importance of Accuracy and Difficulty of Developing Accuracy in Writing

CAF is an important construct of writing proficiency used in various studies. This study focuses on accuracy among them. In this section, the importance of accuracy is first pointed out. Then, current examples from classroom and research are presented, considering the difficulties in developing accuracy skills in learners. Finally, the importance of accuracy measurement is discussed.

Accuracy in writing is a skill that is important in accurately communicating one's ideas to the reader. If the wrong language is used in a text, the content may not be understood and misunderstandings may occur.

Accuracy in writing in English is also one of the assessment elements in various major tests. For example, the International English Language Testing System (IELTS) writing assessment is about grammatical knowledge and accuracy. In the EIKEN rating scale, accuracy of vocabulary and sentence structure are also among the criteria for rating.

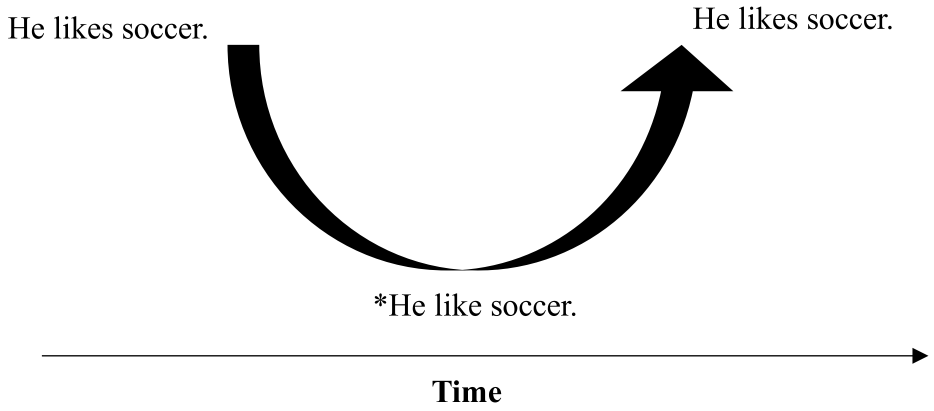
However, developing accuracy is not an easy task. The development of writing skills would require writing teachers or educational institutions. Likewise, the skill to write accurately is difficult to develop by L2 learners because accurate writing in L2 requires various grammatical and lexical knowledge (e.g., Weigle, 2002). Spinner (2021) claims that L2 learners need to know the correct knowledge and context in which it is appropriate to use certain grammatical forms (e.g., passive voice).

Moreover, L2 learners need a lot of time to acquire the knowledge. SLA has found that learners at certain stages of developmental progress tend to perform less accurately.

It is called *U-shaped development* or *U-shaped patterns*, in which accuracy tends to increase and then decrease before increasing again (Gass et al., 2020). Gass et al. (2020) explained that this pattern had three stages (Figure 2.1).

Figure 2.1

U-Shaped Development



In the first stage, learners produce some linguistic forms such as a past tense and subject-verb agreement accurately (*He likes soccer.*). At the second stage, learners tend to produce the less accurate forms (**He like soccer.*). Then learners appear to use correct forms (*He likes soccer.*).

Furthermore, SLA studies (e.g., Gass et al., 2020) showed that some grammatical knowledge that is deviant from the target language norm continues to be produced permanently despite further language input (i.e., *fossilization*). Thewissen (2013) investigated the developmental patterns of accuracy in L2 learners and suggested that some linguistic features (e.g., morphological errors) tend to be produced by learners with higher proficiency. Based on the theory and the results of the studies of SLA, there are many obstacles to the acquisition of correct forms. The ability to produce language correctly is also a struggle for L2 learners.

Even if L2 learners have difficulty performing accurately in written activities, accurate written performance should be important (e.g., Kitamura, 2011). Kitamura (2011) used 134 essays on two topics written by Japanese EFL learners and examined the relationships between essay scores and linguistic errors using a decision tree analysis. Kitamura's study suggests that the number of errors is an influential predictor, although the degree of influence may vary depending on the topic of the English essay. Considering the influence of errors on writing quality, it is necessary to promote writing accuracy in educational settings (e.g., classroom).

Although there are many methods to develop written accuracy, *written corrective feedback* has been the most used method in writing classrooms and explored the effectiveness (e.g., Ferris, 2010). Because there are different types of corrective feedback, many studies have examined which feedback is most effective for developing written accuracy (e.g.,Van Beuningen et al., 2012). Although many studies have examined the effectiveness of corrective feedback, recent studies claim that the effectiveness of corrective feedback varies depending on individual differences, such as performance ability and motivation (e.g., Sheen, 2007).

As another method to develop accuracy, *data-driven learning* (DDL) has attracted the researcher's and teachers' attentions (e.g., Chang, 2014; Mizumoto & Chujo, 2016; Moon & Oh, 2017). DDL is a methodology that uses corpora (a collection of many texts) to not only language teaching but also language learning (Aijimer, 2009). Learners must access a large amount of authentic input on the computer and inductively explore the rules of certain grammatical forms. Some studies have reported that DDL enables learners to notice their errors and apply DDL as error correction (e.g., Gaskelll & Cobb, 2004).

While many types of instructional methods for developing writing accuracy have been proposed in recent years, understanding the learner's accuracy levels should be important for selecting the appropriate teaching methods. Some recent studies about corrective feedback (e.g., Lee, 2019) claimed that written corrective feedback (especially comprehensive one) should not be given because it might be over the learners' readiness, based on Pienemann's processability theory (1998). From the previous studies, it appears that it would be necessary to provide corrective feedback considering the learner's level of development. Also, teachers need to consider learners' abilities when using DDL as a teaching method. Since DDL requires learners to derive certain rules from the corpora, teacher guidance would be necessary for learners with lower proficiency (Moon & Oh, 2017).

As for choosing appropriate instructions, assessing accuracy in learners' writing performance is fundamental to understanding learners' ability. In this way, teachers can then select the most appropriate instructional strategies (e.g., corrective feedback) to improve their students' writing accuracy.

Although studies using CAF are increasingly being conducted in a variety of settings (e.g., language development studies), the CAF framework would present some problems, such as the link between CAF. One of the important challenges concerns definitions and measurements: How can CAF be operationalized and measured? To get an overview of the measurement problems and the latest ideas of CAF, the overviews of the problems are summarized in the next section.

2.2.3 The Measurement of CAF

This section first summarizes the measurement of complexity. Then the measurement of fluency is described, and finally the measurement of accuracy is presented. It can be said that most studies have focused on the measurement of complexity.

Bulté and Housen (2012) summarized the studies that use complexity measures and pointed out the construct specifications for complexity. L2 linguistic complexity can be divided into two types: grammatical variety and sophistication. Grammatical variety is related to syntactic complexity, and syntactic complexity is divided into three subgroups, namely sentence complexity, clause complexity, and phrase complexity. Each subgroup is operationalized by some measures based on linguistic features. For example, sentence complexity is operationalized by words per T-unit, comprising a main clause, consisting of an independent clause and any related dependent clauses . If the number of words in a T-unit is large, syntactic complexity in a sentence is high.

In recent years, Kato (2019) pointed out the complex measurement structures and investigated the measurement models of L2 complexity using a structural equation modeling approach. Kato's study showed that the construct of syntactic complexity would have phrasal (e.g., mean clause length: MLC), clausal (e.g., clauses per clause: C/S), and verb-argument (VAC) sophistication (e.g., dependents per clause) construct.

As for measuring fluency, the previous studies have used a variety of measures (e.g., Wolfe-Quintero et al., 1998). Wolfe-Quintero et al. suggested that T-unit length, error-free T-unit length, and length of clause could be used to measure fluency. Some studies claimed that more proficient learners tend to produce more fluent writing texts (Ortega, 2003). However, the previous studies pointed out that the fluency measures are similar to the measures of accuracy and complexity (Barrot & Agdeppa, 2021) since the T-unit length is used for complexity measures and error-free T-unit is used for accuracy measures. In recent years, studies have started to employ key-stroke software (e.g., Leijten & van Waes, 2013) that can record online writing production. These studies focused on real-time features such as the time of pause, which would be different from T-unit length and error-free T-unit length. Michel (2017) agreed that such measures can be easily identified and distinguished from measures of complexity and accuracy.

Although many studies on complexity measurement have been conducted since the 2000s, Michel (2017) pointed out the need for studies on accuracy measurement. He called for the discussion on accuracy measurement. Accuracy measurement is described below.

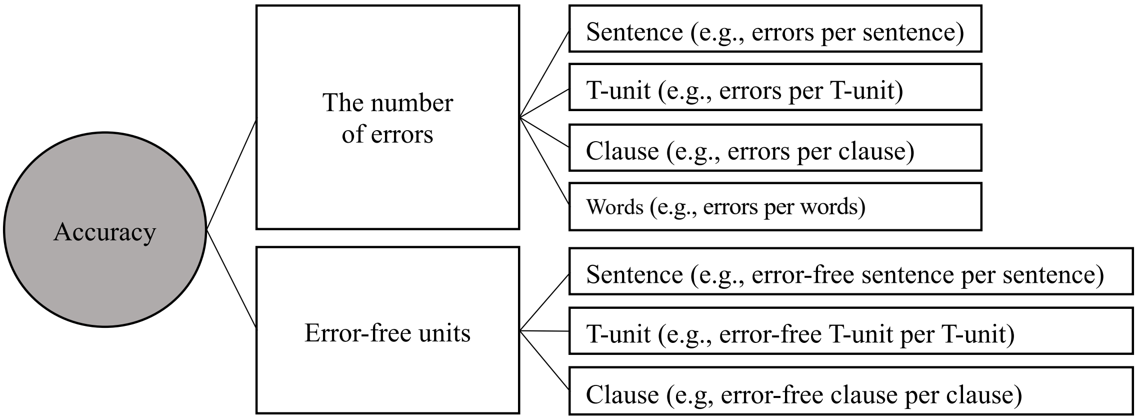
In assessing written accuracy, an *error* is an important linguistic feature to calculate the written accuracy measures. The error is defined as “A linguistic form or combination of forms which, in the same context and under similar conditions of production, would, in all likelihood, not be produced by the speakers’ native speaker counterpart” (Lennon, 1991, p. 182).

There are some types of errors: omission and overgeneralized errors. These error classifications can be applied to errors in all errors such as grammatical, vocabulary choice, and rhetorical errors. For example, if a writer produces **He like soccer,* the sentence has a subject-verb agreement error and should be *He likes soccer*. More specifically, this error is the missing subject-verb agreement error. In addition, if a learner writes a sentence such as **He goed to a supermarket*, an overgeneralization error of past tense inflection. In measuring the written accuracy, the decrease of errors means the increase of accuracy. Therefore, accuracy and errors would be two sides of the same coin.

A variety of measures have been used for measuring accuracy in previous studies (Figure 2.2). Figure 2.2 summarizes the accuracy measure used in previous studies (e.g., Evans et al., 2014).

Figure 2.2

Summary of Accuracy Measures


Accuracy
The number of errors
Sentence (e.g., errors per sentence)
T-unit (e.g., errors per T-unit)
Clause (e.g., errors per clause)
Words (e.g., errors per words)
Error-free units
Sentence (e.g., error-free sentence per sentence)
T-unit (e.g., error-free T-unit per T-unit)
Clause (e.g, error-free clause per clause)

In some studies, written accuracy has been measured by counting the number of errors or the type of errors in written performance (e.g., Chandler, 2003; Truscott & Hsu, 2008). In measuring written accuracy by counting the number of errors, there would be four possibilities: (a) the number of total errors, (b) a ratio of total errors, (c) a value multiplied by the ratio of total errors times 100, and (d) the number of a particular error (e.g., verb-noun collocation). For example, 50 errors in a 200-word written performance would yield a ratio of 0.25 if a ratio of total errors is used. In the case of measure (c), the value using a ratio of total errors is multiplied by 100 and would be 25 in the same situation (b). As research examples, Chandler (2003) used a ratio of total errors to total words to examine the effectiveness of corrective feedback. In addition, Truscott and Hsu (2008) used a value multiplying the ratio of total errors in total words by 100.

While some studies have measured writing accuracy by the number of errors, other studies have measured writing accuracy by various error-free language units such as sentence parts, T-units (a T-unit is a unit, including the main clause, consisting of an independent clause and all related dependent clauses), and sentences (e.g., Evans et al., 2010). For example, error-free clauses per total of all clause (EFCR), error-free T-units per T-unit (EFTR), and error-free sentences per total of all sentence (EFSR) were mainly used. To explain the score calculations and the differences among the measures, here is a writing example ("*Honestly, I think this tremendous writing is*."), which was used in Michel (2017). When the EFCR is used, the first step is to divide the sentence into clauses: (1) *Honestly, I think,* and (2) *this tremendous writing is*. Then the evaluators find errors in each clause. In the case of (1), there is no error, so clause (1) receives 1.0. However, clause (2) has some errors, such as errors in word order and in the article; therefore, the clause receives 0. The total score is calculated as error-free clauses per clause, so the total EFCR score is 0.5.

On the other hand, when the EFTR is used, raters divide the sentence into T-units. In that case, the sentence would also be a T-unit. Then, raters find errors that would include in the T-unit. There are some errors mentioned above so that the T-unit would receive 0. The total score is calculated as error-free T-units per T-unit, so the total EFCR score is 0. The calculation of the EFSR is similar to the EFTR.

Regarding writing studies that examined the development of CAF, some studies have suggested appropriate measures of writing accuracy to capture the changes (e.g., Larsen-Freeman & Storm, 1977; Wolfe-Quintero et al., 1998). Wolfe-Quintero et al. (1998) examined accuracy measures in detail. They concluded that the total number of error-free T-units (a T-unit is a unit, including the main clause, consisting of an independent clause and all associated dependent clauses) is the best measure for measuring accuracy. They also added measures such as error-free T-units over total number of T-units (EFTR).

While various measures have been used, Housen et al. (2012) claimed a lack of consensus in how CAF should be measured. Michel (2017) claimed that methods are necessary to measure accuracy. Recent studies consider measures that use syntactic units of error-free (e.g., EFCR) to be suitable for measuring accuracy because a clause is (a) reliably identified and (b) allows fine-grained analysis of written data (Foster & Wigglesworth, 2016).

Regarding (a), Foster and Wigglesworth (2016) suggested that written accuracy measures that use error-free units are easier to count than those that use error counting. Errors would be not only one linguistic error but also multiple related linguistic errors. However, an error-free unit can be identified for raters because they only need to count the error-free unit.

As for the (b), the clauses would be useful to capture the small difference better than the EFTR because the EFCR is shorter than the EFTR. While the EFCR would be able to show that there is an accurate clause (*Honestly, I think*) in the sentence, the score of the EFTR would not reflect the accurate clause. For these reasons, Foster and Wigglesworth (2016) agreed that the measures using error-free clauses would be appropriate for capturing the changes of written accuracy.

However, some scholars have been critical of using error-free units to measure accuracy because they do not account for the characteristics of *error* or *error gravity,* which may have different degrees of influence on reader comprehension (e.g., Kuiken & Vedder, 2008; Polio, 1997). Errors have two categories depending on reader comprehension: global and local errors (Burt, 1975). Global errors were defined as "errors that significantly hinder communication are those that affect overall sentence organization" (Burt, 1975, p. 56), such as wrong word order and missing errors. In contrast, local errors were viewed as "errors that affect single elements (constituents) in a sentence do not usually hinder communication significantly" (Burt, 1975, p. 57). Local errors include, for example, inflection of nouns and verbs, articles, and auxiliary verbs. While global errors cause readers to misinterpret the writer's message, local errors rarely affect the transmission of messages.

The studies about error gravity have been conducted since the 1980s (e.g., Rao & Li, 2017; Rifkin & Roberts, 1995; Vann et al., 1984). Vann et al. (1984) studied how teachers from different departments at Iowa State College rated the impact of errors in essays. They found that most teachers did not rate all errors as equally serious. For example, while raters tended to rate spelling errors as not influential, they seemed to consider tense errors influential. Rao and Li (2017) also focused on the influence of 10 kinds of errors and compared the differences in ratings between teachers of native and non-native speakers. Rao and Li showed that all errors included in the study had more or less influence on native speakers and non-native speakers.

Although errors would more or less influence the readers' comprehension, the traditional written accuracy measures (e.g., EFCR and EFTR) do not consider the influences. In other words, a clause with minor errors (e.g., subject-verb agreement errors) receives the same score as another clause with serious errors (e.g., serious word choice errors). Using an example, Foster and Wigglesworth (2016) pointed out the limitations that the error-free clause units would have. The slashes indicate the sentence boundaries, and the underlines, which I have added, indicate the errors.

(1a) *In multiple-choiced exams, it is not hard /* (1b) *to make an educated guess, or a random one.*

(2a) *On the other side of the story, I see /* (2b) *their mum see the childs a big map.*

In each case, a sentence has two clauses. In example (1a), the clause has one minor error (i.e., d-affix in *multiple-choiced*). On the other hand, clause (1b) does not have any errors. In addition, clause (2a) has a word choice error (*story* for *picture*), while clause (2b) has several errors: vocabulary choice error (*see* for *show*) and noun inflection (*childs* for *children*). The traditional accuracy measures such as EFCR do not take into account the error severity and consider these errors as equal weight. Therefore, Evans et al. (2014) claimed that using error-free clauses could miss potential differences in accuracy levels, which could affect the reliability and validity of the measure.

2.2.4 The New Accuracy Measure: Weighed Clause Ratio

Based on the limitations of accuracy measures using error-free units, Wigglesworth and Foster (2008) proposed WCR as an alternative accuracy measure. WCR measuring accuracy was developed based on a construct of *adequacy* that ‘‘represents the degree to which a learners’ performance is more or less successful in achieving the task’s goals efficiently” (Pallotti, 2009, p. 7).

The WCR rating scale consists of four categories, definitions, and scores (Table 2.1). After all sentences in an essay are divided into clauses, each clause is categorized by its gravity of error according to the definitions. The category of *No error* means that a clause has no linguistic errors and is accurate. If the clause is accurate, the clause receives 1.0. On the other hand, the category of *Level 1* means that a clause has local errors such as subject-verb agreement and would not influence the readers’ comprehension of the clause. Although the kinds of errors included in Lv.1 would be vague, the morphosyntax errors are included in Lv.1. If the clause has the local errors and is regarded as Lv.1 category, the clause receives 0.8. Next, the category of *Level 2* means that a clause has serious errors (e.g., verb tense) although the meaning is recoverable. The errors corresponding to the category Lv.2 are errors in the tense of the verb and in the word choice, which are divided into global errors. If the clause has global errors and is classified as Lv.2 category, the clause receives 0.5. Finally, the category of *Level 3* means that a clause has very serious errors making the intended meaning far from obvious. Although the kinds of errors to be included in the Lv.3 categories are not written, some global errors would be included. If the clause has global errors and is regarded as Lv.3 category, the clause receives only 0.1. When producing the total WCR score, the final WCR score was calculated as follows: WCR = the number of accurate clauses × 1.0 + the number of Lv.1 clauses × 0.8 + the number of Lv.2 clauses × 0.5 + the number of Lv.3 clauses × 0.1 / all clauses in the essay.

Table 2.1

Rating Scale of a Weighted Clause Ratio

Category	Definition	Score
No error	The clause is accurately constructed.	1.0
Level 1	The clause only has minor errors (e.g., morphosyntax) that do not compromise meaning.	0.8
Level 2	The clause contains serious errors (e.g., verb tense, word choice, or word order), but the meaning is recoverable, though not always obvious.	0.5
Level 3	The clause has very serious errors that make the intended meaning far from obvious and only partly recoverable.	0.1

Here, it should be noted that the WCR has two cautions when providing scores on a clause, unlike the traditional written accuracy measures (e.g., EFCR). Foster & Wigglesworth (2016) noted that 0 should not be awarded because even Lv. 3 clauses have linguistically accurate parts. In addition, when there were multiple errors (e.g., Lv. 1 and

Lv. 3 errors) in the same clause, the clause was categorized according to its worst-level error.

Michel (2017) gave an example about the judgment and calculation of the WCR. When a student writes, "*Honestly, I think this tremendous writing is*.", the first step which a teacher takes is to divide the sentence into clauses; (1) *Honestly, I think,* and (2) *this tremendous writing is*. Then the teacher finds errors and considers how the errors would influence the understanding. Clause (1) would receive 1.0 because it would be the accurate clause. In contrast, clause (2) would have word order errors and an article so that the clause would receive 0.5. Finally, the total score is calculated [e.g., (1.0×1) + (0.5×1) / 2 = 0.75].

In contrast to the binary classification of accuracy and inaccuracy (e.g., EFCR), WCR provides researchers and teachers with considerable insight into assessing accuracy and language performance by categorizing errors according to their severity. Unlike other measures, WCR is sensitive to small changes in written accuracy in written performance because WCR can categorize clauses in written performance in greater detail than traditional measures. Thus, researchers can assess accuracy without overlooking possible differences in accuracy levels when using measures of error-free units.

Moreover, they can track the small developments in writing accuracy in writing performance over time (Evans et al., 2014). While traditional measures of writing accuracy divide clauses into only two types (i.e., accurate clauses or inaccurate clauses), WCR can divide clauses into four types. Therefore, researchers can use it to track the number of correct clauses and three kinds of clauses. For example, researchers could set up a research question about how the clauses, including various errors, would change. The research question on clauses with errors would help researchers understand how the WCR score would increase.

In addition, the WCR can be useful when English teachers provide corrective feedback on errors in written performance. Rather than teaching all kinds of errors, teachers can use WCR to focus on the errors that significantly affect readers' comprehension. While corrective feedback on errors in written performance can increase accuracy, some researchers have claimed that comprehensive corrective feedback that addresses all errors may affect learners' motivation (e.g., Lee, 2019) and is not always helpful in reducing all kinds of errors (e.g.,Oka, 2019). In addition, many previous studies have shown that targeted corrective feedback (i.e., feedback on two or three kinds of errors) is more effective than comprehensive corrective feedback. Therefore, corrective feedback should be given according to the purpose and situation. The WCR could help teachers think about where they should give corrective feedback. By using the WCR, learners can also learn how to write/rewrite the intended messages while keeping readers in mind.

2.2.5 The Discussion About Written Accuracy Measurement

Although WCR has many advantages for accuracy assessment and language learning, some studies doubt that measuring written accuracy while accounting for error severity could not measure written accuracy (e.g., Pallotti, 2009). Pallotti (2009) claimed that "a 100-word production with 10 errors not compromising communication is not more 'accurate' than a text of the same length with 10 errors hindering comprehension, but just more 'understandable' or 'communicatively effective'." The accuracy assessment using the WCR might be debatable from the claims by Pallotti (2009). Furthermore, some studies (e.g., Pallotti, 2009; Wolfe-Quintero et al., 1998) claimed that the developments or changes of accuracy measures' scores would not directly imply language development. Wolfe-Quintero et al. claimed that "the purpose of accuracy measures is precisely the comparison with target-like use. Whether that comparison reveals or obscures something about language development is another question (p. 33)."

On the other hand, the other studies (e.g., Housen & Kuiken, 2009; Housen et al., 2012) showed opposed opinions about written accuracy. Housen and Kuiken (2009) explained that accuracy and complexity would link to the current state of the interlanguage knowledge (partially explicit and partly implicit knowledge). In addition, Housen et al. (2012) also described that accuracy and complexity would relate to L2 knowledge representation. Thus, the way accuracy is perceived differs among researchers and is still under debate.

First, I would like to address what the accuracy measurements reflect. One possible reason for the discrepancy between perceptions in previous studies is the research fields and objectives. Suppose the goal of the study is to examine changes in individual linguistic features. In this case, an accuracy measure that only provides a value for the entire English text is not appropriate. Even if the EFCR value changes from 0.2 to 0.7, it is impossible to know which linguistic feature is changing. Thus, as Pallotti (2009) says, accuracy measures may not perfectly reflect language development. However, in language testing, accuracy is one of the abilities that make up writing proficiency. Therefore, the accuracy measures evaluate the learner's overall ability to be accurate rather than language development.

Second, the problem that written accuracy measures that consider the degree of influence of errors do not reflect accuracy may have a similar cause than the one described above in the way errors are viewed. Since research on second language acquisition focuses on changes in individual language elements, it can be said that it is not necessary to consider the degree of influence of errors on the reader as accuracy. In this case, another measure (e.g., Target-like use method) would be needed to capture accuracy.

On the other hand, writing tests usually assume a “reader.” In argumentative writing tasks, for example, students are expected to write about a topic in a way that persuades the reader rather than the writer. Given the reader, errors will affect the reader to a greater or lesser degree. Therefore, accuracy measured by errors must also evaluate the quality of the writer's performance in anticipation of the reader. Since the term error is used in the definition of accuracy, namely “the ability to produce purposeful and error-free language,” it is assumed that a measure that considers the effects of errors is an accuracy measure.

While improvement in the results of an accuracy measure that takes into account the severity of errors suggests the development of accuracy skills, researchers should be cautious about whether this leads directly to language development. Suppose we want to elucidate language development in writing. In that case, it may be better to focus on the changes in individual errors rather than the numerical values of the accuracy measures (e.g., Thewissen, 2013).

Although discussion of WCR has continued, some scholars (e.g., Barrot & Adgeppa, 2022) have begun to use WCR to study the evolution of accuracy using a corpus, i.e., a collection of written materials stored on computers. Since there is a corpus of written performance of Japanese EFL learners, it would be possible to investigate the development of written accuracy of Japanese EFL learners using WCR. By showing the patterns of accuracy development using WCR, one could find out not only what kinds of sentences Japanese EFL learners produce, but also how they develop their accuracy skills.

However, one issue would be unclear: is the WCR valid for accuracy development studies? Based on the recent theory of validity (Chappelle et al., 2008; Kane, 1992), the validity studies focusing on the WCR would not provide enough evidence (e.g., Evans et al. 2014; Polio & Shea, 2014). If the validity of measurements is not confirmed, the results obtained from experiments are not reliable. It is important to ensure validity when measuring a construct such as an ability and motivation.

Until the studies on the development of written accuracy of Japanese learners EFL are conducted, the next section describes the reviews of corpus studies and the possible limitations. Then the theory and history of validity are described, and the limitations of WCR assessment and use in the corpus studies are summarized.

2.3 Corpus

2.3.1 Definition and History

Before looking at the CAF studies that use corpora, the overviews of the corpora are summarized here. First, the characteristics of the two main types of corpora (native speakers and learners) are discussed. Then, an overview of the CAF studies that use corpora is provided.

Stubbs and Halle (2012) defined a corpus as "a text collection which is large, computer-readable, and designed for linguistic analysis" (p. 1). The texts stored in the corpora are written and spoken language. In language education, there are two types of corpora: a corpus that stores the written/linguistic performance of native English speakers and a corpus that stores the written/linguistic performance of English learners (ESL and EFL). These two corpora have different purposes and unique characteristics.

The British National Corpus (BNC) may be one of the best known corpora storing the output of native English speakers. The BNC contains 100 million tokens of written/spoken English produced by native English speakers in Britten. The BNC has been used for the development of many dictionaries and TESOL materials (Ishikawa, 2011). Moreover, the Corpus of Contemporary American English (COCA, Davis, 2008), which has more than one billion words of text from eight genres (e.g., newspapers), has also been used not only in language development studies (Kyle & Crossely, 2018) but also language testing. In a sample study, Kyle and Crossely (2018) focused on verb-argument construction (VAC, e.g., give + indirect object + direct object) as a linguistic feature and examined not only how many VACs learners used compared to the data in the COCA corpus, but also how VACs correlated with writing performance scores. The study used TOEFL writing data and showed that high-scoring essays tended to contain less frequent VSCs and suggested that VACs were one of the useful features for predicting essay scores.

On the other hand, there are many corpora that store the written/linguistic performance of ESL and EFL learners (i.e., a learner corpus). Ishikawa (2011) suggested that a learner corpus is important to study the “interlanguage”. By developing learner corpora under a variety of conditions (e.g., tasks and topics) and with a wide range of English proficiency, it might be possible to show how learners develop (or acquire) grammatical features (e.g., articles), making corpora a useful tool for studies of second language acquisition and language development.

Some famous learner corpora were made in the 1990s (i.e., The International Corpus of Learner English: ICLE and Louvain International Database of Spoken English Interlanguage: LINDSEI). The ICLE, developed by (Granger et al., 2009), is one of the most influential corpora for learners. This corpus was introduced in 2002 and contained 2.5 million words of essays written by English language learners with 11 different first languages (e.g., Dutch, Swedish, and Turkish). The second version was published in 2009 and added Japanese, Chinese, Norwegian, Tswana, and Turkish performance data to the corpus. The LINDSEI created by Gilquin et al. (2010) would also be as large as the ICLE, as the corpus contains approximately 100,000 learner language words. Unlike the ICLE, the LINDSEI contains oral data produced by advanced learners of English with a variety of native languages (e.g., French, Spanish, and Greek).

While the corpora mentioned above focused on English learners in European countries, many corpora focus on learners in one country (e.g., Japan) and Asian English learners. As the former corpus, the Japanese EFL Learner Corpus (JEFLL), made by Tono (2007), has essays produced by Japanese junior and high school students. The JEFLL Corpus English Composition is a 20-minute free English composition conducted in a classroom without a dictionary. Data were collected on six subjects, ranging from first to third grade. In addition, the JEFLL corpus project "CLAWS" (Lancaster College's automatic tagging software) was used to create the corpus data for online searching, and part-of-speech tagging (C5 Basic Tagset) was done.

As the latter corpus, the International Corpus Network of Asian Learners of English (ICNALE) was developed by Ishikawa (2013). The corpus includes written and spoken data from 2800 college students in Asian countries/regions (China, Hong Kong, Indonesia, Japan, Korea, Pakistan, Philippines, Singapore/Malaysia, Taiwan, and Thailand). In Ishikawa's (2013) study, all participants reported their scores in English tests such as TOEFL, TOEIC, or IELTS and also took a vocabulary test (Nation & Beglar, 2007). Based on their proficiency and vocabulary test scores, participants were classified into four CEFR-linked (Common European Framework of Reference for Language) proficiency levels: *A2*, *B1_1* (lower level of the B1), *B1_2* (upper level of the B1), and *B2+* (a merger of the B2, C1, and C2 levels). Writing performance in ICNALE was determined in the controlled experiments. Learners had to complete each writing task for 20-40 minutes with a length of 200-300 words (Ishikawa, 2011, 2013). They could not use a dictionary during the experiments, but spell-checking was possible.

As various useful corpora have been developed, the studies using those corpora have also been conducted and provide researchers and teachers with insightful suggestions (e.g., Thewissen, 2013). While accuracy in written performance was assumed to increase as English proficiency increased, it was not clear how linguistic errors (e.g., grammatical, punctuation, and lexical errors) would decrease. Thewissen (2013) used the ICLE corpus and examined how L2 learners developed their written accuracy as their CEFR level increased. Thewissen's study showed the developmental patterns of about 40 error types. It was found that lexical errors decreased with increasing English proficiency, while tense errors showed no development. Their study also suggests that the error development patterns include the changes of progress and stabilization, but the development is often seen between B1 and B2 levels.

In addition to the many corpus studies that have examined changes in specific linguistic features, writing studies have also been conducted in recent years that focus on CAF and provide meaningful evidence about language development (e.g., Barrot & Gabinet, 2019; Lu, 2011). Studies from CAF can be divided into two types: Studies that use all CAF measures or focus on a specific measure.

Barrot and Gabinet (2019) used all CAF measures and compared the CAF measures between the writing performance of ESL learners (e.g., in Hong Kong and Singapore) and EFL (e.g., in Indonesia and Japan). Barrot and Gabinet used 1870 essays from the ICNALE corpus. The study used independent samples t-tests and showed that ESL learners produced more complex and accurate texts than EFL learners. In addition, measures of fluency (average number of clauses per text and average number of words in the text) were higher for the EFL learners than for the ESL learners. In addition, the analysis showed that the CAF essays produced by ESL learners varied widely, even when they were at the same language level. Therefore, Barrot and Gabinet (2019) claimed that the CAF measures would be influenced by their English proficiency and their L1 backgrounds and suggested that it would be necessary to investigate the development studies for L2 learners with different L1 backgrounds.

Barrot and Adgeppa (2021) investigated how L2 learners developed their writing performance using all CAF measures. In this study, the WCR, which is the focus of this dissertation, was used to examine the developmental patterns of writing accuracy. Results showed that some of the complexity measures increased with CEFR levels, but some measures (e.g., T-units per sentence) did not change. In addition, WCR scores differed significantly, but the score did not change between A2 and B1_1 levels. Regarding fluency, three measures of fluency differed significantly between most CEFR levels except B1_2-B2 levels.

While Barrot and Ganinet (2019) used all CAF measures to investigate the differences between EFL and ESL learners' writing performance, Lu (2011) focused on syntactic complexity. Lu (2011) focused on 14 syntactic complexity (e.g., MLC) and investigated how ESL learners developed their syntactic complexity using the data of Written English Corpus of Chinese Learners (Wen et al., 2005). By conducting the study, researchers could determine what linguistic structures second language learners use as they improve their language skills. This would be relevant not only to second language writing, but also to second language acquisition studies. In addition, it would be possible to predict learners' writing proficiency without spending a lot of time and money, such as training raters. Lu's study showed that complex nominals per clause (CN/C) and MLC were the best syntactic complexity measures to distinguish two or more adjacent levels.

With the development of computers and technology, writing researchers can store a lot of written data in a corpus. For studies of writing development, the corpus should be an essential tool to examine how writing performance develops as a function of increasing English proficiency. In addition, CAF measures can be analyzed by adding error tags and calculating their values. In particular, complexity measures could be created by a computerized tool such as a L2 syntactic analyzer (Lu, 2010). Combining the CAF framework and the corpora would enable researchers to investigate the development patterns of writing performance and apply the results to a variety of areas such as test development, task complexity, and corrective feedback. However, writing development research using CAF has some limitations, even though it uses the useful corpus.

2.3.2 The Limitation of Developmental Studies of CAF and Measurement Issues of WCR

While the use of CAF measures to identify differences between proficiency levels can provide a more holistic picture of L2 learners' writing performance, such studies may be unable to compare their results and provide an in-depth discussion, based on previous research, on individual changes in CAF measures. For example, Barrot and Adgeppa (2021) showed that complexity measures significantly differed among the CEFR levels, but the discussion in the study was limited to stating that the developments of some complexity measures corresponded to the previous studies. As for the written accuracy, Barrot and Adgeppa only reported that the WCR could distinguish among L2 learners because there were no studies that investigated the accuracy development using the written accuracy measures. The provision of the theoretical warrants or experiment results toward the changes of CAF in the writing studies would enable researchers to develop the theory of language development and generalize the results in future studies.

To address the limitations and provide the theoretical warrants, studying a specific construct or measure deeply would be important. In the CAF framework, it may be said that the previous studies focusing on complexity have provided the characteristics of complexity in writing performance more than studies of written accuracy (e.g., Biber et al., 2016).

The number of studies focusing on written accuracy are fewer than those handling complexity. Michel (2017) pointed out the scarcity of studies of written accuracy and claimed that the investigations of written accuracy should be conducted. While the WCR, which is the new written accuracy measure, has been proposed to reveal accuracy development patterns in detail, none of the studies have investigated the written accuracy development of Japanese English as a foreign language (EFL) learners using WCR.

With the advent of the WCR, the WCR proposal allows us to capture differences that were not captured by the Error-Free clause unit of accuracy measures. For this reason, it has become widely used in developmental and other writing research. However, there are some issues regarding the validity of accuracy measurement by WCR. Polio and Shea (2014) examined how reliable the accuracy assessments were and compared various evaluation methods (a holistic rubric, accuracy measures, and specific grammatical features). Their results revealed significant reliability in all methods. For example, the reliability of the written accuracy measures, which were EFCR, weighted T-units (similar to the WCR but the units were T-units), and EFTR, were over .80. Clearly, this study focused on the reliability of the written accuracy measures and did not use WCR in their study.

Evans et al. (2014), in their study, used three accuracy measures (EFTR, EFCR, and WCR) and examined the validity and reliability of the WCR based on the content validity, criterion-related validity, and construct validity. They converted the measure scores to whole numbers between 0 and 10 because they used a multi-faced Rasch model to analyze only the whole numbers. Evans et al.'s study showed that the separation reliability of the WCR, which means the rater severity in the Rasch model, was .00 points, indicating significant reliability similarity to the EFCR and EFTR. Additionally, they suggested that the WCR correlated with the EFCR (*r* = .88) and EFTR (*r* = .78), meaning that the WCR confirmed the criterion-related validity. As for the content and construct validity, they claimed that the WCR would be within the definitions of written accuracy and concluded that the content and construct validity were also confirmed positively.

While previous studies have examined validity and reliability (Evans et al., 2014; Polio & Shea, 2014), there may have the following limitations: (1) no purposes of measurement of written accuracy and (2) partial validation. Regarding the former limitation, based on the latest validation theory (e.g., Chappelle et al., 2008), validity would differ depending on the situation and purpose (see the next section in 2.4 for details). The validity of the WCR has not been confirmed in the accuracy development studies; therefore, it is necessary to conduct a validation study of the WCR in the use of the corpora.

Furthermore, the previous studies provided only a few warrants to show the validity and reliability of the WCR. While they revealed significant reliability of the WCR (e.g., Evans et al., 2014), the generalizability of the WCR score obtained from tasks remains unclear. Additionally, it is unclear whether the WCR would reflect the same construct as the traditional written accuracy measures (e.g., EFCR). The necessary evidence to build the validity could differ depending on the purpose; hence the WCR still has measurement limitations. The history and theory of validity is reviewed in the subsequent sections to discuss the validation of the WCR at the end of the section.

2.4 Theory of Language Testing

2.4.1 Construct and Measurement in L2 Writing

This section analyses the ideas of construct and measurement using an example of an experiment in L2 writing. In conducting experimental studies of L2 writing such as corrective feedback, two types of variables are important: *independent* and *dependent*.

Independent variables are also called treatments or interventions; hence corrective feedback is the independent variable. If researchers seek to ascertain the effectiveness of certain corrective feedback on the written accuracy in writing performance, they have to prepare at least two groups (a treatment and a control group) and manipulate the corrective feedback (i.e., independent variables). This manipulation is necessary to confirm that the results are attributable to the independent variables (e.g., corrective feedback).

These independent variables are used even beyond experimental studies. In the creation of the corpora, the kinds of tasks, topics, and writing time would serve as independent variables. For example, the ICNALE corpus sets two argumentative writing tasks and requires participants to engage in the writing tasks in 30 minutes.

The dependent variables, on the other hand, are the outcome variables or scores. They are produced by rubrics, measures, and Likert scales. For example, when researchers use a measure of the written accuracy to examine the effectiveness of the corrective feedback, the scores produced by measures of written accuracy are the dependent variable and would reflect the degree of accuracy in the writing performance. Therefore, although researchers have to manipulate the independent variables to identify the cause, they also have to select appropriate dependent variables (e.g., measures) to understand how corrective feedback is effective. As another example, the dependent variables are important in investigating the development of writing performance using a corpus. If researchers seek to investigate written accuracy development with increasing writing proficiency, they have to produce the scores in writing performance in each proficiency group.

However, it should be noted that researchers can only know the degree of accuracy through the numbers or scores and cannot see the ability directly. These invisible characteristics and abilities reflected in test performance are called *constructs* (Cronbach & Meehl, 1955). For example, written accuracy is a construct defined as "the ability to produce the target-like and error-free language" (Housen et al., 2012, p.2). As shown in Figure 2.1, the written accuracy has been manipulated by a variety of measures (e.g., EFCR and EFTR).

Additionally, it would be difficult for researchers to know the appropriate measures. It is possible that a measure that a researcher chooses may not reflect the intended construct. If so, the results or outcomes obtained from the study would be meaningless for judging the effectiveness of the independent variables.

Furthermore, even if a measure is relevant to the intended construct, it may be narrow in the scope that it reflects. This problem is called *construct underrepresentation;* in such cases, the assessment remains incomplete because important aspects cannot be included (Messick, 1996). Messick also claimed that *construct-irrelevant variance*, which are factors not related to the construct, would be not desirable in the measurement.

These problems mentioned above are the range that the measures would reflect. There are other cautions for the measurement: the coherence of the scores. If the scores produced by raters are significantly different between raters, it would be difficult to trust the scores because the scores would be changeable at some points. Therefore, researchers could not draw a coherent conclusion in their study.

In summary, it will be apparent that there are many problems in obtaining the dependent variable from the measurement tools. Therefore, it is important to investigate the quality of the measurements to obtain reliable results. In language testing areas, the quality of the measurements is examined from two perspectives: *validity* and *reliability*.

2.4.2 The Introduction of Reliability and Validity in Language Testing

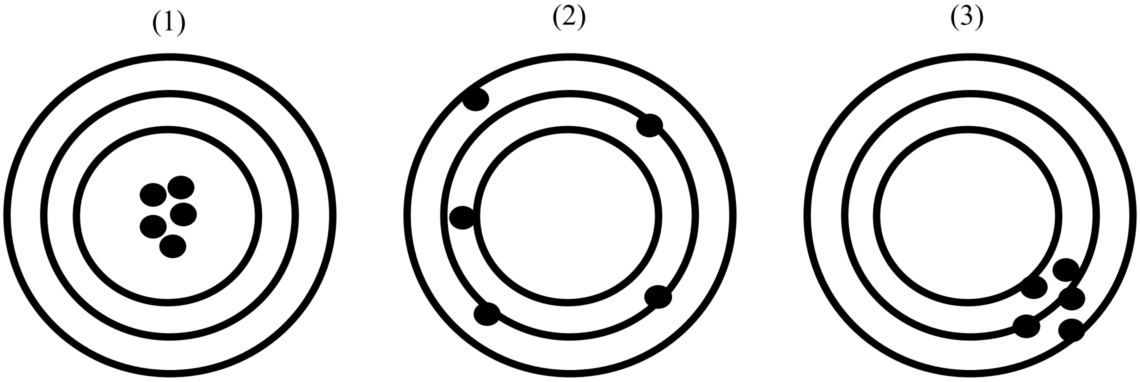
In this section, I will review the definitions of reliability and validity and then present an overview of their history. These concepts have long been considered by researchers in the field of language testing (e.g., Messick, 1996).

As mentioned above, the two concepts are important to examine the quality of measurement (i.e., validity and reliability). The reliability refers to the degree of consistency among the scores across different times, raters, and test forms (e.g., Bachman, 1990). Meanwhile, validity, overall, refers to the degree to which the measuring instrument measures the intended construct. That is, a measure is deemed valid when it measures what it says it measures (Polio & Friedman, 2016). The definitions of validity have been discussed for decades; the details of these discussions are outlined below.

The relationships between these two important concepts have sometimes been discussed with an analogy of a dartboard (Figure 2.3).

Figure 2.3

Dart Analogy in Reliability and Validity



Suppose a person shoots five arrows toward the center of the dart. In example (1) in Figure 2.3, most of the dots are at the center of the dartboard, and the range of the dots is close, meaning that the measurement is valid and reliable. On the other hand, example (2) shows that all dots are placed in different places. This means that the measurement is not valid and reliable. Finally, in example (3) in Figure 2.3, most of the dots are at similar points but are not at the center of the dartboard. This result suggests that the measurement would be reliable but not valid. When researchers draw a firm conclusion from their study, the measures should be valid and reliable.

2.4.3 Overview of Reliability and its Approaches

First, the reliability is reviewed. As mentioned above, reliability means the quality of test scores (Bachman, 1990). If a test score is reliable, the score would be similar when the test is evaluated across different times and raters. For example, if a learner obtains a high score on a writing test one day and a low score on the same test three days later, the score would not be a reliable measure for inferring the learner's ability. The reliability of a multiple-choice test is also important. If the test is reliable, it is assumed that a participant obtains a similar score at different times.

When investigating reliability, the *classical test theory* (CTT) has been applied. The CTT assumes that an observed score is affected by a *true score* and an *error* (or a *measurement error*). According to studies (e.g., Bachman, 1990; Brown, 2017), the true score is derived from a learner's ability, and the error indicates factors that are not relevant to the learner's ability (e.g., test conditions, temperature, and raters). The relationship is described in equation [1].

$$\text{True score} = \text{Observed score} + \text{Error} \qquad [1]$$

Similarly, the variance of a set of test scores is characterized below:

$$\text{Total test variance} = \text{True score variance} + \text{Error variance} \qquad [2]$$

According to equation [2], the total test variance consists of the true score variance and error variance (Brown, 2017).

The second assumption is that error scores are random or unsystematic, and the scores are not correlated with the true scores (Bachman, 1990). If the assumption were not true, researchers could not distinguish between the true scores and error scores. In the CTT, the reliability would defined as follows:

$$\text{Reliability} = \frac{\text{True score variance}}{\text{True score variance} + \text{error variance}} = \frac{\text{True score variance}}{\text{Total test variance}} \qquad [3]$$

Equation [3] suggests that the closer the true score variance and total test variance are, the higher the reliability coefficient is.

However, it is impossible to calculate the reliability since the true score variance is used to produce the coefficient. Therefore, we cannot *know* the reliability but can *estimate* the reliability of the observed score. In the CTT, the ideas of the *parallel test* or *items,* which are two tests, were introduced. This idea is based on the assumption that parallel tests measure the same ability. Clearly, it is impossible to know the true score; hence we might not be able to know whether the two tests are truly parallel. The CTT defined the parallel test from two points. If a test is a parallel, (1) the true score on one test is equivalent to the true score on the other test, and (2) the error variances for the two tests are also equivalent.

According to Bachman (1990), based on the assumptions, the three equations would be derived as:

$$\mathrm{M_a = M_b} \qquad [4]$$

$$\mathrm{Var_a = Var_b} \qquad [5]$$

$$\mathrm{Cor_a = Cor_b} \qquad [6]$$

The small letters in the equations (i.e., a and b) mean tests A and B. If the two tests are parallel, which indicates same ability, the mean scores and variances are equal as shown in equations [4] and [5]. Therefore, the correlations between tests A and B should also be equal [6]. Furthermore, error scores on the two tests are assumed to be random and not to be correlated. If the errors would influence the observed scores, the correlation between tests A and B would be lower. However, if the influence of the errors is small, tests A and B are highly correlated. Therefore, if the observed scores between tests A and B are highly correlated, this result suggests that the influence of the errors is minimal. Thus, this correlation would be a reliable indicator of the ability to be measured.

Based on the ideas in the CTT, a variety of methods are evolved to estimate the reliability of a test and items (e.g., Cronbach's α, split-half reliability, and Spearman-Brown split-half estimate). In this section, Cronbach's α is reviewed because the coefficient is also related to the writing performance evaluation.

Cronbach (1951) invented a general formula for estimating internal consistency, called Cronbach's α:

$$\alpha = \frac{t}{t-1}\left(1 - \frac{\sum Var^2{}_i}{Var^2{}_x}\right) \qquad [7]$$

α = the coefficient of Cronbach's α

t = the number of items

ΣVar^2_i = the sum of the variance of the different parts of the test

Var^2_x = the variance of the test scores

Other methods, such as Kuder-Richardson reliability and the Guttman split-half of estimate, estimate the reliability coefficient by dividing items into two blocks and showing the different coefficients if the splits were different. In contrast, Cronbach's α overcomes the limitation and can estimate the reliability by calculating the scores obtained from as many splits tests as possible.

As for the evaluation by raters, Cronbach's α has been used in many studies in writing. For example, Cronbach's α can be used when two or three raters evaluate the written accuracy with a rubric which was produced by 100 participants. It is necessary to investigate how consistent raters evaluate the written accuracy. When producing Cronbach's α coefficient, equation [7] can be used although the meaning of terms is different. According to Ebel (1961), t is the number of raters, Var^2_i is the variances of the ratings for a rater, Var^2_x = the variance of the summed ratings.

The Cronbach's α coefficient would be useful since it can be produced in one test. However, there are some problems with reliability: source of errors. In the CTT, the error score is regarded as one score which cannot be divided. However, the source of errors can be divided into two types: *systematic* and *random* errors. The systematic errors are such as raters, tasks, and the rating criteria. The combinations errors can be a source of errors, for example, the Rater × Task errors. On the other hand, random errors are the ones that might occur randomly. For example, the temperature, the health conditions of test-takers, and the motivation can be random errors. While the observed scores might be affected by a variety of errors, the CTT has the limitation that the framework cannot divide the source of errors in test scores (Gebril, 2010).

To overcome the limitation, *generalizability theory* (G theory) can be useful because the G theory can divide the source of errors into systematic and random errors (e.g., Shavelson & Webb, 1991). In the following, an overview of G theory will be given using specific experimental designs because the present dissertation used the G theory (see Study 1).

In a study, fifty participants worked on two writing tasks with different topics, and two raters measured the participants' writing proficiency using a holistic evaluation rubric with a five-point scale (Table 2.2). For example, while rater A provides participant A with four points on task A, rater B provides participant A with three points on task A. Finally, two raters assigned scores to all the English essays.

Table 2.2

Example of Evaluation

	Task A		Task B	
ID	Rater A	Rater B	Rater A	Rater B
1	4	3	5	4
2	2	3	3	2
⋮	⋮	⋮	⋮	⋮
49	4	4	3	2
50	2	1	4	3

In this situation, seven sources of errors can be set, as Table 2.3 shows. By using the G theory, the influence of errors and combined errors (e.g., $p \times r$) can be estimated.

Table 2.3

Summary of Error Sources

Sources	Interpretation
Participant (*p*)	The difference in participants' ability
Rater (*r*)	The difference in severity of ratings among raters
Task (*t*)	The difference of tasks
$p \times r$	The extent to which performances are rated differently by each rater
$p \times t$	The extent to which participants perform differently based on each task
$r \times t$	The extent to which raters differ on scores of each task
$p \times r \times t$	Interaction among Participant, Rater, and Task

When using the G theory, there are two types of sections: *G study* and *D study*. The G study focuses on the calculations of the influence of errors. The analysis of variance (ANOVA) is used to estimate the influence of each source in Table 2.3 by partitioning the total variation of sources into each variable, such as the Rater factor (Shavelson & Webb, 1990).

On the other hand, the D study estimates the reliability coefficients. In the G study, there are two types of reliability: *generalizability* and *index of dependability* (Schoonen, 2013). The generalizability coefficient is calculated by equation [9] (Brennan, 2001):

$$\mathrm{G} = \frac{\mathrm{Var}^2(p)}{\mathrm{Var}^2(p)+\mathrm{Var}^2(\delta)} \qquad [9]$$

and can be used for *relative decisions*. For example, if the purpose of an evaluation is to rank order participants, the generalizability coefficient can be appropriate. The index of dependability can be used for *absolute decisions*. For example, if a teacher wants to know whether a student passes a criterion or not.

In sum, an overview of reliability and its estimation methods were reviewed, and the CTT framework has made it possible to estimate reliability using Cronbach's α. On the other hand, the drawback of CTT is that it fails to distinguish and treat errors. Therefore, generalizability theory was developed, and it became possible to know in detail, the kinds of errors affecting the variance of scores. While a highly reliable assessment is a demand in measurement and evaluation, there is another important aspect: validity. In the following sections, I present an overview of validity, its history, and methods of verification.

2.4.4 Overview of Validity and its Approaches

While the reliability reflects the score quality (Bachman, 1990) and is a necessary aspect in language testing, the validity is also important to infer the learners' ability with the test scores. In this section, the changes in the definitions of the validity will be reviewed chronologically. According to Kane (2001, 2013), there are three models of validity: the criterion-based and content-based model, the construct model, and the argument-based model (or approach). In this section, we review how researchers investigate the validity and what kinds of components they proposed.

In the 1940s and early 1950s, the validity model was regarded as the criterion-based and content-based model (Kane, 2001). Validity was thought to be what we now call criterion-related validity (Strauss & Smith, 2009). In addition, with the establishment of statistical methods capable of investigating criterion relevance among tests, criterion-related validity has become an absolute method for investigating validity (Kane, 2013). As for the methods to investigate the validity, Anastasi (1950) claimed that "It is only as a measure of a specifically defined criterion that a test can be objectively validated at all…. To claim that a test measures anything over and above its criterion is pure speculation" (p. 67). For example, to determine the validity of an intelligence test, the criterion was whether or not the test was highly relevant to external standards. Hence, the question of what construct an intelligence test measures was ignored. This idea at the time was described as "the correlation between the actual test scores and the 'true' criterion score"(p. 623) in the first edition of *Educational Measurement.* It can be said to have had a significant influence on the validity ideas of that era (Cureton, 1951).

However, it has begun to be pointed out that there are several problems with the method of validation based on the relationship between tests (e.g., Strauss & Smith, 2009). The first problem is that the quality of the criteria used to investigate the relevance of a test cannot be guaranteed. This is because the creation of the external criteria was based on some judgments (vague diagnostic classifications, teacher evaluations), and these judgments had to be made on a knowledge base that was not well developed. The second problem is that the theory is not expected to be developed. In this framework, the validity of a test focuses only on whether or not it is highly predictive of external criteria. Also, as Anastasi (1950) says, it is only speculative to think that a test measures more than a criterion. Hence, despite significant effort to research validity, it has been difficult to develop theories.

In the middle of the 1950s, a validation method based on content was also proposed (e.g., Guion, 1977). According to Guion, this validation method was also described in the *Technical Recommendations* issued in 1954. It is a method for verifying the appropriateness of what a test measures, but the judgment depends on the evaluator's bias and subjectivity.

As Kane (2001) noted, in the 1950s, the American Psychology Association (APA)

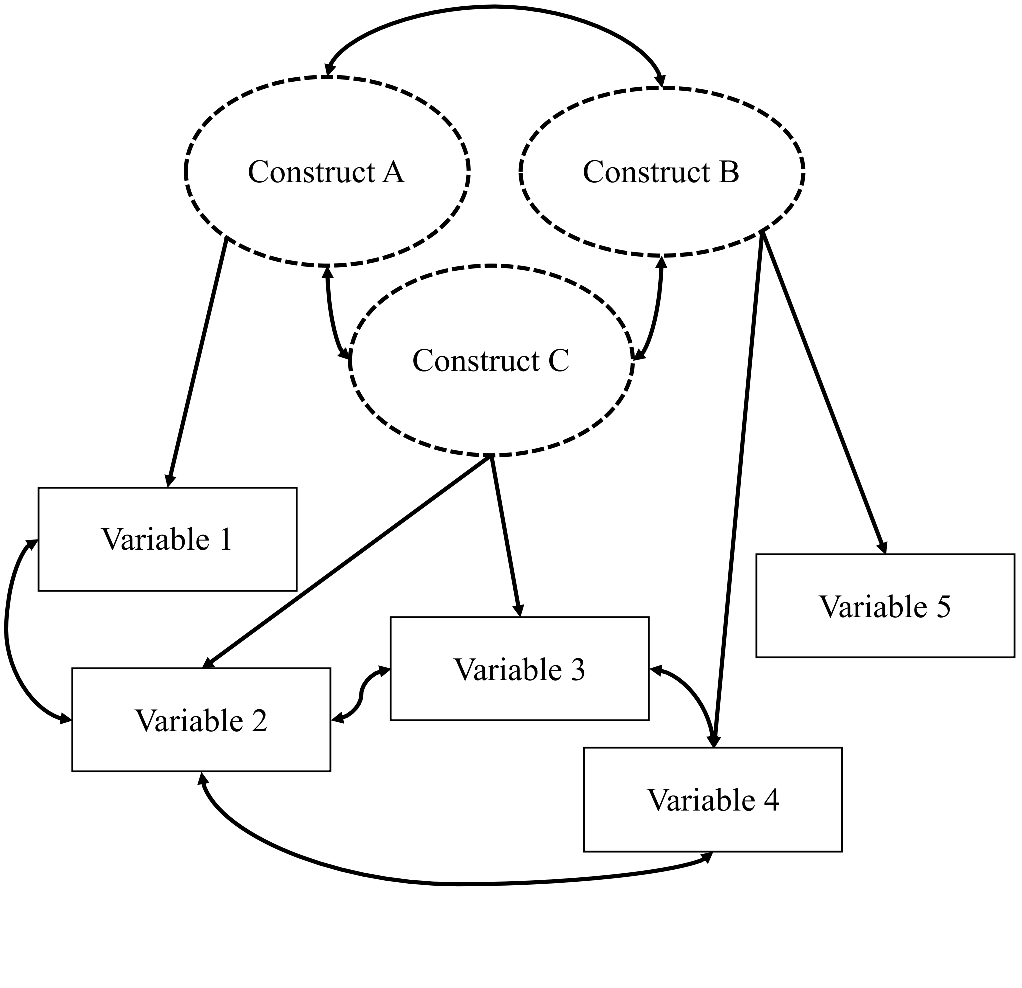
Committee pointed out the need for a broader view of validity at that time. Thereafter, the members of the committee (e.g., Paul Meehl) proposed the idea of construct validity, which they described in their *Technical Recommendations* (APA, 1954), which was mentioned above. Moreover, the idea of the construct validity was further developed by Cronbach and Meehl (1955). The construct model, coined by Kane (2013), was an important model for validation.

According to Cronbach and Meehl (1955), Technical Recommendations divided validity into four types: predictive validity, concurrent validity, content validity, and construct validity. The predictive and concurrent validity is similar to criterion-related validity. The major difference is that while predictive validity focuses on the accuracy of prediction by using the same measures before and after a given time period, concurrent validity focuses on the correlations to the external criteria. In addition, the content validity is confirmed positively by showing whether items in a test would reflect an intended sample. For example, if researchers want to measure the ability to write documents in English in a company, they need to compare and examine the content of the writing documents activities in that company. Finally, the construct validity (or validation), which Cronbach and Meehl (1955) mainly focused on, "is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not 'operationally defined.' (p. 175). "

Cronbach and Meehl (1955) developed construct validity, which was theoretically supported by the hypothetico-deductive model that was central in the 1950s (e.g., Kane, 2001; Murayama, 2012). In this model, a theory is considered to be a system consisting of several axioms (Suppe, 1977). The set of axioms that connect the concepts defined by the theoretical constructs is considered to be the center of the theory. Some axioms are then interpretable by associating their constructs with observable variables. Once an axiom becomes interpretable, it can make predictions about observable relationships between variables. This theoretical relationship between constructs and the correspondence between each construct and measurement is called the *nomological network* (Figure 2.4).

Figure 2.4

Relationships Between Constructs and Variables



As Kane (2001) explained, the concept of composition is not always explicitly defined. In other words, not all constructs are defined based on observation. The validity of linking constructs to the interpretation of scores depends on how well the scores satisfy the theory. Suppose the observations are consistent with the theory, it means that both the validity of the theory and the measures used to estimate the constructs defined by the theory have been validated. On the other hand, if the observations do not agree with the theory, then parts of the nomological network will be modified.

The *Standard*, which is a test guideline made by APA, the American Educational Research Association (AERA), and the National Council on Measurement in Education (NCME), focused on the three types of validity (content validity, criterion-related validity, and construct validity) (APA, AERA, & NCME, 1966), meaning that the *trinitarian view* on the idea of validity was dominant and was a decisive trigger for the spread of the idea that validation is nothing more than the examination of these three types of validity. (Murayama, 2012).

However, it has gradually become clear that this framework has three main problems (e.g., Landy, 1986; Murayama, 2012). The first problem is that validation can be completed with a simple survey (Landy, 1986). While the framework of three types of validity is easy for researchers to understand, even a superficial survey on the three types of validity is enough to ensure validity (Murayama, 2012). In fact, it is clear that in order to satisfy construct validity, not only correlation analysis, which analyzes the relationships among variables, but also analyses such as factor analysis and structural equation modeling, which estimate the constructs reflected by the observed variables, are required.

The second problem is that the differences between the three types of validity are ambiguous, making it difficult for the researchers to distinguish between them (Murayama, 2012). For example, content validity is examined in terms of whether the content of the test is consistent with experts' assumption. However, these investigations would examine the match between the observed data and the theory. In other words, it is very similar to a construct validity study, which examines the degree of agreement between data and theory.

Finally, Murayama (2012) pointed out that this framework cannot be examined at all with respect to whether or not the three types of validity are appropriate. Many studies have investigated validity from a variety of perspectives (e.g., factorial validation). It is unclear whether these perspectives are included in any of the three types of validity. As these limitations were highlighted, support for the idea that construct validity research is itself validity research, came to the fore (e.g., Kane, 2001).

Around the end of the 1970s, as Brennan (2013) said that "the construct validity is all of the validity (p. 75)", the investigation of construct validity expanded to what validation was all about. The idea of incorporating three types of validity into construct validity has become widespread. This idea did not suddenly appear at the end of the 1970s. It had been conceived in the 1950s, as Loevinger (1957) claimed that "since predictive, concurrent, and content validities are all essentially ad hoc, construct validity is the whole of validity from a scientific point of view" (p. 636).

However, the unified concept had some drawbacks (e.g., Brennan, 2013; Kane, 2001, 2013; Strauss & Smith, 2009). In the framework of Cronbach and Meehl (1955), the hypothetico-deductive method is used to construct a theory of validity. In this framework, a theory is assumed to consist of a number of axioms. The group of axioms that connect the words defined as theoretical constructs was assumed to be the center of the theory (e.g., the nomological network). The framework version based on highly formal and theory-dependent values is called "strong program" (e.g., Kane, 2001). While the strong program version is elegant, many social sciences do not have such a set of axioms, unlike hard sciences such as physics and mathematics, as Cronbach and Meehl (1955) recognized.

In contrast, there was the other program, which was called the "weak program" version (Kane, 2001). The characteristic of the program is that any evidence can be relevant to validity. In addition, there is no explicit guidance for identifying the appropriate evidence. Therefore, Kane (2001) described that the weak program of construct validity pulled everything and did not provide researchers with any suggestions. Some studies agreed that these developments of two competing versions might be inevitable (e.g., Brennan, 2013; Kane, 2001, 2013). The strong program version of construct validity was made based on theory and elegance, but the applications to education and social sciences might be difficult. Brennan (2013) also explained that evaluation became a public and large-scale activity in the 1960s. Therefore, these evaluations appeared to be pragmatic and would not match the strong program version of validity. As a result, the weak program version took on much of the vagueness or abstractness of the strong program version without the base of the formal theory. Therefore, the construct validity resulted in “sheer exploratory empiricism (Cronbach, 1988, p.12).”

The unified concept of construct validity did not provide researchers with any explicit guidance (Kane, 1992, 2001). The *Standards* in 1985 (AERA, APA, & NCME, 1985) described the unified concept of construct validity. Still, the explanations were limited to the general discussion in terms of three types of validity (i.e., content, criterion, and construct validity).

While some scholars criticized the Standards in 1985 (e.g., Messick, 1988), it regarded validity as the inferences based on test scores and focused on the test score use (Shimizu, 2004). According to the *Standards* in 1985, the validity is “the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores (AERA, APA, & NCME, 1985, p.9).” Shimizu also suggested that the definition of validity in the *Standards* was influenced by Samuel Messick, a member of the *Standards* development team. Messick (1989) defined validity as “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of methods (p. 13).”

In the definition proposed by Messick, there are points worth mentioning (e.g., Kane, 2006, 2013; Murayama, 2012). In particular, Murayama (2012) summarized key points of Messick’s definition. The first point is that validity is not something that is inherent in testing. In the conventional validity framework, it is claimed that "the test has been validated. However, based on Messick’s definition, this claim is false, because validity is a judgment about “the interpretation of test scores”. Depending on the purpose and interpretation, the degree of validity may vary (Cronbach, 1971). For example, if a writing test shows a ceiling effect, it may not be a valid test if its purpose is to measure the relative ability of test takers. However, if the test is used as a criterion-referenced test, at least the ceiling effect is unlikely to cause the test to be less valid (e.g., Murayama, 2012).

The second point is that evidence of validity (e.g., criterion-relevant validity or content validity), which was previously considered to be a different type of validity, is now seen as evidence for assuring construct validity. Examining correlations with other tests and checking item content with experts became another form of evidence for more appropriate interpretation of test scores. In addition, Messick’s definition of validity eliminated the need to answer the question of whether three types of validity are sufficient. Depending on the interpretation of the test scores, more evidence may be needed.

So far, we have reviewed the historical background on validity. It can be seen that we have arrived at today's definition by solving various problems related to the idea of validity one by one. In the context of CAFs, the scores obtained from the measures are also used to infer the abilities that learners possess. For example, by using the accuracy measures to measure accuracy in English writing and calculating the scores, researchers estimate English learners' accuracy level. However, there are many accuracy measures that are supposed to reflect accuracy (e.g., EFCR and EFTR). Depending on the type of measure, it is quite possible that the scores obtained from the indicator may be misinterpreted. Therefore, it is necessary to examine which of the many measures is appropriate in different contexts. In order to achieve this goal, it can be said that it is necessary to conduct validation according to Messick's (1989) definition of the interpretation of test scores. This study will conduct a validation study for written accuracy measures (i.e., WCR) according to Messick's (1989) definition. In the next section, we will review the methods of validation.

2.4.5 Approaches to Validation

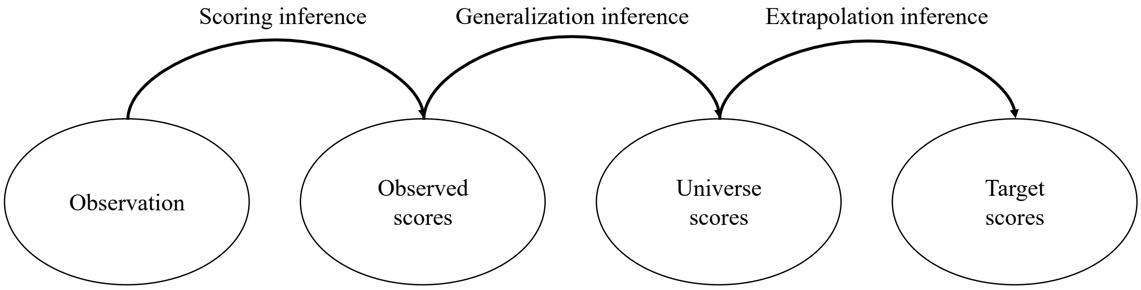
This chapter provides an overview of validation methods that have been proposed based on Messick's (1989) definition. These include the works of Bachman and Palmer (1996), Chappelle et al. (2008), Kane (1992), and Weir (2005).

First, we introduce the argument-based approach proposed by Kane (1992). This approach, which has been applied in a number of studies in recent years (e.g., Chapelle et al., 2008), originated from the ideas of Cronbach (1980) and House (1980). Cronbach (1988) claimed the necessity of *the logic of evaluation argument*, which was proposed by House (1977), when the importance of argument began to attract attention in the field of language testing. Moreover, Cronbach (1998) also proposed a *validity argument*, which gives an overall evaluation of the interpretation intended by researchers, as well as on the way test scores are used. Later, Kane (1992) proposed an *interpretive argument*, presenting inferences and assumptions and laying out the intended interpretations as a chain of inferences. The interpretive argument plays an important role in gathering evidence which would support intended score interpretations (Bachman, 2005).

Kane et al. (1999) showed an example of an interpretive argument based on the discussion in Kane (1992). They showed three inferences (i.e., scoring, generalization, and extrapolation) in the context of high-stakes assessments in order to examine the usefulness of the argument-based approach. The first one is *scoring inference*, which goes from actual performance to observed scores. The inference can be confirmed positively by the two assumptions: (1) criteria are appropriate and used as intended, and (2) the elicited performance can be interpreted as intended. The second is *generalization inference*, which links observed scores to universe scores. The necessary evidence for the generalization inference is that the scores would be equivalent to the scores on multiple tasks similar to the test in the assessment. The third is *extrapolation inference*, which goes from the universe scores to *target scores* referring to an interpretation of what a test taker knows or can do (Kane, 1992). In Kane et al. (1999), a validation study presents a number of inferences and can be done by a chain of reasoning (Figure 2.5).

Figure 2.5

Three Inferences in Kane et al. (1999)

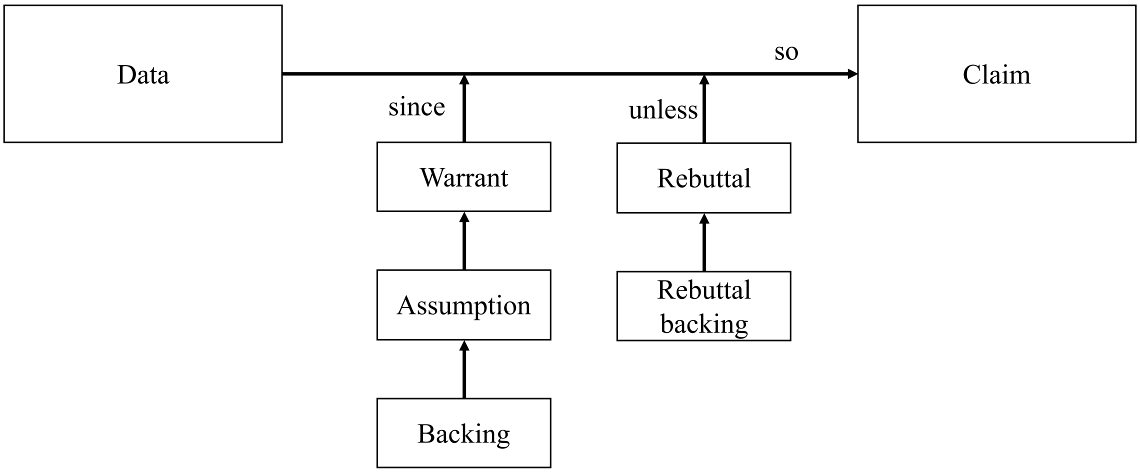


The argument-based approach proposed by Kane (1992) is a framework easily understood by researchers and test developers because it specifies the types of inferences and the assumptions that support them. While there were detailed explanations of interpretive arguments and inferences (e.g., Kane et al., 1999), few explanations of theories about the argument itself can be found. As Kane (1992) only explained that inferences and assumptions constitute an argument, no model about arguments was referred.

Mislevy (2003) proposed that language testing researchers usually rely on argument structures, which were developed by Toulmin (2003). From this point on, we will overview the structure of the argument, following Toulmin's (2003) description. His idea of an argument is to make claims on the basis of data and warrants (Figure 2.6).

Figure 2.6

Structure of an Argument



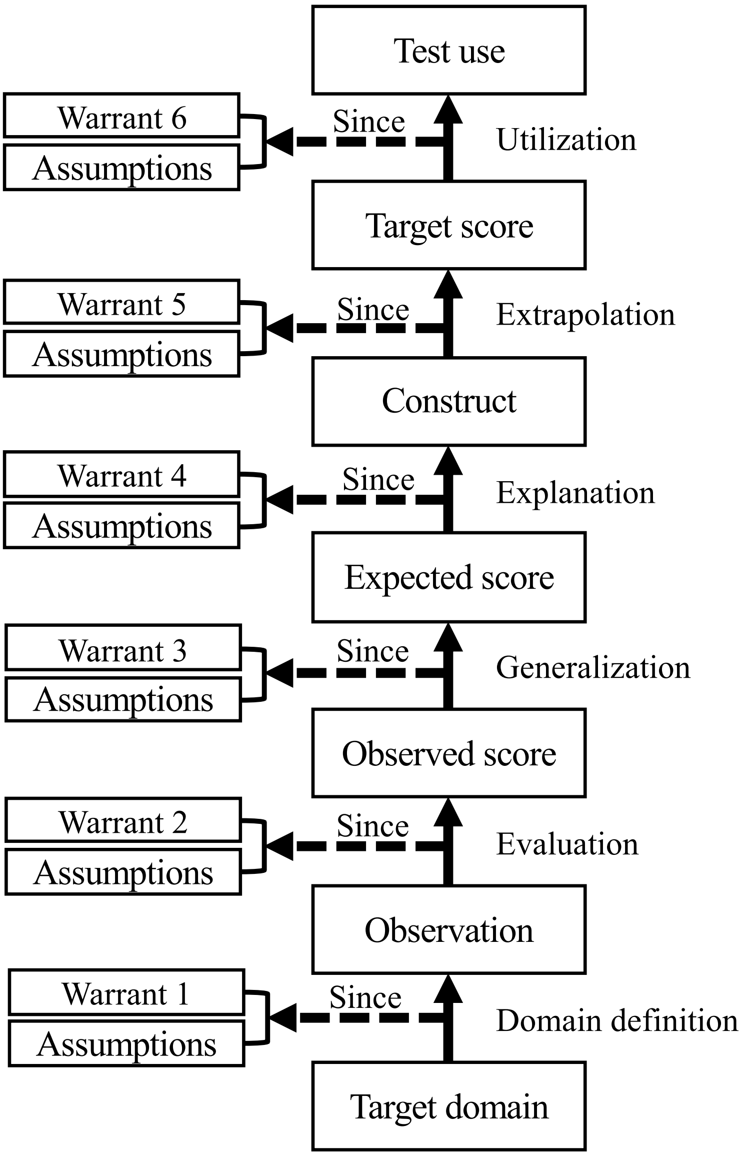
According to Toulmin (2003), the *data* consist of "information on which the claim is based (Toulmin, 2003, p. 90)." In the context of language assessment, the data would be the performance which test-takers produced, such as writing performance. It should be relating to a *claim*, which is "a conclusion whose merits we are seeking to establish"

(Toulmin, 2003, p. 90). Each claim is regarded as the *interpretation* that researchers choose (e.g., Bachman, 2005). That is, the claims are built based on the data in Toulmin's model. However, a *warrant* and *backing* are necessary to establish the claim from the data. A warrant is "a general statement that provides legitimacy of a particular step in the argument" (Toulmin, 2003, p. 92). The warrants justify the inference based on a certain data to a certain claim (Mislevy, 2003). A backing is "other assurance, without which the warrants themselves would possess neither authority nor currency" (Toulmin, 2003, p. 96). The backings would come from a particular theory, prior experience, or analysis (e.g., Bachman, 2005). However, if a rebuttal is shown in a validation process, the inference link between a data and a claim would become weak. If a rebuttal is proven, a stronger backing would be necessary to confirm an inference (Kane, 1992).

In later years, a modified version of the model given in Kane (1992) and Kane et al. (1999) was proposed (e.g., Chapelle et al., 2008). In Chapelle et al.'s (2008) model, three new inferences were added (Figure 2.7): domain definition inference, explanation inference, and use inference.

Figure 2.7

Framework of Argument-Based Approach in Chapelle et al. (2008)



Domain inference is the inference that connects the observations obtained in the test from the target domain. This reasoning allows for the creation of tests that clarify the constructs. In addition, *explanation inference* represents the reasoning that connects constructs from scores that can now be generalized. This reasoning makes it possible to measure, with an awareness of the constructs. Finally, *utilization inference* links target scores to test use.

This inference makes it possible to use test scores in a purposeful manner.

Following Messick's (1989) definition, Kane (1992) advocated an argument-based validity approach, while other researchers advocated validation models following Messick's (1989) definition (e.g., Bachman & Palmer, 1996, Weir, 2005). Bachman and Palmer (1996) pointed out that what matters most in test development and test specification is the usefulness of the test. In addition, it is said that the usefulness can be expressed by an equation consisting of six components [11].

Usefulness = reliability + construct validity + authenticity + instructiveness + impact +practicality [11]

Reliability represents consistency of measurement as reviewed above. Constructs represent the extent to which or whether a given test score can be interpreted as a construct of the ability intended to measure. Authenticity refers to whether the characteristics of the task in the domain of use are consistent with the characteristics of the task used in the test. Impact refers to the effect of the test being developed on society and education. Finally, practicality refers to the relationship between the resources that will be needed in the design and development of the test and the resources that will actually be available. While several studies have been conducted using the Backman and Palmer (1996) model (e.g., Chapelle et al., 2003), difficulties in validating individual constructs (e.g., authenticity) were also noted (Bachman & Palmer, 2010). The checklist created to satisfy the individual components had a total of 42 items. Therefore, it was very difficult to fulfill all the items.

Weir (2005) developed a *socio-cognitive framework* for a test development and validation study. He presented five validities that are involved in the overall process of the test taker responding to the test and the raters scoring it to produce a score: context validity, cognitive validity, scoring validity, consequential validity, and criterion-related validity. However, a limitation is that it is difficult to see how validity is assembled as a whole (Koizumi, 2018). The model of Weir appears to be similar to the six-factor framework proposed by Messick.

2.4.6 Measuring Accuracy Based on Argument-Based Validation

This section describes measurement issues in writing development research (especially accuracy) using CAFs, based on the theories of validation reviewed above. Many studies have used CAFs to elucidate developmental patterns of writing proficiency and linguistic traits (e.g., Barrot & Gabinete, 2019). Many of the studies that have set this objective have often used complexity (e.g., Biber et al., 2016). This is because complexity is also closely related to intermediate language (Housen et al., 2012) and is significantly useful for predicting writing proficiency and English proficiency.

Although no studies have been conducted on the measurement of complexity based on a validation approach, it can be said that each of these studies has been tested for reliability and validity. For example, Lu (2010) created a program that can automatically produce complexity measures' values. As a result, its reliability can be said to be very high. This is highly relevant to scoring inference in argument-based validation approaches. Research has also been recently conducted on which constructs that are assumed to reflect complexity actually reflect it (e.g., Kato, 2019). This research would fall under explanation inference. In addition, there is a great deal of research on the relationship between complexity measures and writing proficiency and English proficiency. These studies are highly relevant to extrapolation inference. It can be confirmed that even for the complexity measures that have been studied, there have been no inputs on generalization inference or utilization inference, but previous studies have partially verified the content related to validity.

In contrast, there are few studies on the measurement of accuracy using accuracy measures or on development (Michel, 2017). Very few validity studies have focused on the WCR, which is assumed to be one of the accuracy measures (e.g., Evans et al., 2014); Evans et al. (2014) investigated the validity of the WCR using three accuracy measures (EFCR, EFTR, and WCR) In that study, the validity of inferences using scores from the WCR is satisfied when two aspects are met: (1) the measure reflects how accurate the English composition is, and (2) the measure can effectively discriminate the level of accuracy in the learner. The validation framework used by this study then consisted of three perspectives (content validity, criterion-related validity, and construct validity).

English compositions (N = 4) produced by 97 ESL learners were sampled in the study. A multi-faceted Rasch model was used to test the reliability of the ratings and the degree of discrimination. Correlation analysis was also used to investigate associations between accuracy measures. Two raters participated in the evaluation. As part of the inter-rater training, the raters (1) received guidance on T-units, clauses, and WCRs, (2) rated 10 English compositions, and (3) discussed any differences or areas of disagreement in their ratings. The raters then assessed accuracy in 87 English compositions.

First, with regard to content validity, it is argued that the WCR measures accuracy based on the definition of accuracy and previous studies. Criterion-related validity was also met, as the results of the correlation analysis showed that the WCR correlated with two accuracy measures (EFCR: $r = .88$, EFTR: $r = .78$). Finally, because construct validity incorporates content and criterion-related validity (Evans et al., 2014), these two aspects are discussed. In addition, multi-faceted Rasch model analyses revealed that the WCR is able to discriminate learners' accuracy levels into approximately four groups. This value was higher than the EFTR (3.76) but lower than the EFCR (4.92).

The Evans et al. (2014) study was the first validation of the WCR. However, there are several points that need to be addressed when conducting developmental research using a learner corpus, such as in the present study.

First is the method of converting the scores obtained from the WCR: in the study, the WCR values were converted to integer values (0 to 10) in order to use the multi-faceted Rasch model, which only allows analysis of integers. However, in corpus-based studies, it is common to run analyses through the use of the original values of the WCR (e.g., Barrot & Adgeppa, 2021).

The second is the generalizability of the scores: Since they only examined rating severity among raters, it is not known to what extent the values are generalizable. If the generalizability is low, it is difficult to trust the values, since WCR scores may be calculated very differently in other situations.

Third, only correlation analysis was used to verify that the measures reflect the degree to which English writing is accurate. Correlations between measures alone do not tell us whether they reflect the same construct or not (i.e., accuracy). From the WCR scores, we can infer that for the purpose of validating the development of accuracy with a learner corpus, validation through an argumentation-based validation approach (Chappelle et al., 2008) is useful for three main reasons.

First, it allows us to systematically gather the evidence necessary to use the WCR for the purpose of validating the development of accuracy with a learner corpus. Argument-based validation sets up an interpretive argument that clearly specifies the inferences and assumptions necessary for the validity of the inferences of the scores. Next, based on the interpretive argument, evidence is presented to support the premises and a validity argument is constructed to evaluate the interpretive argument. These two arguments clarify the overall validity of the interpretation and use based on the test score;

in Evans et al., (2014) the degree to which learners discriminate between levels of accuracy is examined, but it may be difficult to tell which inferences this evidence is meant to satisfy.

Second, using validation through an argument-based validation approach allows us to test whether the accuracy ratings from the WCR provide more information to the investigators. Validity validation, based on the triadic view, does not examine what information scores and ratings give investigators. However, the extent to which the scores provide more information is a very important aspect of understanding the development of learners' abilities in detail, and while accuracy assessment with the WCR has the potential to do so, no relevant research has yet been conducted.

Third, although not limited to the WCR, the use of validation through an argument-based validation approach can clarify the context in which the measure can be used, and thus, it allows for comparisons across studies. It is clear that many types of accuracy measures exist; this has led to the use of different measures across studies. As a result, it is difficult to compare results across studies, and it is also difficult to integrate studies through meta-analysis. However, the discrepancies in the employed measures could be resolved by using a validation approach that examines the validity of the use and interpretation of the scores according to their purpose. Thus, I believe this study is a first step in this direction.

2.4.7 The Interpretive Arguments in the Validation Study

The validation study was based on the argument-based approach proposed by Chappelle et al. (2008). In this approach, there are six inferences to produce the validity argument (i.e., domain description, evaluation, generalization, explanation, extrapolation, and utilization). Here, the inference, warrants, assumptions, and ways of supporting the assumptions will be explained.

The domain description inference links written performances in the target domain to performance observations in the writing test. The warrant necessary to confirm the inference is that the observations in test performance are related to the relevant knowledge, skills, and abilities in situations that are representative in the target domain. In the context of written accuracy, the warrant is that the WCR represents the written accuracy of the writing performance obtained from the argumentative tests. In the context of the present study, it is necessary to show whether the WCR would be related to the written accuracy in the argumentative writing tests. Therefore, the assumption is that the WCR represents the written accuracy domain obtained in the argumentative writing tests. The literature review was used to provide the backings of the domain description inference.

The evaluation inference links the observed performance to the scores. The warrant is that observation of writing performance evaluated by the WCR can be noted as the observed scores. The assumption for the warrant is that when raters evaluate the accuracy using the WCR, the reliability of the evaluation is appropriate. Although the raters can classify the clauses by using the rating scale in the WCR, it is critical that the scores are reliable.

While reliability is an essential aspect in language testing (Bachman, 1990), some studies (e.g., Polio & Shea, 2014) claimed that even the current studies do not report the reliability coefficients and suggested that the reliability coefficients are necessary for rigorous research. The rating scale of the WCR is vague and developing. Therefore, it is necessary to investigate the reliability coefficient. In this study, Cronbach's α was used to examine the assumption.

It should be noted that in each validation study, the WCR and written accuracy measures analyzed in the previous studies were included because the validation study of the WCR should be conducted with the measures confirming the inferences.

The generalization inference links the observed scores of the WCR to expected scores, which refer to the scores one would expect to obtain across different tasks, tests, and rating conditions. The warrant is that observed scores of the WCR are estimates of expected scores over the parallel versions of tasks and across raters. The assumption is that a sufficient number of tasks and a rater is included to provide stable estimates of participants' performance. To examine the assumption, the present study used a generalizability theory.

While the reliability coefficients such as Cronbach's α can be useful since the value would be applicable to one test or performance, it is unclear whether the scores could be generalized. In addition, enough tasks and raters have not been conducted. Furthermore, the degree of influence of measurement error on scores has not been investigated when measuring accuracy using WCR.

The explanation inference links the expected scores of the WCR to a construct (i.e., accuracy). There are two assumptions in the explanation inference. The first assumption is that the relationship between a WCR and a construct (i.e., accuracy) corresponds to the theory. The present study used factor analysis to examine the first assumption. Moreover, the second assumption is that there is little variation in the scores due to text factors (e.g., the number of clauses) that can seriously affect the interpretation of the scores. As all measures are calculated by textual features, all measures in CAF would be affected by the textual features (e.g., the number of clauses); it might therefore be difficult to interpret the scores and the factor structures if the WCR is highly correlated with textual factors. The present study used a correlation analysis to examine the relationships.

The extrapolate inference links the expected scores of the construct to the target score, which can be obtained from the language performance over the test situations. The warrant is that expected scores of the construct are correlated with language performance over test situations. The assumption is that WCR is correlated with English proficiency. The present study used the CEFR levels as the English proficiency levels set in the ICNALE corpus.

In studies examining the development of linguistic and L2 proficiency using the CAF, performance and test-taker scores obtained from large-scale tests are often used. For example, a study by Kyle and Corssley (2016) used performance and scores obtained from the Test of English as a Foreign Language (TOEFL) to investigate the relationship between its complexity measures and writing proficiency. These studies can be used to predict writing proficiency and identify each proficiency level's characteristics.

Meanwhile, in recent years, studies investigating the completion of CAF with proficiency assessments such as the CEFR have received greater attention (e.g., Barrot & Adgeppa, 2021; Gaillat et al., 2021). The reason is that educational institutions (e.g., universities) often use the CEFR, and there is a growing need to understand the differences between the English proficiency levels (Hawkings & Buttery, 2010). This study uses the ICNALE corpus, a collection of English essays written by university students in Asian countries. This corpus differs from the standard CEFR in how it translates the levels, but the levels are tailored to Japanese and other Asian contexts. This corpus will be suitable for studying how Japanese EFL learners' accuracy develops as their proficiency level develops.

The present study used a correlation analysis to examine the relationships.

Finally, the utilization inference links the target score and test use. The warrant is that the target scores provide detailed information about language development. The assumption is that the WCR provides a more detailed picture of writing accuracy development in Japanese learners of English, than traditional measures. WCR can potentially subdivide writing performance and show differences between proficiency levels. In that case, the WCR can be recognized as a measure that provides detailed information about the development of writing proficiency. Descriptive statistics and non-parametric tests were used to examine the assumption. The summary of the interpretive argument is described in Table 2.4.

Table 2.4

Interpretive Argument in the Validation Study

Inference	Warrants	Assumptions	Support
Domain definition	The WCR represents the written accuracy of the writing performance obtained from the argumentative tests.	1. The WCR is representative of the written accuracy domain obtained in the argumentative writing.	1. Review of related literature
Evaluation	Observation of writing performance evaluated by the WCR can be noted as the observed scores.	2. When raters evaluate the accuracy using the WCR, the reliability of the evaluation is appropriate.	1. Cronbach's α
Generalization	Observed scores of the WCR are estimates of expected scores over the parallel versions of tasks and across raters.	3. A sufficient number of tasks and raters are included to provide stable estimates of participants' performance.	1. Generalizability theory
Explanation	Expected scores of the WCR represent a construct of the written accuracy.	4. The relationship between the WCR and a construct corresponds to the theory. 5. There is little variation in the scores due to texts factors that can seriously affect the interpretation of the scores.	1. Factor analysis 2. Correlation analysis

Interpretive Argument in the Validation Study (Continued)

Inference	Warrants	Assumptions	Backing
Extrapolation	Expected scores of the construct are correlated with language performance over test situations.	6. WCR is correlated with English proficiency.	1. Correlation analysis
Utilization	The target scores provide the detail information about the language development.	7. The WCR provides a more detailed picture of the development of writing accuracy development in Japanese learners of English than traditional measures.	1. Descriptive statistics 2. Non-parametric test

2.5 The Purposes of the Present Dissertation

The assessment of written accuracy has been developed from the previous studies (e.g., Evans et al., 2014; Foster & Wigglesworth, 2016). While such studies used some written accuracy measures based on the error-free units such as clauses, recent studies have begun using the WCR, which considers the influence of the linguistic errors, to measure the effectiveness of writing instructions (e.g., Barrot, 2021) and explore accuracy development (e.g., Barrot & Adgeppa, 2021) and task complexity (e.g., Michel et al., 2019). In particular, studies of the development of the written performance could be informative for researchers and teachers since they could not only reveal how L2 learners develop their performance as the proficiency increases but also how teachers should use their time for teaching. In the CAF framework, the studies of written accuracy are scarce (Michel, 2017), and the current situation is that it is difficult to discuss the development of accuracy.

While some studies used the WCR to examine the development of CAF using written corpus (e.g., Barrot & Adgeppa, 2021), there would remain some issues to be solved: the validation of the WCR. Although the validation studies focusing on the WCR have been conducted (e.g., Evans et al., 2014), these studies would not show the use of the WCR and would fail to provide the necessary evidence. Based on the recent ideas in validation studies, there would be no one-size-fits-all measures and tests. Therefore, it would be necessary to investigate the validity of the WCR for the development studies using the corpus.

Moreover, the developmental process of accuracy in Japanese learners of English is still incompletely understood. Although the development of specific grammatical features (e.g., tense) have been conducted (e.g., Abe, 2007), it is still unknown how Japanese EFL learners develop their accuracy as a whole and what nature of errors (i.e., Lv.1 in the WCR rating scale) are reduced. Therefore, the final purpose of the dissertation is to show how Japanese EFL develop written accuracy as English proficiency increases. To explore the development, the present dissertation conducted validation studies of the WCR (Table 2.5).

Table 2.5

Summary of the Dissertation

Validation study	**Study 1**	Reliability of measurements using the accuracy measures (Evaluation and generalization inference)
	Study 2	Relevance of factors and textual factors reflected by accuracy measures, including WCR (Explanation inference)
	Study 3	Relationship between accuracy measures and writing proficiency (Extrapolation inference)
		On the use of WCR in corpus-based accuracy development studies (Utilization inference)
Development study	**Study 4**	The development of accuracy in writing texts composed by Japanese EFL learners
	Study 5	The development of each clause of the WCR in writing texts composed by Japanese EFL learners

The validation study comprises three studies. First of all, Study 1 focused on the reliability of the measurement of the written accuracy measures. After checking the reliability, Study 2 examined the factor structures of the written accuracy measures. Then, Study 3 investigated the relationships between the written accuracy measures and English proficiency (i.e., CEFR). In addition to the investigation, Study 3 examined the usefulness of the WCR and showed extent to which it was useful in investigating written accuracy development using the corpus. The written accuracy development study was conducted after the validation study and had two sub-studies. Study 4 focused on the relationships between the WCR and other domain measures (i.e., complexity and fluency) and examined how and whether the development of the written accuracy correlated with the other domains. Then, Study 5 investigated how each clause in the rating scale of the WCR developed.

The dissertation provides two main theoretical implications: first, it indicates the availability of the WCR in the corpus studies; and second, it shows the development patterns of the written accuracy. The study is the first to clearly establish the purpose of use based on recent validation approaches for accuracy measures. The present study, especially the validation study, aids writing researchers in using the WCR with reliable evidence to investigate the development of the written accuracy using the corpus. Moreover, as this study can show the developmental patterns of how Japanese learners of English improve their accuracy while relating it to complexity and fluency, it can provide a solution to a particular issue in the CAF framework, "How is CAF component related? (Housen et al. 2012)."

As for the pedagogical implications, the present study provides specific points of focus for teachers in English classes. Based on previous findings, it would be difficult to identify the parts of the English texts produced by Japanese EFL learners, that teacher should spend time on. However, by showing written accuracy development using the WCR, the present study can provide English teachers with suggestions regarding where they should devote their time to improve learners' accuracy levels. The present studies of written accuracy development would also be informative when choosing a certain instruction method and enable optimum use of time.

Chapter 3

Study 1: Investigating the Reliability of Accuracy Assessment With Accuracy Measures

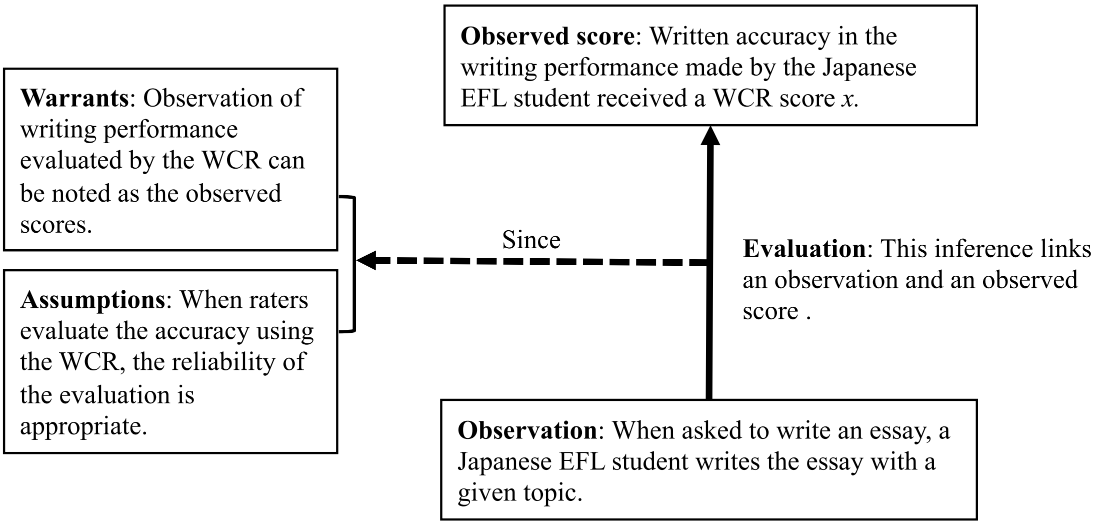
3.1 Evaluation and Generalization Inferences

3.1.1 Purposes and Research Questions

For Study 1, the evaluation inference was initially examined. The evaluation inference is necessary to claim that writing task performance observations are evaluated to provide observed scores (Figure 3.1).

Figure 3.1

Summary of Evaluation Inference

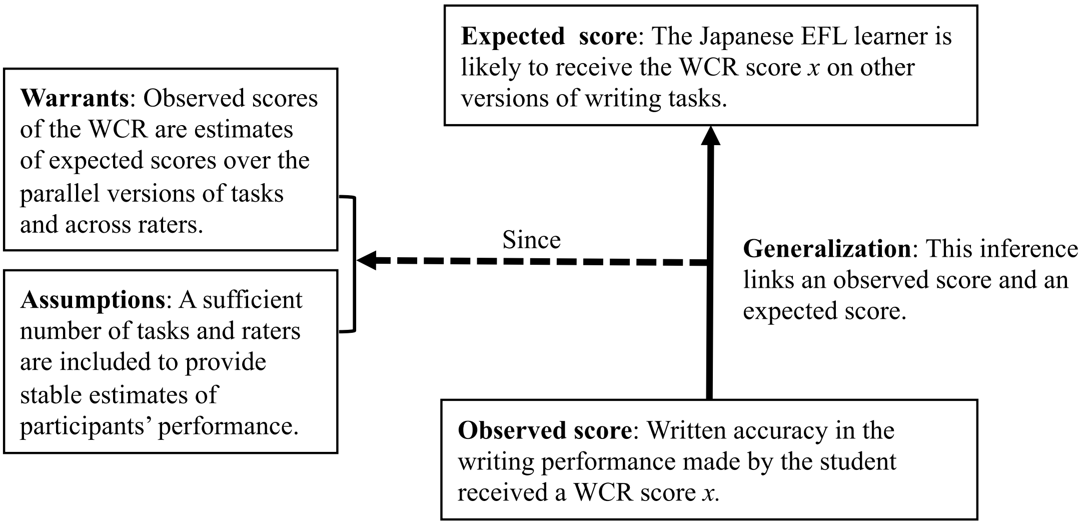

Observed score: Written accuracy in the writing performance made by the Japanese EFL student received a WCR score *x*.
Warrants: Observation of writing performance evaluated by the WCR can be noted as the observed scores.
Since
Evaluation: This inference links an observation and an observed score .
Assumptions: When raters evaluate the accuracy using the WCR, the reliability of the evaluation is appropriate.
Observation: When asked to write an essay, a Japanese EFL student writes the essay with a given topic.

In this study, Cronbach's α and the adjusted Cronbach's α was used to verify the inference. By using Cronbach's α, it is possible to show how consistent the evaluation of raters is. In addition, the adjusted Cronbach's α can provide information on how inconsistencies of evaluation affected inter-rater reliability.

Thereafter, the generalization inference was tested to claim that observed scores are estimates of expected scores over the relevant parallel versions of tasks and raters (Figure 3.2).

Figure 3.2

Summary of Generalization Inference


Expected score: The Japanese EFL learner is likely to receive the WCR score *x* on other versions of writing tasks.
Warrants: Observed scores of the WCR are estimates of expected scores over the parallel versions of tasks and across raters.
Assumptions: A sufficient number of tasks and raters are included to provide stable estimates of participants' performance.
Since
Generalization: This inference links an observed score and an expected score.
Observed score: Written accuracy in the writing performance made by the student received a WCR score *x*.

The present study applied the generalizability theory and investigated the reliability (i.e., G coefficient) with all accuracy measures. As mentioned in the literature review, it is also important to note here that the reliability and impact of factors affecting accuracy measures' scores have not yet been investigated. These elements matter because they impact the reliability of assessments of written accuracy.

Notably, there were issues to resolve in WCR. As clauses with more than five errors were excluded in previous studies (Polio & Shea, 2014), the WCR rating method was limited by the vague rating scale. In response, I detailed the descriptors for the different errors in each level. In sum, this study addressed the following three research questions (RQs):

RQ 1-1: If raters assess written accuracy using accuracy measures, to what extent could inter-rater reliability be obtained?

RQ 1-2: If raters assess written accuracy using accuracy measures, to what extent could the score variances be explained by the factors?

RQ 1-3: If raters assess written accuracy using accuracy measures, what is the degree of reliability (G coefficient) obtained?

RQ1-1 was set for confirming the evaluation inference, and the generalization inference would be investigated from RQ1-2 to RQ1-3.

3.1.2 Method

3.1.2.1 Participants

The present study used data from the International Corpus Network of Asian Learners of English developed by Ishikawa (2013). The corpus includes written and spoken data produced by 2,800 university students in Asian countries/regions (China, Hong Kong, Indonesia, Japan, Korea, Pakistan, the Philippines, Singapore/Malaysia, Taiwan, and Thailand). In Ishikawa (2013), all participants reported their scores on English proficiency tests such as TOEFL, TOEIC, or IELTS, and they also took a vocabulary test (Nation & Beglar, 2007). Based on their proficiency and vocabulary test scores, participants were classified into four CEFR-linked (Common European Framework of Reference for Language) proficiency levels: *A2*, *B1_1* (lower level of the B1), *B1_2* (upper level of the B1), and *B2+* (a merger of the B2, C1, and C2 levels).

After obtaining permission, the present study used 100 out of 400 Japanese EFL university students' writing data in his corpus. Specifically, In this study, 50 students were selected in A2 and B1_1 groups and 50 in B1_2 and B2+ groups. The total number of participants was 100 students. I randomly selected participants from each CEFR-linked proficiency-level group: 25 students from A2, 25 students from B1_1, 32 students from B1_2, and 18 students from B2+.

The participants (44 females and 56 males, average age = 18.84 years) were majoring in various fields, including business, engineering, and economics. Their essays on two topics were analyzed. Note that the topics were (a) *It is important for college students to have a part-time job* (PTJ) and (b) *Smoking should be completely banned at all the restaurants in the country* (SMK). The average lengths of the PTJ and SMK essays were 223 words (*SD* = 24.1) and 219 words (*SD* = 26.1), respectively.

3.1.2.2 Data Collection

When collecting the writing data, many factors (e.g., L1 background) might influence writing performance. Written performance in ICNALE was obtained in the controlled experiments. Learners were required to write 200-300 words for each writing task for 20-40 minutes (Ishikawa, 2011, 2013). While they could not use a dictionary during the experiments, it was possible to use a spell checker.

As for levels of proficiency, collecting written data from students with different L2 proficiency was also controlled. To do this, the ICNALE project team investigated the writers' scores in the L2 vocabulary size test (VST), TOEIC, TOEFL, and IELTS as an objective measurement of their proficiency levels. The validity of these proficiency tests for discriminating proficiency has been validated by a number of studies (e.g., Beglar, 2010; Fleckenstein, Keller, Krüger, Tannenbaum, Koller, 2020; Schmidgall, 2017; Schoepp, 2018).

3.1.2.3 Traditional Written Accuracy Measures

Study 1 used seven written accuracy measures, which have been used in previous studies (Table 3.1). Although this study included *error-free clause* (EFC) and *error-free T-unit* (EFT) in the analysis, these features were not included in the written accuracy measures because they are used for calculating the written accuracy measures. However, the reliability of the EFC and EFT is important to judge the reliability of the written accuracy measures. Therefore, the present study examined their reliability. All measures were not normalized because the raw scores of accuracy measures have been used in the previous studies.

Table 3.1

Description of Traditional Written Accuracy Measures

Measure description	Code
Error-free clauses per total of all clause	EFCR
Error-free T-units per total of all T-unit	EFTR
Error-free sentences per total of all sentence	EFSR
Error-free T-units per total of all sentence	EFT/S
Error-free T-units per total of all word	EFT/W
Error-free clauses per total of all T-unit	EFC/T
Words in error-free clauses per total of all words in clauses	WEFC/WC

3.1.2.4 Raters for the WCR

Four raters—majoring in English education and applied linguistics—also participated in this study. All raters (R1–4) had an MA degree and three of them were Ph.D. students (R1, 2, and 3) when the present study was conducted. R1 was a returnee who attended the last three years of elementary school in the U.S. Upon graduating from high school, R1 attended university in Japan and studied at a university in the UK for one year. At the time of the study, R1 taught English in a Japanese university. R2 is from Ukraine and was an ESL learner. R2 completed an MA degree in Belgium and taught English to Japanese EFL learners for four years. R3 is a Japanese EFL learner with experience teaching English in two high schools and a university. R4 is a returnee from the Republic of Singapore, who previously studied in an international school, and is currently pursuing a Ph.D. in the UK.

3.1.2.5 Revising the WCR Rating Scale and Training Raters

Before training the raters, I developed a tentative rating scale that maps the scores and types of errors for each level; this was similar to the final version presented in the Appendix. Fifteen (five from each of the three levels except B2+) participants' data were selected and 30 essays, including two essays with different topics, were used to develop the tentative rating scale in Japanese and English. R2 used the English version through the rater training.

Subsequently, all raters participated in revising the tentative scale to construct a final, detailed version. This process was also treated as training for the raters; rater training procedures were based on Evans et al. (2014). Initially, the raters discussed the error code guidelines and rating scale, definition of different errors, accuracy, and WCR. Although Polio and Shea's (2014) guidelines for coding errors were used, new types of errors were added (e.g., conjunction error). In addition, the current study defined *error* as "A linguistic form or combination of forms which, in the same context and under similar conditions of production, would, in all likelihood, not be produced by the speakers' native speaker counterpart." (Lennon, 1991, p. 182).

After finishing the discussion, raters evaluated the written accuracy of six sample essays written by one participant (two essays per participant) from each level, from A2 to B1_2. Essays from the B2+ level were not included in this training exercise because essays at that level were used for analysis. I divided the sentences in the essays into clauses in advance. Using the tentative rating scale, the raters identified errors in each clause, appropriately tagged them, and evaluated the clauses. Subsequently, all instances of disagreements were discussed, and the researcher summarized the points for improving the rating scale.

After the evaluation, the rating scale was revised again. Then, all raters evaluated the written accuracy of four essays written by one participant from each of the A2 and B1_1 levels, which were assumed to contain a variety of errors. The appropriateness of the rating scale and procedures for tagging errors were discussed until all raters were in agreement. Following the pilot/rater training, the raters independently evaluated 200 essays (100 × two topics) utilizing the agreed-upon revised rating scale (see Appendix).

3.1.2.6 Scoring

Before the scoring, the sentences were divided into clauses in each essay by the researcher and checked by raters. The clauses were based on Evans et al. (2014), and comprised a subject and a predicate. In addition, predicates could be divided into a verb phrase and complement. Then, raters independently evaluated all essays using the final version of the rating scale.

Following the same procedure used during the rater training, raters were required to find errors in each clause and score the severity of errors according to the extent to which the error affects readers' comprehension. It should be noted that the same errors (e.g., word errors) could be categorized under different levels because the severity of these errors was often contextual. The raters had agreed that they would read the definitions of the rating scale upon finding such errors to categorize them. Moreover, when there were multiple errors (e.g., Level 1 and Level 3 errors) in the same clause, the clause was categorized according to its worst-level error, as suggested by Foster and Wigglesworth (2016).

The final WCR score was calculated as follows: WCR = the number of accurate clauses × 1.0 + the number of Lv.1 clauses × 0.8 + the number of Lv.2 clauses × 0.5 + the number of Lv.3 clauses × 0.1 / all clauses in the essay. As for the traditional written accuracy measure, the data obtained in the accuracy evaluation of the WCR were used.

3.1.2.7 Data Analysis

For RQ1, the study produced Cronbach's α for all accuracy measures in terms of inter-rater reliability. In addition, to reveal how individual inconsistencies affected inter-rater reliability, the study produced the adjusted α by eliminating each rater per task. The *psych* package in R (R Core Team, 2018) was used to calculate Cronbach's α and the adjusted α.

For the G theory, the data were analyzed using the *gtheory* package (Huebner & Lucht, 2019) in R (R Core Team, 2018). For RQ1-2, I used a G-study to compute the variance components of the main and interacting factors. The following seven variance components were established: (a) Person, (b) Rater, (c) Topic, (d) Person × Rater, (e) Person × Topic, (f) Rater × Topic, and (g) Person × Topic × Rater interactions (Table 3.2). It should be noted that Person × Topic × Rater interaction is regarded as a residual and covers other unsystematic or systematic sources of variations that were not included in the present study (Shavelson & Webb, 1991).

In addition, I used a D-study to calculate the G coefficient for RQ1-3. A G coefficient is similar to a CTT reliability coefficient and is used for a relative decision. That is, a G coefficient is interpreted as the consistency in the ranking order of individuals. The maximum possible value of the reliability coefficients is 1.0, and the generally acceptable value is 0.80 (Shavelson & Webb, 1991).

Table 3.2

Factors and Interpretations of Variance Components

Factors	Interpretation
Participant (*p*)	Universe-score variance, which shows how much accuracy differs at tasks
Topic (*t*)	The main effect of tasks, which shows whether tasks are more difficult than others
Rater (*r*)	The main effect for raters, which shows whether raters are more lenient than others in scores of accuracy
$p \times t$	Interaction between learners and tasks, which shows whether the relative standing of learners differs across tasks
$p \times r$	Interaction between learners and raters, which shows whether the relative standing of learners differs across raters
$t \times r$	Interaction between tasks and raters; which shows the inconsistency of raters' average ratings from one task to the next
$p \times t \times r$	Three-way interaction plus remaining unmeasured error

3.1.3 Results

3.1.3.1 Descriptive Statistics

Tables 3.3 and 3.4 shows the descriptive statistics of the WCR scores of the two essays written by 100 Japanese EFL learners assessed by four raters using the revised WCR rating scale. As shown, the mean scores of all the raters were similar within and between the topics. Similarly, the standard deviations of scores were small within and between the topics.

Table 3.3

Descriptive Statistics for All Accuracy Measures in PTJ

	PTJ							
	Rater A		Rater B		Rater C		Rater D	
Measures code	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
EFC	12.92	5.30	13.43	5.79	12.47	5.10	10.74	4.14
EFT	4.94	3.26	5.24	3.52	4.55	3.02	4.55	2.63
EFCR	0.45	0.16	0.47	0.18	0.44	0.16	0.38	0.13
EFTR	0.28	0.16	0.30	0.18	0.26	0.16	0.25	0.14
EFSR	0.26	0.16	0.28	0.19	0.24	0.16	0.24	0.14
EFT/S	0.30	0.18	0.32	0.20	0.28	0.17	0.27	0.15
EFT/W	0.02	0.01	0.02	0.02	0.02	0.01	0.02	0.01
EFC/T	0.77	0.33	0.80	0.37	0.74	0.32	0.64	0.27
WEFC/WC	2.05	1.44	2.29	1.66	1.81	1.31	1.93	1.23
WCR	0.86	0.05	0.89	0.04	0.84	0.06	0.81	0.08

Note. N = 100; EFC = Error-free clauses; EFT = Error-free T-units; EFCR = Error-free clauses per total of all clause; EFTR = Error-free T-units per total of all T-unit; EFSR = Error-free sentences per total of all sentence; EFT/S = Error-free T-units per total of all sentence; EFT/W = Error-free T-units per total of all word; EFC/T = Error-free clauses per total of all T-unit; WEFC/WC = Words in error-free clauses per total of all words in clause; WCR = weighted clause ratio; PTJ = It is important for college students to have a part-time job;

Table 3.4

Descriptive Statistics for All Accuracy Measures in SMK

	SMK							
	Rater A		Rater B		Rater C		Rater D	
Measures code	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
EFC	13.12	5.48	13.70	5.99	12.22	5.18	10.20	4.22
EFT	4.76	2.73	4.80	3.07	4.22	2.70	3.04	2.09
EFCR	0.42	0.16	0.44	0.17	0.39	0.15	0.33	0.12
EFTR	0.26	0.14	0.26	0.16	0.23	0.15	0.17	0.11
EFSR	0.23	0.13	0.24	0.16	0.20	0.14	0.15	0.11
EFT/S	0.30	0.17	0.30	0.20	0.26	0.18	0.19	0.14
EFT/W	0.02	0.01	0.02	0.01	0.02	0.01	0.01	0.01
EFC/T	0.74	0.34	0.77	0.37	0.68	0.32	0.57	0.27
WEFC/WC	1.74	1.14	1.91	1.46	1.48	1.15	1.07	0.89
WCR	0.84	0.07	0.87	0.06	0.81	0.08	0.81	0.07

Note. N = 100; EFC = Error-free clauses; EFT = Error-free T-units; EFCR = Error-free clauses per total of all clause; EFTR = Error-free T-units per total of all T-unit; EFSR = Error-free sentences per total of all sentence; EFT/S = Error-free T-units per total of all sentence; EFT/W = Error-free T-units per total of all word; EFC/T = Error-free clauses per total of all T-unit; WEFC/WC = Words in error-free clauses per total of all words in clauses; WCR = weighted clause ratio; SMK = Smoking should be completely banned at all the restaurants in the country.

3.1.3.2 Inter-Rater Reliability

Tables 3.5 and 3.6 show the results of Cronbach’s α and adjusted Cronbach’s α in all accuracy measures and features. The results indicate that the inter-rater reliability in all measures and features was over .80 in both writing tasks. Moreover, the results show the adjusted Cronbach’s α in the two writing tasks. According to the results, while some Cronbach’s α would be higher if some raters are excluded, the values were over .80.

Table 3.5

Inter-Rater Reliability and Adjusted Cronbach’s α in PTJ

	PTJ				
		Adjusted α			
Measures code	α	Rater A	Rater B	Rater C	Rater D
EFC	0.95	0.93	0.93	0.94	0.96
EFT	0.91	0.86	0.85	0.87	0.94
EFCR	0.94	0.91	0.91	0.92	0.95
EFTR	0.89	0.82	0.81	0.83	0.93
EFSR	0.89	0.82	0.82	0.84	0.93
EFT/S	0.89	0.83	0.82	0.84	0.93
EFT/W	0.89	0.83	0.82	0.84	0.93
EFC/T	0.96	0.94	0.94	0.95	0.97
WEFC/WC	0.88	0.81	0.90	0.81	0.93
WCR	0.91	0.92	0.92	0.91	0.93

Note. PTJ = It is important for college students to have a part-time job.

Table 3.6

Inter-Rater Reliability and Adjusted Cronbach's α in SMK

	SMK				
		Adjusted α			
Measure code	α	Rater A	Rater B	Rater C	Rater D
EFC	0.95	0.93	0.94	0.93	0.95
EFT	0.89	0.86	0.86	0.85	0.90
EFCR	0.93	0.90	0.91	0.90	0.93
EFTR	0.89	0.85	0.85	0.83	0.90
EFSR	0.89	0.85	0.85	0.83	0.90
EFT/S	0.91	0.88	0.87	0.86	0.92
EFT/W	0.89	0.85	0.86	0.84	0.89
EFC/T	0.96	0.94	0.95	0.95	0.96
WEFC/WC	0.90	0.88	0.88	0.86	0.92
WCR	0.93	0.92	0.92	0.91	0.94

Note. SMK = Smoking should be completely banned at all the restaurants in the country.

3.1.3.3 G-Study: Estimating Variance Components

Table 3.7 summarizes seven variance components and the percentages of their variability in EFC score as estimated by the G-study. Results are summarized in order of the percentage of variability.

Table 3.7

Estimated Variance Components and Percentage of Variabilities in EFC Score

Factor	Variance component	% Variability
Person (*p*)	8.649692	30.2
Topic (*t*)	0.000000	0.0
Rater (*r*)	1.777293	6.2
Person × Rater (*pr*)	0.715374	2.5
Person × Topic (*pt*)	13.53799	47.2
Rater × Topic (*rt*)	0.031437	0.1
Person × Topic × Rater (*ptr*)	3.944178	13.8

Note. EFC = error-free clause.

The largest variance was attributed to the Person × Topic (*pt*) factor (47.2%), which indicated that the topic affected learners' written accuracy scores in written performances. This was followed by the variance component of Person (*p*) (30.2%), indicating that the learners' accuracy ability accounted for a relatively large portion of the variance. The Person × Topic × Rater (*ptr*) interaction (13.8%) indicated that there were some unexplained influences beyond the factors treated in the present study.

The variance component of rater (*r*) (6.2%) indicated that the severity of ratings among the raters did not differ. The variance component of Person × Rater (*pr*) was almost negligible (2.5%), indicating that the raters ranked the learners in a similar manner. In addition, the variance of Rater × Topic (*rt*) was also small (0.1 %), indicating that the raters' severity or leniency was similar between the topics. Finally, the variance component of topic (*t*) was also nonexistent (0%), suggesting that the difficulty of the writing topics was similar.

Table 3.8 summarizes seven variance components and the percentages of their variability in EFT score as estimated by the G-study.

Table 3.8

Estimated Variance Components and Percentage of Variabilities in EFT Score

Factor	Variance component	% Variability
Person (*p*)	2.129478	23.9
Topic (*t*)	0.097178	1.1
Rater (*r*)	0.244165	2.7
Person × Rater (*pr*)	0.366664	4.1
Person × Topic (*pt*)	3.805410	42.6
Rater × Topic (*rt*)	0.133640	1.5
Person × Topic × Rater (*ptr*)	2.146344	24.1

Note. EFT = Error-free T-unit.

The largest variance was attributed to the Person × Topic (*pt*) factor (42.6%), which indicated that the topic affected learners' written accuracy scores in written performances. This was followed by the variance component of the Person × Topic × Rater (*ptr*) interaction (24.1%), which indicated that there were some unexplained influences beyond the factors treated in the present study Person (*p*) (23.9%). This indicated that the learners' accuracy ability accounted for a relatively large portion of the variance.

The variance component of Person × Rater (*pr*) was almost negligible (4.1%), indicating that the raters ranked the learners in a similar manner. The variance component of rater (*r*) (2.7%) indicated that the severity of ratings among the raters did not differ. In addition, the variance of Rater × Topic (*rt*) was also small (1.5 %), indicating that the raters' severity or leniency was similar between the topics. Finally, the variance component of topic (*t*) was also nonexistent (1.1%), suggesting that the difficulty of the writing topics was similar.

Table 3.9 summarizes seven variance components and the percentages of their variability in EFCR score as estimated by the G-study.

Table 3.9

Estimated Variance Components and Percentage of Variabilities in EFCR Score

Factor	Variance component	% Variability
Person (*p*)	0.009842	36.8
Topic (*t*)	0.000629	2.4
Rater (*r*)	0.001916	7.2
Person × Rater (*pr*)	0.000676	2.5
Person × Topic (*pt*)	0.009195	34.3
Rater × Topic (*rt*)	0.000004	0.0
Person × Topic × Rater (*ptr*)	0.004517	16.9

Note. EFCR = Error-free clause per total of all clause.

The largest variance was attributed to the Person (*p*) (36.8%), indicating that the learners' accuracy ability accounted for a relatively large portion of the variance. The Person ×

Topic (*pt*) factor (34.3%) indicated that the learners' written accuracy scores in written performances were affected by the topic. This was followed by the variance component of the Person × Topic × Rater (*ptr*) interaction (16.9%), which indicated that there were some unexplained influences beyond the factors treated in this study.

The variance component of rater (*r*) (7.2%) indicated that the severity of ratings among the raters did not differ. The variance component of Person × Rater (*pr*) was also almost negligible (2.5%), indicating that the raters ranked the learners in a similar manner. In addition, the variance component of topic (*t*) was also almost negligible (2.4%), suggesting that the difficulty of the writing topics was similar. Finally, the variance of Rater × Topic (*rt*) was also small (0 %), indicating that the raters' severity or leniency was similar between the topics.

Table 3.10 summarizes seven variance components and the percentages of their variability in EFTR score as estimated by the G-study. The largest variance was attributed to the Person × Topic (*pt*) factor (33.1%), which indicated that the learners' written accuracy scores in written performances were affected by the topic. Person (*p*) was (28.4%), indicating that the learners' accuracy ability accounted for a relatively large portion of the variance. This was followed by the variance component of the Person × Topic × Rater (*ptr*) interaction (27.5%), which indicated that there were some unexplained influences beyond the factors treated in this study.

Table 3.10

Estimated Variance Components and Percentage of Variabilities in EFTR Score

Factor	Variance component	% Variability
Person (*p*)	0.007019	28.4
Topic (*t*)	0.000716	2.9
Rater (*r*)	0.000677	2.7
Person × Rater (*pr*)	0.000916	3.7
Person × Topic (*pt*)	0.008168	33.1
Rater × Topic (*rt*)	0.000407	1.6
Person × Topic × Rater (*ptr*)	0.006775	27.5

Note. EFTR = Error-free T-units per T-unit.

The variance component of Person × Rater (*pr*) was also almost negligible (3.7%), indicating that the raters ranked the learners in a similar manner. The variance component of topic (*t*) was also almost negligible (2.9%), suggesting that the difficulty of the writing topics was similar. In addition, the variance component of rater (*r*) (2.7%) indicated that the severity of ratings among the raters did not differ. Finally, the variance of Rater × Topic (*rt*) was also small (1.6 %), indicating that the raters' severity or leniency was similar between the topics.

Table 3.11 summarizes seven variance components and the percentages of their variability in EFSR score as estimated by the G-study. The largest variance was attributed to the Person × Topic (*pt*) factor (38.3%), which indicated that the learners' written accuracy scores in written performances were affected by the topic. This was followed by the variance component of the Person × Topic × Rater (*ptr*) interaction (28.4%), which indicated that there were some unexplained influences beyond the factors treated in the present study. Person (*p*) was (23.7%), indicating that the learners' accuracy ability accounted for a relatively large portion of the variance.

Table 3.11

Estimated Variance Components and Percentage of Variabilities in EFSR Score

Factor	Variance component	% Variability
Person (*p*)	0.005725	23.7
Topic (*t*)	0.000902	3.7
Rater (*r*)	0.000472	2.0
Person × Rater (*pr*)	0.000606	2.5
Person × Topic (*pt*)	0.009249	38.3
Rater × Topic (*rt*)	0.000332	1.4
Person × Topic × Rater (*ptr*)	0.006867	28.4

Note. EFSR = Error-free sentences per sentence.

The variance component of topic (*t*) was also almost negligible (3.7%), suggesting that the difficulty of the writing topics was similar. The variance component of Person × Rater (*pr*) was also almost negligible (2.5%), indicating that the raters ranked the learners in a similar manner. In addition, the variance component of rater (*r*) (2.0%) indicated that the severity of ratings among the raters did not differ. Finally, the variance of Rater × Topic (*rt*) was also small (1.4 %), indicating that the raters' severity or leniency was similar between the topics.

Table 3.12 summarizes seven variance components and the percentages of their variability in EFT/S score as estimated by the G-study. The largest variance was attributed to the Person × Topic (*pt*) factor (33.7%), which indicated that the learners' written accuracy scores in written performances were affected by the topic. This was followed by the Person (*p*) (31.5%), indicating that the learners' accuracy ability accounted for a relatively large portion of the variance. The variance component of the Person × Topic × Rater (*ptr*) interaction (25.8%) indicated that there were some unexplained influences beyond the factors treated in this study.

Table 3.12

Estimated Variance Components and Percentage of Variabilities in EFT/S Score

Factor	Variance component	% Variability
Person (*p*)	0.010225	31.5
Topic (*t*)	0.000314	1.0
Rater (*r*)	0.000858	2.6
Person × Rater (*pr*)	0.001180	3.6
Person × Topic (*pt*)	0.010935	33.7
Rater × Topic (*rt*)	0.000563	1.7
Person × Topic × Rater (*ptr*)	0.008372	25.8

Note. EFT/S = Error-free T-units per sentence.

The variance component of Person × Rater (*pr*) was also almost negligible (3.6%), indicating that the raters ranked the learners in a similar manner. The variance component of rater (*r*) (2.6%) indicated that the severity of ratings among the raters did not differ. In addition, the variance of Rater × Topic (*rt*) was also small (1.7 %), indicating that the raters' severity or leniency was similar between the topics. Finally, the variance component of topic (*t*) was also almost negligible (1.0%), suggesting that the difficulty of the writing topics was similar.

Table 3.13 summarizes seven variance components and the percentages of their variability in EFT/W score as estimated by the G-study.

Table 3.13

Estimated Variance Components and Percentage of Variabilities in EFT/W Score

Factor	Variance component	% Variability
Person (*p*)	0.000044	22.7
Topic (*t*)	0.000001	0.7
Rater (*r*)	0.000004	2.4
Person × Rater (*pr*)	0.000008	4.2
Person × Topic (*pt*)	0.000080	41.0
Rater × Topic (*rt*)	0.000003	1.9
Person × Topic × Rater (*ptr*)	0.000053	27.0

Note. EFT_W = Error-free T-units per word.

The largest variance was attributed to the Person × Topic (*pt*) factor (41%), which indicated that the learners' written accuracy scores in written performances were affected by the topic. This was followed by the variance component of the Person × Topic × Rater (*ptr*) interaction (27%), which indicated that there were some unexplained influences beyond the factors treated in this study. The Person (*p*) was (22.7%), indicating that the learners' accuracy ability accounted for a relatively large portion of the variance.

The variance component of Person × Rater (*pr*) was also almost negligible (4.2%), indicating that the raters ranked the learners in a similar manner. The variance component of Rater (*r*) (2.4%) indicated that the severity of ratings among the raters did not differ. In addition, the variance of Rater × Topic (*rt*) was also small (1.9 %), indicating that the raters' severity or leniency was similar between the topics. Finally, the variance component of Topic (*t*) was also almost negligible (0.7%), suggesting that the difficulty of the writing topics was similar.

Table 3.14 summarizes seven variance components and the percentages of their variability in EFC_T score as estimated by the G-study.

Table 3.14

Estimated Variance Components and Percentage of Variabilities in EFC/T Score

Factor	Variance component	% Variability
Person (*p*)	0.042113	37.6
Topic (*t*)	0.000513	0.5
Rater (*r*)	0.005718	5.1
Person × Rater (*pr*)	0.002175	1.9
Person × Topic (*pt*)	0.048352	43.2
Rater × Topic (*rt*)	0.000033	0.0
Person × Topic × Rater (*ptr*)	0.01296	11.6

Note. EFC/T = Error-free clauses per T-unit.

The largest variance was attributed to the Person × Topic (*pt*) factor (43.2%), which indicated that the topic affected the learners' written accuracy scores in written performances. This was followed by the variance component of the Person (*p*) (37.6%), indicating that the learners' accuracy ability accounted for a relatively large portion of the variance. Person × Topic × Rater (*ptr*) interaction (11.6%) indicated that there were some unexplained influences beyond the factors treated in this study.

The variance component of rater (*r*) (5.1%) indicated that the severity of ratings among the raters did not differ. In addition, The variance component of Person × Rater (*pr*) was also almost negligible (1.9%), indicating that the raters ranked the learners in a similar manner. The variance component of topic (*t*) was also almost negligible (0.5%), suggesting that the difficulty of the writing topics was similar. Finally, the variance of Rater × Topic (*rt*) was also small (0 %), indicating that the raters' severity or leniency was similar between the topics.

Table 3.15 summarizes seven variance components and the percentages of their variability in the WEFC_WC score as estimated by the G-study. The largest variance was attributed to Person (*p*) (32.8%), indicating that the learners' accuracy ability accounted for a relatively large portion of the variance. This was followed by the variance component of the Person × Topic × Rater (*ptr*) interaction (27.8%), which indicated that there were some unexplained influences beyond the factors treated in this study. The Person × Topic (*pt*) factor was (27.4%), which indicated that the learners' written accuracy scores in written performances were affected by the topic.

Table 3.15

Estimated Variance Components and Percentage of Variabilities in WEFC/WC Score

Factor	Variance component	% Variability
Person (*p*)	0.616329	32.8
Topic (*t*)	0.096533	5.1
Rater (*r*)	0.052251	2.8
Person × Rater (*pr*)	0.046683	2.5
Person × Topic (*pt*)	0.515018	27.4
Rater × Topic (*rt*)	0.029448	1.6
Person × Topic × Rater (*ptr*)	0.522178	27.8

Note. WEFC/WC = Words in error-free clause per word in clauses.

The variance component of topic (*t*) was also almost negligible (5.1%), suggesting that the difficulty of the writing topics was similar. In addition, the variance component of rater (*r*) (2.8%) indicated that the severity of ratings among the raters did not differ. The variance component of Person × Rater (*pr*) was also almost negligible (2.5%), indicating that the raters ranked the learners in a similar manner. Finally, the variance of Rater × Topic (*rt*) was also small (1.6 %), indicating that the raters' severity or leniency was similar between the topics.

Table 3.16 summarizes seven variance components and the percentages of their variability in WCR as estimated by the G-study. The largest variance was attributed to the Person (*p*) factor (38.0%), indicating that the learners' accuracy ability accounted for a relatively large portion of the variance. This was followed by the variance component of Person × Topic (*pt*) (20.3%), which indicated that the topic affected the learners' written accuracy scores in written performances. The Person × Topic × Rater (*ptr*) interaction (19.7%) indicated that there were some unexplained influences beyond the factors treated in this study. The variance component of Rater (*r*) (18.3%) indicated that the severity of ratings among the raters differed to a certain degree.

Table 3.16

Estimated Variance Components and Percentage of Variabilities in WCR Score

Factor	Variance component	% Variability
Person (*p*)	0.00206	38.0
Topic (*t*)	0.00013	2.5
Rater (*r*)	0.00099	18.3
Person × Rater (*pr*)	0.00000	0.0
Person × Topic (*pt*)	0.00110	20.3
Rater × Topic (*rt*)	0.00006	1.2
Person × Topic × Rater (*ptr*)	0.00107	19.7

Note. WCR = weighted clause ratio.

The variance component of topic (*t*) was almost negligible (2.5%), suggesting that the difficulty of the writing topics was similar. The variance of Rater × Topic (*rt*) was also small (1.2%), indicating that the raters' severity or leniency was similar between the topics. Finally, the variance component of Person × Rater (*pr*) was also nonexistent (0.0%), indicating that the raters ranked the learners in a similar manner.

3.1.3.4 D-Study: Simulating the G Coefficients

Table 3.17 shows the G coefficient for all written accuracy measures and features obtained under the evaluations by the four raters and two tasks. According to Table 3.17, G coefficient values in all written accuracy clauses and features were .80 or above.

Table 3.17

G Coefficient Values for Each Written Accuracy Measure

Measure name	Code	G coefficient
Error-free clauses	EFC	0.93
Error-free T-units	EFT	0.84
Error-free clauses per total of all clause	EFCR	0.90
Error-free T-units per total of all T-unit	EFTR	0.83
Error-free sentences per total of all sentence	EFSR	0.80
Error-free T-units per total of all sentence	EFT/S	0.87
Error-free T-units per total of all word	EFT/W	0.83
Error-free clauses per total of all T-unit	EFC/T	0.95
Words in error-free clauses per total of all word in clauses	WEFC/WC	0.83
Weighted clause ratio	WCR	0.91

3.1.4 Discussion

RQ 1-1: If raters assess written accuracy using accuracy measures, to what extent could inter-rater reliability be obtained? (See p.78)

RQ1-1 is for testing whether the observations of performance on writing tasks are evaluated to provide observed scores. According to the results, Cronbach's α in all evaluations with accuracy measures was high, meaning that raters evaluated the written accuracy reliably. In addition, the adjusted Cronbach's α showed that the coefficients were similar even if the number of raters was eliminated. Most of the accuracy measures' scores were based on binary judgment (i.e., accuracy or inaccuracy), thus it would be easy to reach high reliability (e.g., Polio & Shea, 2014). These results would confirm the warrants, which are necessary to make the evaluation inference.

These results would correspond to the previous studies (e.g., Evans et al., 2014; Foster & Wigglesworth, 2014). Polio and Shea (2014) evaluated the written accuracy using three different methods (e.g., holistic evaluation, accuracy measures, and specific errors) and examined the reliability. Polio and Shea reported that the reliability of *weighted error-free T-units*, which would be similar to the WCR, was high (Pearson's *r* = .84). In addition, the reliability of the EFCR and EFTR were also high (*r* in the EFCR = .88, *r* in the EFTR = 88). Moreover, Evans et al. (2014) used three written accuracy measures (EFCR, EFTR, and WCR) and examined the reliability using the Rasch model. Evans et al.'s study indicated that the reliability of the WCR was high (Separation reliability = .00), meaning that the severity among raters was quite similar. These tendencies would be theoretically natural because some studies agreed that the construct of the accuracy would be the most coherent construct in the CAF framework (e.g, Pallotti, 2019). The evaluation consistency among raters would be easy for raters. Therefore, it should be reasonable that the reliability of all written accuracy measures in this study was also high.

RQ 1-2: If raters assess written accuracy using accuracy measures, to what extent could the score variances be explained by the factors? (See p.78)

In RQ 1-2, this study used the generalizability theory (G theory) and examined the factors which would influence the score variances. The discussion will be summarized from the seven factors.

The percentages of the Person factor indicate how much the variances derived from the persons' ability (i.e., accuracy). The analysis showed that the percentages of the Person factor varied among written accuracy measures. The results showed that the percentage in the WCR was the highest in all written accuracy measures (38.0%). This indicated that differences in accuracy levels accounted for the variance of WCR scores. This is fairly reasonable because the differences among levels of accuracy should reflect learners' intended ability (i.e., accuracy).

This result could also be reasonable because the WCR can include the error gravity and capture the small changes of the written accuracy in the writing performance (Evans et al., 2014). Moreover, compared to the impact of the Person factor (29.4%) in analytic ratings of language use in Schoonen (2005), the effect of the Person factor in this study was relatively high; however, these may be difficult to compare directly because the study used an analytic scale.

Interestingly, the percentages of the Person factor tended to decrease as the linguistic units (e.g., clauses, T-units, and sentences) were longer. For example, while the percentages of the Person factor in the EFCR and EFC/T were 36.8% and 37.6%, respectively, the percentages in the EFTR and EFSR were 28.4% and 22.7%, respectively. This could be explained by the fact that the percentage of the Person factor in the EFC was higher than the EFT. As Foster and Wigglesworth (2014) claimed that an accuracy measure using EFC could be better than an accuracy measure using EFT because the length of the unit of the EFC is shorter than the EFT. Hence, the EFC could reflect the persons' accuracy ability more than the EFT. However, the different tendencies were also shown (e.g., WEFC/WC), so it should be careful to discuss the generalizability.

Second, the percentage of the Topic factor was relatively low in all written accuracy measures and features (Max = 5.1, Min = 0), suggesting that the written accuracy scores would not be influenced by the topics. However, other studies related to written accuracy assessments have also indicated that the written accuracy scores (EFTR, EFCR, and WCR) might be affected by the topics used in the experiments (Evans et al., 2014). The possible reasons would be genres of writing topics. The two topics used in the present study are social, but the topics in Evans et al. (2014) were totally different (e.g., graduation, too much freedom, and farmers). Hence, the influence of topics would be large in Evans et al. (2014).

Additionally, although the Topic factor in all written accuracy measures and features was negligible, the percentages of the Person × Topic factor were relatively high (e.g., EFCR: 34.3% and EFT/W: 41%). These results indicate that learners' performance would be more or less different among the topics. As for the results, In'nami and Koizumi's (2016) suggestion that task-related interactions cause more score variance than rater-related interactions advises that this effect is possible in writing performance assessments. These results suggest that writing topics should be chosen carefully when multiple tasks are used—a claim also put forth by Evans et al. (2014).

Third, the percentage of the Rater factor in the written accuracy measures was quite low, meaning that the severity of the ratings among the raters was quite similar. In the accuracy evaluations using the traditional accuracy measures, raters determine whether a clause has an error. Therefore, the severity of the judgment would be similar if the raters have high-proficiency levels.

However, a different result was obtained in the percentage of the WCR. The analysis showed that the percentage of the Rater factor in the WCR was 18.3%, which was the highest among all measures. This suggests differences in the severity of the ratings between the raters even though all raters went through the same rater training. Although it might be difficult to compare the results directly with other studies using G theory, Brown (2011), in summarizing generalizability studies, showed that the impact of the rater factor was between 0.00% and 61.10%. Therefore, the impact of the factor obtained in this study was relatively low.

A possible reason for this occurrence could be that the raters had to go through several steps to assign scores. To measure the written accuracy of written essays using the revised WCR rating scale, the raters were required to find errors in clauses, assess their severity, and categorize them under appropriate levels. Small differences in judgments during the process might lead to deviations in ratings.

Another reason could be that some of the assessed contextual errors could be placed under any level depending on their severity. For example, word choice errors were categorized under all levels. Therefore, differences in interpreting the definitions and severities of errors according to the scale could have caused the differences in the severities of each rater's ratings. Additionally, Rao and Li's (2017) suggestion that the backgrounds and English proficiencies of raters affect judgments of error gravity encourages attentiveness to the ways in which other rater characteristics, such as teaching experience, rating experience, and academic background would impact inconsistency.

While the Rater factor affected the variances of WCR scores, the results showed that other factors related to the rater did not affect, not only WCR score variances, but also accuracy features and traditional accuracy measures. The interaction of Person × Rater did not affect the variance of written accuracy scores (Max = 4.2%, Min = 0%), indicating that raters awarded scores for each learner in a similar manner. In addition, the influence of the Rater × Topic interaction factor was also small (Max = 1.9%, Min = 0%), suggesting that the evaluations were similar between topics.

Although most variances in WCR scores can be explained by the six aforementioned factors, the percentage of Person × Topic × Rater variance was 19.7%. The traditional written accuracy measure also had relatively high percentages of Person × Topic × Rater variance (e.g., EFCR = 16.9). This indicates that factors not included in this study may also affect the variance of written accuracy scores, that is, other unsystematic or systematic sources of variations not included in the present study may affect score variance (Shavelson & Webb, 1991).

RQ 1-3: If raters assess written accuracy using accuracy measures, what is the degree of reliability (G coefficient) obtained? (See p.78)

The purpose of RQ1-3 was to explore the degree of reliability (G coefficient) of the written accuracy measures (e.g., EFCR) and features (e.g., EFC). In addition, the present study examined the G coefficient of the WCR score based on the detailed descriptors included in the revised WCR rating scale. The results of the D-study showed that the G coefficient in all written accuracy measures and features was over .80 if four raters evaluated two tasks. It should be natural because the Cronbach's α of all written accuracy measures and the feature was quite high in both tasks.

It should be noted that this study used the error-coding data coded in the written accuracy evaluation with the WCR when calculating the traditional written accuracy measures' scores (e.g., EFCR). The key difference between the WCR and the traditional written accuracy measures is whether raters consider the error gravity. For example, what raters should do is find errors in a clause when using the EFCR. By contrast, raters have to find errors in a clause and judge the influence when they use the WCR. Therefore, the present study used the error-coding data coded in the written accuracy evaluation with the WCR when producing the traditional written accuracy measures' scores.

As for the traditional written accuracy measures and features, the results showed that the G coefficient values in the traditional written accuracy measures (e.g., EFCR) and features were quite high. Although some studies about the writing evaluation claimed that rater training would be doubtful to obtain high reliability (e.g., Weigle, 2002), detecting errors in a linguistic unit (e.g., clauses) would be reliable if raters' English proficiency is high (e.g., C1 level). In recent years, computer-assisted systems, such as Grammarly, have been used to help detect errors (e.g., Barrot & Adgeppa, 2022). It would become an alternative way to find linguistic errors.

As for the WCR, the D-study estimated the value of the G coefficient as 0.91 under four raters and two writing tasks with different topics, that is, the G coefficient obtained in this study condition was considerably high. It would be the encouraging results since judging the error gravity would be influenced by the raters' backgrounds, such as teaching experience (Rao & Li, 2017). One possible reason that enables the high reliability would be the detailed descriptors of the WCR. The present study pointed out the vagueness of the WCR rating scale and revised it. As descriptors should be fundamental in rating writing performance, the descriptors of the WCR rating scale would account for this high value because the raters were able to reliably evaluate written accuracy based on the revised scale.

In addition to the revised descriptor, the four raters in the current study had adequate time for rater training. They participated in all the processes from revising the descriptors to completing the final version, that would enable them to consistently rate the written accuracy with WCR. While the computer-assisted systems would be useful to find errors, it would still be difficult to judge which errors would be serious. Therefore, the detailed descriptor and rater training should be important for raters to judge the error gravity.

In addition, unlike Polio and Shea’s (2014) study, which excluded incomprehensible T-units with more than five errors, the present study did not exclude any clauses. Regardless, this study successfully established the high reliability of the WCR rating scale with detailed descriptors. Consequently, it can be concluded that the WCR rating scale revised in this study can be utilized to evaluate the written accuracy of Japanese EFL learners’ written essays.

3.1.5. Conclusion

Study 1 examined the reliability of the accuracy measures using Cronbach’s α coefficient and G-study. The findings in this study can be summarized as follows. First, the inter-rater reliability of the accuracy assessment by WCR was found to be very high. This means that it is possible to conduct a highly reliable evaluation. This will be obtained when the descriptors of error types in the WCR rating scale were described in detail. In addition, inter-rater reliability was also shown to be high for the other accuracy measures. Based on the results, scoring inference has been satisfied.

The second conclusion is that the accuracy assessment by WCR is highly generalizable by G-theory analysis. The G coefficient in the measurement design, in which four raters rated two writing tasks with different topics, was quite high (0.91). High generalization coefficients were obtained for other measures as well, although there were variations. From these results, the present study can say that the generalized inference was also satisfied.

Chapter 4

Study 2: Investigating the Factors Which Accuracy Measures Would Reflect

4.1 Explanation Inference

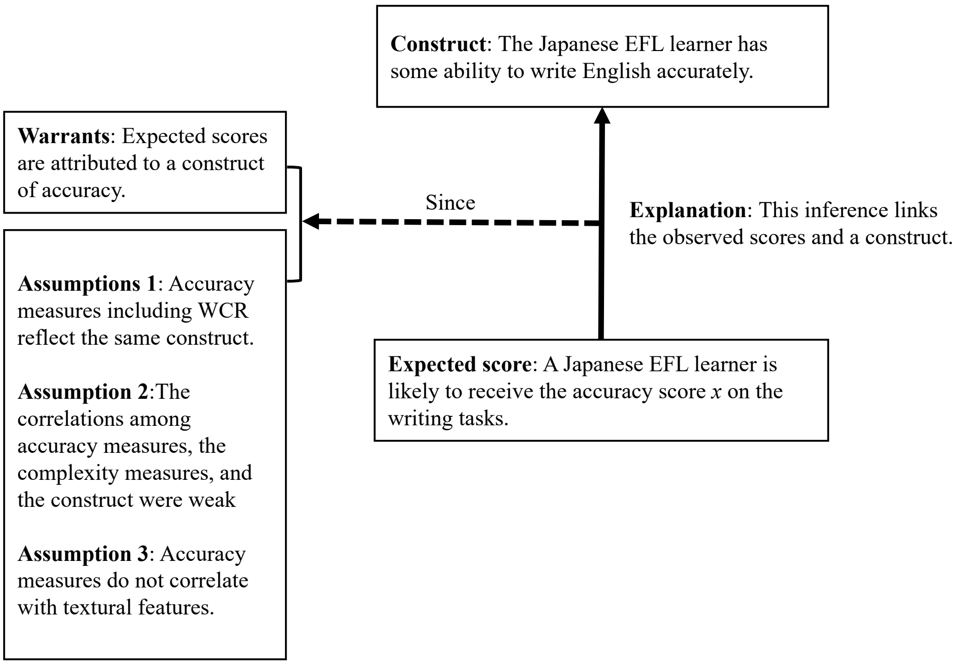
4.1.1 Purposes and Research Questions

Study 1 showed that the Cronbach's α and G coefficient of all accuracy measures were high (α > .80 and G > .80), hence the evaluation and generalization inferences were confirmed positively. In Study 2, the explanation inference was examined in order to link the observed scores and a construct.

Figure 4.1 shows the specific part of the argument-based approach which Study 2 addressed.

Figure 4.1

Summary of Explanation Inference



Construct: The Japanese EFL learner has some ability to write English accurately.

Warrants: Expected scores are attributed to a construct of accuracy.

Since

Explanation: This inference links the observed scores and a construct.

Assumptions 1: Accuracy measures including WCR reflect the same construct.

Assumption 2:The correlations among accuracy measures, the complexity measures, and the construct were weak

Assumption 3: Accuracy measures do not correlate with textural features.

Expected score: A Japanese EFL learner is likely to receive the accuracy score *x* on the writing tasks.

In order to determine whether a participant writes an essay accurately, it is necessary to build the explanation inference, which plays a role to connect the expected scores and a construct. In addition, the essential warrant for the explanation inference is that expected scores are attributed to a construct of accuracy. Therefore, the assumption should be that all accuracy measures would reflect the same construct (i.e., accuracy).

Moreover, this investigation would be important to build the explanation inference and answer a critique in the previous studies (e.g., Pallotti, 2009). The *error gravity*, which is the degree to which they influence a reader's comprehension, is used for classifying the clauses in essays when WCR is calculated (e.g., Foster & Wigglesworth, 2016). While WCR has been regarded as the accuracy measures in the present studies (e.g., Evans et al., 2014; Foster & Wigglesworth, 2016), some studies claimed that accuracy measures using error gravity would not measure accuracy but other constructs, such as comprehensibility (Pallotti, 2009). Moreover, Fox (2019) claimed that WCR might measure the dimensions of complexity. If WCR would not be an accuracy measure, it would be difficult to use WCR as the accuracy measure and interpret how written accuracy would change over time. In sum, this study used EFA to reveal whether WCR and other accuracy measures would reflect the same construct.

In addition, it would be necessary to investigate the relationships between the accuracy measures and linguistic units to calculate the scores (e.g., the number of clauses, T-units, and words) because the interpretations of scores might be doubtful. Even though an accuracy measure reflects the construct of accuracy, the increase of the score would not necessarily mean the increase of accuracy ability if an accuracy measure would be correlated with the number of words. In order to investigate the relationships between them, this study applied a correlation analysis. In sum, this study addressed the following three research questions (RQs):

RQ2-1: To what extent does the WCR reflect the factor that traditional accuracy measures do?

RQ2-2: To what extent do extracted factors and measures correlate with each other?

RQ2-3: To what extent do the accuracy measures correlate with textual features in essays?

4.1.2 Method

4.1.2.1 Participants

The present study used data from Study 1, which is extracted from the International Corpus Network of Asian Learners of English developed by Ishikawa (2013). The 100 participants (44 females and 56 males, average age = 18.84 years) were majoring in various fields, including business, engineering, and economics. The participants' essays on two topics were analyzed. Note that the essay topics were (a) *It is important for college students to have a part-time job* (PTJ) and (b) *Smoking should be completely banned at all the restaurants in the country* (SMK). The average lengths of the PTJ and SMK essays were 223 words (*SD* = 24.1) and 219 words (*SD* = 26.1), respectively.

4.1.2.2 Written Accuracy Measures

Study 2 used eight accuracy measures (i.e., WCR and EFCR) based on the results in Study 1. Study 1 showed that these accuracy measures would evaluate written accuracy with high reliability. The WCR rating scale consists of four categories, definitions, and scores (Table 4.1). After all sentences in an essay are divided into clauses, each clause is categorized by its gravity of error according to the definitions. It should be noted that zero should not be awarded because even Level 3 clauses are linguistically accurate to a certain degree (Foster & Wigglesworth, 2016). However, traditional accuracy measures such as

EFCR do not take into account the gravity of errors. Table 4.2 is the description of all accuracy measures used in Study 2.

Table 4.1

Rating Scale of a Weighted Clause Ratio

Category	Definition	Score
No error	The clause is accurately constructed.	1.0
Level 1	The clause has only minor errors (e.g., morphosyntax) that do not compromise meaning.	0.8
Level 2	The clause contains serious errors (e.g., verb tense, word choice, or word order), but the meaning is recoverable, though not always obvious.	0.5
Level 3	The clause has very serious errors that make the intended meaning far from obvious and only partly recoverable.	0.1

Table 4.2

Description of Written Accuracy Measures

Measures	Code
Error-free clauses per total of all clause	EFCR
Error-free T-units per total of all T-unit	EFTR
Error-free sentences per total of all sentence	EFSR
Error-free T-units per total of all sentence	EFT/S
Error-free T-units per total of all word	EFT/W
Error-free clauses per total of all T-unit	EFC/T
Words in error-free clauses per total of all word in clauses	WEFC/WC
Weighted clause ratio	WCR

4.1.2.3 Complexity Measures

The present study used three complexity measures based on the previous studies (Barrot & Agdeppa, 2021; Kato, 2019); Dependent clauses per total of all clause (DC/C);

Clauses per total of all sentence (C/S); Verb phrases per total of all T-unit (VP/T). Table 4.3 shows the descriptions of the complexity measures.

Table 4.3

Description of Written Complexity Measures

Measures	Code	Descriptions
Dependent clauses per total of all clause	DC/C	Number of dependent clauses per total of all clause
Clauses per total of all sentence	C/S	Number of clauses per total of all sentence
Verb phrases per total of all T-unit	VP/T	Number of verb phrases per total of all T-unit

The reason why the present study used those measures is that they would be assumed to be the same construct (i.e., Clausal Complexity). Although many complexity measures have been developed in previous studies (e.g., Bulté & Housen, 2012; Lu, 2011), it was not investigated which factors each complexity measure reflected. According to Kato's (2019) study, which investigated the factor structures of complexity using EFA and confirmatory factor analysis, the three measures reflected the same construct of Clausal Complexity (CFI = .997, RMSEA = .044, and SRMR = .032). In addition, some studies claimed that WCR might measure complexity as well (e.g., Fox, 2019). Although WCR was proposed to measure written accuracy (Foster & Wigglesworth, 2016), it would be necessary to investigate whether WCR would measure written accuracy and complexity.

4.1.2.4 Textual Features

The present study used three textual features (i.e., number of words, clauses, T-units, and sentences), which are necessary to calculate written accuracy measures. This study chose the features because they are used for calculating the written accuracy measures (e.g., EFCR).

4.1.2.5 Scoring

Weighted clause ratio

Study 2 used WCR as an accuracy measure which was obtained in Study 1. Before the scoring, the sentences were divided into clauses in each essay by the researcher and checked by raters. Then, raters independently evaluated all essays using the final version of the rating scale. Following the same procedure used during the rater training, raters were required to find errors in each clause and score the severity of errors according to the extent to which the error affects readers' comprehension.

It should be noted that the same errors (e.g., word errors) could be categorized under different levels because the severity of these errors was often contextual. The raters had agreed that they would read the definitions of the rating scale upon finding such errors to categorize them. Moreover, when there were multiple errors (e.g., Level 1 and Level 3 errors) in the same clause, the clause was categorized according to its worst-level error, as suggested by Foster and Wigglesworth (2016). The final WCR score was calculated as follows: WCR = (the number of accurate clauses × 1.0 + the number of Lv.1 clauses × 0.8 + the number of Lv.2 clauses × 0.5 + the number of Lv.3 clauses × 0.1) / all clauses in the essay.

Traditional written accuracy measures

The traditional written accuracy measures (e.g., EFCR) were calculated based on the WCR data. The accuracy measures with the linguistic units (e.g., clauses, T-units, and Sentences) were calculated by dividing the error-free linguistic units (e.g., Error-free clauses) by the linguistic units (e.g., all clauses). Therefore, after calculating the score of WCR, these traditional accuracy measures were produced.

Complexity measures

The complexity measures were calculated by the automated tool L2 Syntactic Complexity Analyzer (L2SCA; Lu, 2010). The reliability and validity were investigated in previous studies (Lu, 2010, Polio & Yoon, 2018). According to the study of Lu (2010), the reliability of the L2SCA ranged from 0.834 to 1.00. Moreover, recent studies (Polio & Yoon, 2018) confirmed the reliability and validity of the L2SCA.

4.1.2.6 Data Analysis

This study used EFA for RQ2-1 and RQ2-2. For all analyses for the two RQs, the *psych* packages (Revelle, 2018) in R (R Core Team, 2018) were used. Before conducting EFA for RQ2-1 and RQ2-2, this study checked the correlation above .90 among accuracy measures to avoid problems of multicollinearity. In addition, this study used the Shapiro-Wilk normality test to check the normality of complexity and accuracy measures.

Then Kaiser-Meyer-Olkin's measure of sampling adequacy (KMO) and Bartlett's test of sphericity were checked. KMO ranges from zero to one. The more the score approaches one, the better the reliable extractions of factors and the validity of the sampling (Field, 2009). The interpretations of the KMO score were summarized in Table 4.4.

In addition, Bartlett’s test of sphericity suggests whether the observed variables correlate. The null hypothesis is that there is no correlation among the observed variables. If the hypothesis is rejected, it means that there are correlations among them and it would be possible to conduct EFA.

Table 4.4

Summary of Kaiser-Meyer-Olkin Measure

KMO Score	Interpretation
.90 – .99	Marvelous
.80 – .89	Meritorious
.70 – .79	Middling
.60 – .69	Mediocre
.50 – .59	Miserable
Below .50	Unacceptable

After checking the KMO and Bartlett’s test of sphericity, the present study used the eigen value, scree plot, parallel analysis, and very simple structure (VSS) criterion and examined the number of factors. When the eigen value was used, the number of factors with an eigen value greater than 1 were adopted. In addition, when the scree plot was used for examining the number of factors, the factors up to the point where the eigen value drop is large were adopted.

Moreover, in order to decide the number of factors, this study used the parallel analysis and VSS criterion, which have been recommended in recent studies. The parallel analysis used a random number of data of the same size as the actual data and compared the eigen values of the correlation matrix for random data and eigen values of the correlation matrix for actual data. However, VSS criteria is also one of the tools for deciding the number of factors. This criterion can suggest the simple structure of factors in EFA.

As for the rotation, this study used the *promax* rotation because it was assumed that all of the measures would correlate with each other. For the estimation, the *generalized weighted least squares* (GLW) method was used to conduct EFA because some of the measures would not follow the normal distribution. The GLW estimation can be used when the observed variables do not follow the normal distribution (Toyoda, 2012).

As for RQ2-3, this study used a correlation analysis to investigate the relationships between the written accuracy measures and textual features. It would be assumed that the accuracy measures and textual features are correlated nonlinearly because the accuracy scores may or may not decrease as the word count increases. Therefore, this study used the maximal information coefficient (MIC, Reshef et al., 2011) with *minerva* packages.

MIC is a statistical method to investigate the relationships between two variables that are nonlinear. The MIC score will be close to one when the relationships between the two variables are strong. On the other hand, the MIC score will be close to zero when the relationships between the two variables are weak. In addition, the MIC score will be a positive score even if the relationships are downward to the right. Therefore, it is necessary to use the MIC score and shape of the relationships in the plot when the meaning of the MIC score is interpreted. Moreover, the present study used the HSIC test (Hilbert-Schmidt Independence Criteria; Gretton et al., 2005) to investigate the significance of the correlation.

4.1.3 Results

4.1.3.1 Descriptive Statistics

Table 4.5 is the descriptive statistics of all measures used in Study 2.

Table 4.5

Descriptive Statistics for Accuracy and Complexity Measures

Measures	Code	*M*	*SD*	*Min*	*Max*
Clauses per total of all sentence	C_S	1.8	0.35	0.83	2.95
Dependent clauses per total of all clause	DC_C	0.37	0.08	0.18	0.6
Verb phrases per total of all T-unit	VP_T	2.15	0.37	1.25	3.12
Weighted clause ratio	WCR	0.84	0.05	0.62	0.95
Error-free clauses per total of all clause	EFCR	0.41	0.12	0.11	0.79
Error-free T-units per total of all T-unit	EFTR	0.25	0.11	0.08	0.64
Words in error-free clauses per total of all word in clauses	WEFC/WC	1.78	0.98	0.4	5.99
Error-free clauses per total of all T-unit	EFC/T	0.71	0.26	0.15	1.48
Error-free T-units per total of all word	EFT/W	0.02	0.01	0.01	0.06
Error-free T-units per total of all sentence	EFT/S	0.28	0.13	0.1	0.84
Error-free sentences per total of all sentence	EFSR	0.23	0.11	0.04	0.6

As for the written accuracy measures, the WCR score was higher than the written accuracy measures produced by the linguistic units (e.g., EFCR, EFTR) because based on the rating scale of WCR, all clauses have scores more than zero. Therefore, the score would be higher than other written accuracy measures using linguistic units such as EFCR.

The mean score of WCR was 0.84, meaning that essays in the present study would be easy to understand.

4.1.3.2 Results in EFA

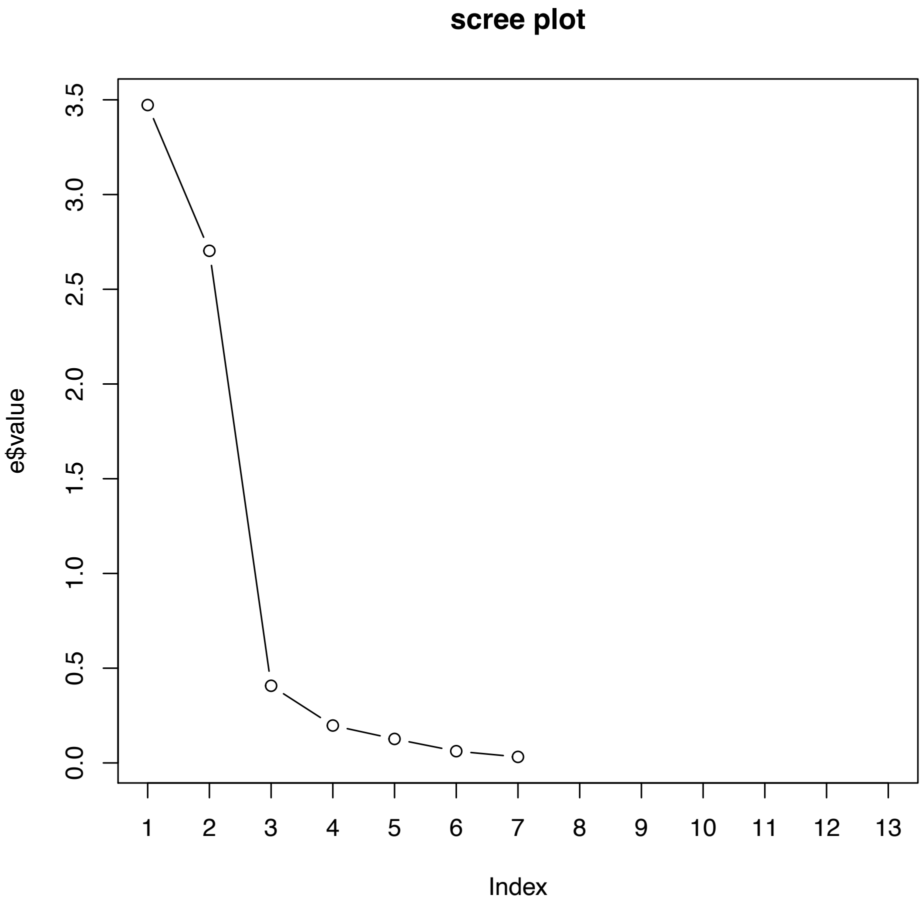
Table 4.5 is the descriptive statistics of all measures. Before examining the number of factors, the present study checked the multicollinearity based on the correction analysis. According to the results of the correlation analysis, some of the measures were highly correlated. Therefore, the present study excluded five accuracy measures (WEFC/WC, EFC/T, EFT/S, and EFS/S). The second correlation analysis showed that correlations over .90 were not observed, suggesting that multicollinearity did not occur.

After checking the multicollinearity, the KMO and Bartlett's test of sphericity were examined. The KMO test score was .72 (Middling, Kaiser, 1974), meaning that the sample in the present study was regarded to be valid as a sample. Moreover, Bartlett's test of sphericity (χ^2 = 822.42, *df* = 21, *p* < .001) was statistically significant, suggesting that the correlations among observed variables rejected the null hypothesis.

Then, the number of factors to be extracted were examined referring to the eigen value, scree plot, parallel analysis, and VSS criterion. According to the eigen value, it suggested two factors be adequate. In addition, the scree plot showed that the point where the eigen value drop is large was before the three-factor and indicated the two factors should be adequate (Figure 4.2).

Figure 4.2

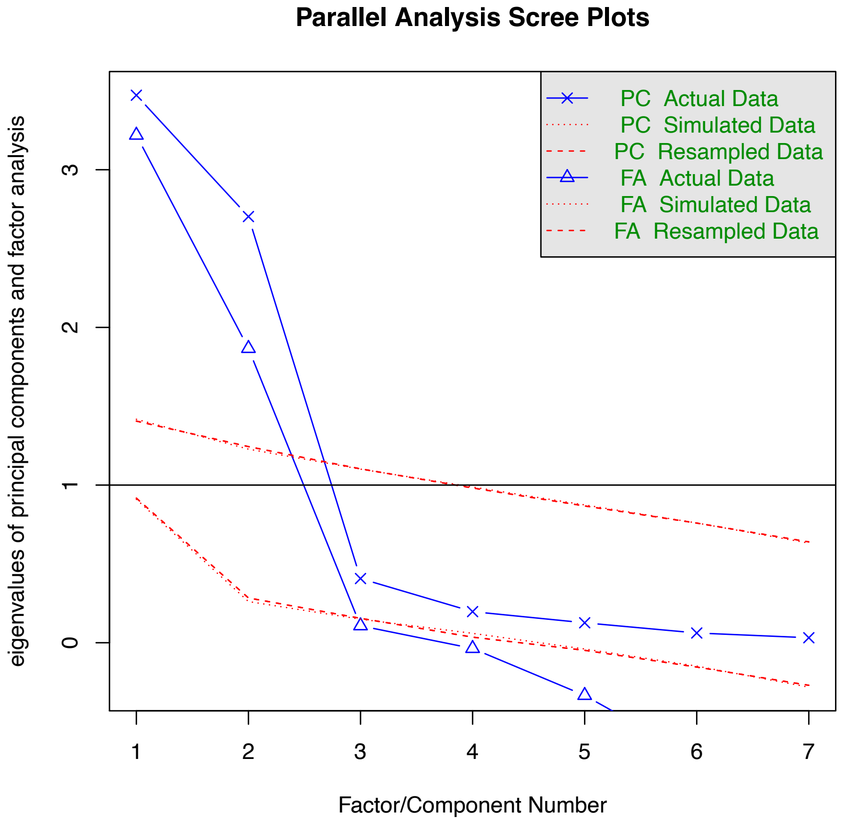
Scree Plot



Moreover, the parallel analysis showed that the dotted line of FA Simulated Data at the bottom of the figure crossed with the left straight line of FA Actual Data. This result suggested that there are two factors. Although the VSS criterion suggested the three-factor structure, the present study accepted the two-factor model based on the whole results.

Figure 4.3

Parallel Analysis



After confirming the assumptions and deciding the number of factors, EFA was conducted to answer RQ2-1 and RQ2-2. Table 4.6 shows the results of EFA and indicated that the first factor seemed to capture the construct of accuracy. This factor included four accuracy measures that operationalized written accuracy. The second factor seemed to be the factor of Clausal Complexity since the three measures operationalizing Clausal Complexity.

The results showed that the proportion of variance explained, which shows the percentage contribution of each factor to the total observed variable. The proportions of variance explained were 45% and 38% in Factors 1 and 2. The results also showed the cumulative proportion of variance explained, which is the factor contribution ratio added in order from the first factor. The cumulative proportion of variance explained was 84% in all variances could be explained by the two factors. Moreover, the result indicated that the correlation between Factors 1 and 2 was .08.

Table 4.6

Summary of Factor Loading With Accuracy and Complexity Measures in EFA

Measures	Code	Factor loading		Communality
		1 (α = .82)	2 (α = .81)	
Error-free T-units per total of all T-unit	EFTR	**.96**	-.02	.92
Error-free clauses per total of all clause	EFCR	**.95**	.21	.97
Error-free T-units per word	EFT/W	**.85**	-.35	.80
Weighted clause ratio	WCR	**.81**	.16	.71
Dependent clauses per total of all clause	DC/C	-.04	**.91**	.83
Verb phrases per total of all T-unit	VP/T	.06	**.91**	.84
Clauses per total of all sentence	C/S	.03	**.90**	.82

	Factor correlations	
Factor 1	-	
Factor 2	.08	-

Note. *N* = 100; Factor loading ≧ .40 are in boldface.

4.1.3.3 Results in the Correlation Analysis

Table 4.7 is the descriptive statistics of words in essays produced by the participants in A2 to B2+ levels. Although the number of essays was different, the number of words among the levels was similar.

Table 4.7

Descriptive Statistics of Words in Essays

Levels	No. of essays	Number of words			
		M	*SD*	*Min*	*Max*
A2	25	213.2	11.7	196	252
B1_1	25	212.8	21.8	182	303
B1_2	32	227.7	22.6	185	291
B2+	18	234.1	29.6	196	294

Table 4.8 shows the number of clauses in essays produced by the participants in A2 to B2+ levels. The number of clauses was also similar among all proficiency levels. Moreover, the standard deviations in each proficiency level among them were also small, hence the writing performances were quite similar in the perspectives in clauses.

Table 4.8

Descriptive Statistics of Clauses in Essays

Levels	Number of clauses			
	M	*SD*	*Min*	*Max*
A2	30.3	3.2	25	39
B1_1	30.1	4.7	24	47
B1_2	28.5	3.8	23	38
B2+	30.5	3.3	25	37

Table 4.9 shows the number of T-units in essays produced by the participants in A2 to B2+ levels. The number of T-units was similar among all proficiency levels. Moreover, the standard deviations in each proficiency level and among them were also small, hence the writing performances were also quite similar in the perspectives in T-units.

Table 4.9

Descriptive Statistics of T-Units in Essays

Levels	Number of T-units			
	M	*SD*	*Min*	*Max*
A2	18.4	2.8	13	23
B1_1	18.1	3.5	12	25
B1_2	17.7	2.8	12	23.5
B2+	17.4	3.4	13	27

Table 4.10 shows the number of sentences in essays produced by the participants in A2 to B2+ levels. The number of sentences was similar among all proficiency levels.

The standard deviations in each proficiency level and among them were also small.

Table 4.10

Descriptive Statistics of Sentences in Essays

Levels	Number of sentences			
	M	*SD*	*Min*	*Max*
A2	17.0	3.5	9.5	22.5
B1_1	16.8	3.8	9.5	24
B1_2	16.1	3.0	11	22
B2+	15.9	3.9	11	26.5

Figure 4.4 is the plot describing the relationships among the measures and textual factors. It suggested that the measures and textual factors would be correlated nonlinearly, hence the present study used MIC, which can be used for nonlinear correlations.

Figure 4.4

Relationships Among Measures

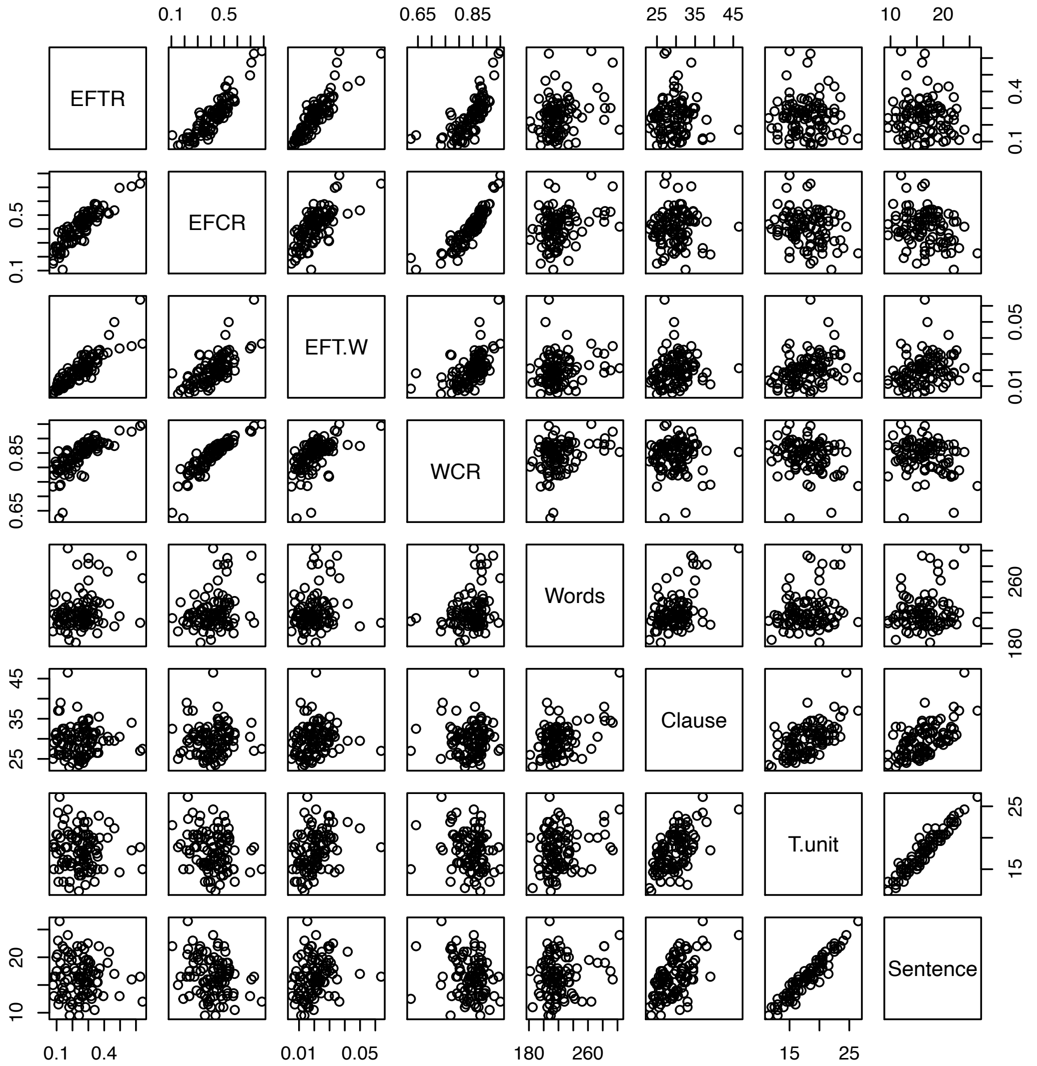


Table 4.11 shows the correlations among the measures and textual factors (i.e., word, clause, T-unit, and sentence). It should be noted that the stronger the relationship between the two variables, the closer the MIC is to one. The weaker the relationship between the two variables, the closer the MIC is to zero. In addition, the MIC score will be a positive score even if the relationships are downward to the right. Therefore, it is

necessary to use not only the MIC score but also the shape of the relationships in the plot (Figure 4.4) when the meaning of the MIC score is interpreted.

These results describe that the relationships between the written accuracy measures and textual factors were mostly low (MIC = .18–.26).

Table 4.11

Correlation Values

Measures	Word (MIC–R^2)	Clause (MIC–R^2)	T-unit (MIC–R^2)	Sentence (MIC–R^2)
WCR	.24 (.15)	.22 (.22)	.21 (.16)	.21 (.16)
EFCR	.22 (.12)	.20 (.20)	.24 (.18)	.26 (.20)
EFTR	.20 (.12)	.19 (.19)	.18 (.18)	.19 (.18)
EFT/W	.24 (.21)	.26* (.24)	.21** (.16)	.23** (.16)

Note. WCR = weighted clause ratio; EFCR = Error-free clauses per all clause; EFTR = Error-free T-units per all T-unit; EFT/W = Error-free T-units per word. *p** < .05, *p*** < .01, *p**** < .001 for the HSIC test.

4.1.3 Discussion

RQ2-1: To what extent does the WCR reflect the factor that traditional accuracy measures do? (See p.111)

The purpose of RQ2-1 is to investigate whether WCR would reflect the factor that traditional accuracy measures do. This RQ would be shed light on developing the validation in the present study and answering the questions that Pallotti (2009) claimed (i.e., the accuracy measures calculating by error gravity might not reflect the accuracy construct). Previous studies examined only the correlation between WCR and a few accuracy measures (e.g., Polio & Shea, 2014) and did not show the construct structures by factor analysis. Based on these limitations, the present study sought to uncover the factor which WCR would reflect.

The result by the EFA showed that WCR was grouped into the accuracy factor including other traditional measures (Table 4.6). According to the results, WCR would reflect the construct of written accuracy in second language writing. In addition, the Cronbach's α in the construct of accuracy was also high (α = .82), meaning that the accuracy measures as the observable variables would consistently measure the written accuracy. These results would confirm that WCR reflects the written accuracy construct. Theoretically, the construct of the written accuracy is the simplest and most internally coherent construct in the framework of CAF (Pallotti, 2009).

Given the definitions of the accuracy and nature of the error, the way of considering the error gravity and the fact that the WCR would reflect the written accuracy could be reasonable. In this study, accuracy is defined as "The ability to produce the error-free and target-like languages" (Housen et al., 2012, p.2), and error is also defined as "A linguistic form or combination of forms which, in the same context and under similar conditions of production, would, in all likelihood, not be produced by the speakers' native speaker counterpart" (Lennon, 1991, p.182). The traditional written accuracy measures have primarily focused on errors of a linguistic form or combination of forms (e.g., subject-verb agreements). However, it has also been thought that there are two types of errors (i.e., global and local errors) as the subclass of errors (Burt & Kiparsky, 1974). Therefore, it might be natural that the written accuracy measures could consider the types of errors when measuring the accuracy in writing. Housen et al. (2012) assumed that accuracy relates to interlanguage and L2 knowledge. This kind of knowledge (e.g., grammatical knowledge) would not only be linguistically correct, but also meaningful (Purpura, 2004; Spinner, 2016). Based on this definition, the construct of accuracy can capture the extent to which meaning is conveyed to the reader because the scope of accuracy includes not only linguistic features, but also meaning. Therefore, accuracy measures including WCR would be grouped into the same factor even though the WCR were calculated by the error gravity.

RQ2-2: To what extent do extracted factors and measures correlate with each other? (See p.111)

In RQ 2-2, the present study investigated the relationships between written accuracy and complexity measures to reveal whether the construct of the written accuracy would highly correlate with the construct of the complexity. There have been many discussions about the high correlations between the WCR and the complexity measures (e.g., Fox, 2019) although the WCR was proposed to measure the written accuracy. The present study can provide suggestions for the discussions. If the WCR measures reflect not only the written accuracy but also the complexity, the correlations between them might be high.

However, the results of the EFA showed that the relationship between the constructs of the written accuracy and the complexity was low (r = .08), meaning that WCR would not measure the complexity in the second language. These results would correspond to the one in the present studies (e.g., Barrot & Agdeppa, 2021) although the results of the study were obtained by the correlation analysis. Barrot and Agdeppa investigated the relationships among CAF measures and showed that the measures of the WCR and the complexity (e.g., C/S, DC/C) did not correlate (*r* = .02–.15). Similar tendencies between the written accuracy and complexity measures were shown in the Barrot and Agdeppas' study and other previous studies (e.g., Barrot & Gabinete, 2019). Based on the results of the correlation analysis and the EFA, it would be possible to claim that WCR would reflect the written accuracy construct.

RQ2-3: To what extent do the accuracy measures correlate with textual features in essays? (See p.111)

In RQ2-3, the present study showed the relationships between measures and textual features (i.e., words, clauses, T-units, and sentences) to investigate whether the written accuracy measures would highly correlate with the textual features. If the written accuracy measures significantly correlated with them, it would be doubtful that the changes of a score of an accuracy measure mean the development of the ability of accuracy.

The results showed that most of the written accuracy weakly correlated with the textual features (MIC = 18–.26). It should be noted that EFT/W correlated significantly with T-units (MIC = .21) and sentences (MIC = .23) in spite of the similar MIC score with other written accuracy measures.

One of the reasons for this result is the strict word limit: in the creation of the ICNALE corpus, participants were asked to write English essays of no more than approximately 200 to 300 words. The results of the descriptive statistics of the word count show that the maximum standard deviation of the word count is about 30. The relatively narrow range of word counts suggests that the correlation with the total word count was low.

The same is true for the correlation between the accuracy measures and other text features. The standard deviation of the text features is relatively small in all proficiency bands; since all the indicators in CAF are calculated based on text features, the influence of text features would be unavoidable. However, the results of this study did not show any significant correlation, so the influence of text factors may be considered small. These results suggest that the interpretations of the changes of the written accuracy measures would not be biased by the textual features, hence it would be the robust assumption backing the warrants in the explanation inference (Figure 4.1).

4.1.4 Conclusion of Study 2

Study 2 examined the explanation inference which connected the written accuracy measures and a construct (i.e., written accuracy). Given that the construct which the WCR reflects has been controversial (e.g., Fox, 2019), confirming the explanation inference is essential for building the chain of logic in the argument-based approach and clarifying the constructs that the WCR is measuring.

The present study set the warrant (i.e, expected scores are attributed to a construct of accuracy) and three assumptions to provide backings to the warrant. The first assumption was that accuracy measures including WCR reflect the same construct (RQ2-1). The second assumption was that the correlations among accuracy measures, complexity measures (e.g., C/S), and construct were weak (RQ2-2). The third and last assumption was that the relationships between written accuracy measures and textual features (e.g., clause) are weak (RQ2-3).

As for RQ2-1, the present study conducted the EFA and showed that the written accuracy including WCR was grouped in the same construct (i.e., accuracy). In addition, Cronbach's α was also high (α = .82). Therefore, it would be possible to conclude that the

WCR would reflect the written accuracy which the traditional accuracy measures reflect. As Pallotti (2009) suggested, the present study also agreed that the construct of the written accuracy is the most internally coherent construct in the framework of CAF.

As for RQ2-2, the present study investigated the relationships between the constructs of the written accuracy and complexity based on the results of the EFA. The results showed that the correlation was quite low (r = .08), meaning that they would reflect the distinct constructs. Although several studies claimed that the WCR could measure the written accuracy and complexity (e.g., Fox, 2019), the present study showed the weak correlations between the constructs of the written accuracy and complexity.

In RQ2-3, the present study conducted the correlation analysis based on the MIC, which can investigate the nonlinear relationships of two observed variables and investigated the relationships between the written accuracy measures and the textual features. The analysis showed that the correlations between the written accuracy measures and the textual features were weak (MIC = .18–.26), suggesting that the interpretations of the changes of the written accuracy measures would not be biased by the textual features. These results have been obtained in previous studies (e.g., Barrot & Agdeppa, 2021), hence the present study also corresponded to previous studies.

Based on these results, the present study concluded that the explanation inference was confirmed positively; the expected score of written accuracy measures is attributed to a construct of accuracy. The construct which the WCR reflects has been discussed in the previous studies, hence the present study provides suggestions for future studies about the measurement of the written accuracy.

In Study 3, it is necessary to confirm the two inferences (i.e., extrapolation and utilization inferences). The extrapolation inference is the one that links expected scores to the target score or the proficiency levels. In addition, the utilization inference is the one that connects the target score and test score use. These two inferences are quite important to use the WCR when examining the development of the writing performance using the learner corpora.

Chapter 5

Study 3: Investigating the Relationships Between WCR and CEFR and Interpreting the Usefulness of the WCR

5.1 The Extrapolation and Utilization Inferences

5.1.1 Purposes and Research Questions

Study 2 showed that the WCR would be grouped in the written accuracy factor, which the traditional accuracy measures also reflected. Therefore, the explanation inference was confirmed positively. In Study 3, two inferences were examined (i.e., the extrapolation and utilization inferences. The extrapolation inference was investigated to link the scores of WCR and the target score. Moreover, the utilization inference was also tested to show whether the WCR can provide researchers and teachers with detailed and insightful information.

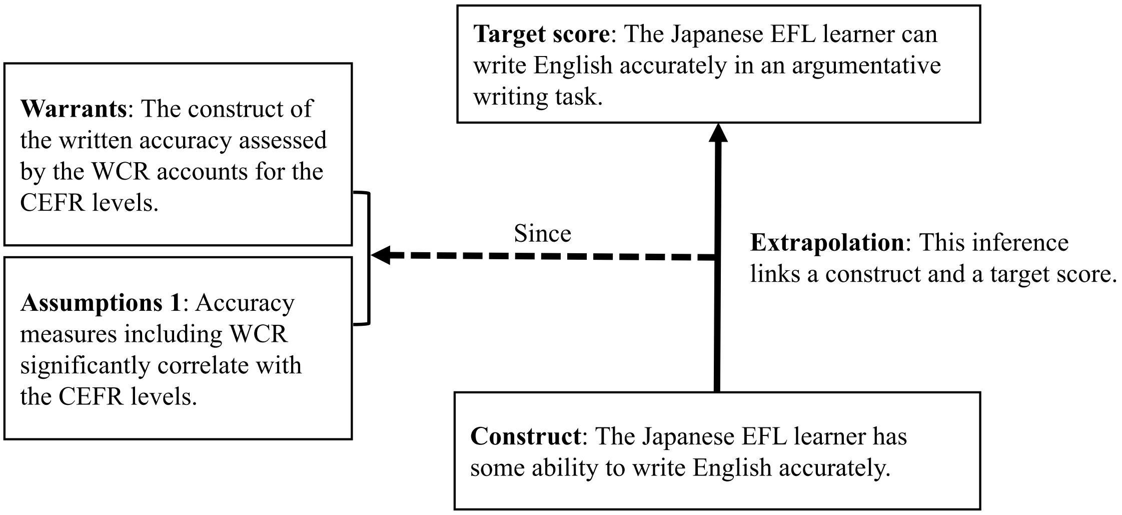
Figure 5.1 shows the extrapolation inference in the argument-based approach. The extrapolation inference is the one that links the construct of written accuracy to the target score. The target score in the present study is the CEFR levels. To verify the extrapolation inference, the warrant is that the construct of written accuracy as assessed by WCR relates to developing CEFR levels in the ICNALE corpus. Therefore, the assumption to be confirmed is that the score of WCR would correlate with CEFR levels.

The confirmation of the extrapolation inference is necessary to confirm the validation and essential for the studies of the performance development in L2 writing using learner corpora. Although numerous studies have investigated the relationships between the complexity in CAF and language proficiency (e.g., Barrot & Gabinete, 2019; Khushik & Huhta, 2019), written accuracy has not been explored.

Kojima and Kaneda (2020) have pointed that there is an opinion that written accuracy would not significantly correlate with English or writing proficiency, whereas others have reported the correlations between written accuracy and writing proficiency. Kojima and Kaneda (2020) investigated the relationships between the CAF measures and the writing proficiency using meta-analysis and reported that the written accuracy correlated with the writing proficiency (r = .44) stronger than the syntactic complexity (e.g., the mean length of T-units, r = .15). Significant correlations between the WCR and the CEFR levels would provide English teachers with insights regarding how to manipulate the instruction time and on which parts to focus in English classes. Therefore, it is necessary to investigate the relationships between the WCR and the CEFR levels to determine how L2 learners develop their written performances.

Figure 5.1

Summary of Extrapolation Inference



Target score: The Japanese EFL learner can write English accurately in an argumentative writing task.

Warrants: The construct of the written accuracy assessed by the WCR accounts for the CEFR levels.

Assumptions 1: Accuracy measures including WCR significantly correlate with the CEFR levels.

Extrapolation: This inference links a construct and a target score.

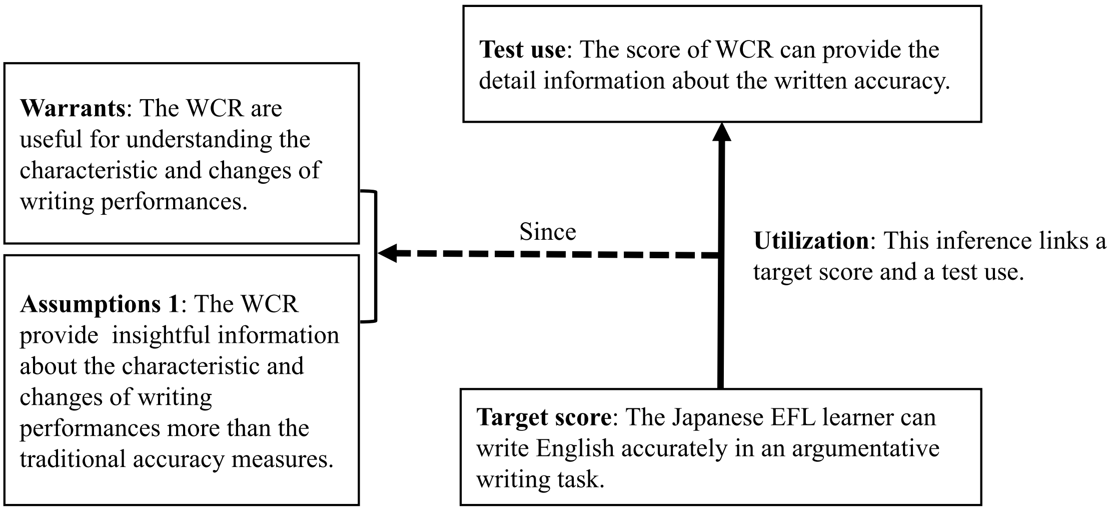
Construct: The Japanese EFL learner has some ability to write English accurately.

In addition, Figure 5.2 shows the utilization inference in the argument-based approach, which Study 3 also examined. The utilization inference plays a role in connecting the target score to test use. Given that the ICNALE was created to investigate the L2 writing developments of Asian learners, the WCR should provide researchers with detailed information about the written accuracy and be used as an accuracy measure to interpret the meaning of the score easily.

The warrant to confirm the utilization inference is that the evaluation of the written accuracy by WCR would provide more detailed information on small changes in the written accuracy. The amount of information obtained from evaluations should be indispensable for researchers to track even the slight changes in the development.

Figure 5.2

Summary of Utilization Inference


Test use: The score of WCR can provide the detail information about the written accuracy.
Warrants: The WCR are useful for understanding the characteristic and changes of writing performances.
Since
Utilization: This inference links a target score and a test use.
Assumptions 1: The WCR provide insightful information about the characteristic and changes of writing performances more than the traditional accuracy measures.
Target score: The Japanese EFL learner can write English accurately in an argumentative writing task.

As for confirming the utilization inference, the present study proved that the WCR can provide instructors and researchers with much information about the changes in writing performance in each CEFR level using descriptive statistics and graphs. In addition, the present study also showed a qualitative example of the writing performance evaluated with the WCR to discuss how well researchers and teachers can know and understand the characteristics. In sum, Study 3 investigated the extrapolation and utilization inferences. This study set the following two research questions (RQs) to examine the two inferences.

RQ 3-1: To what extent does WCR correlate with CEFR levels?

RQ 3-2: Does WCR provide more detailed information than the traditional accuracy measures in assessing the writing accuracy development?

5.1.2 Method

5.1.2.1 Participants

The present study used data used in Study 1, which is extracted from the International Corpus Network of Asian Learners of English developed by Ishikawa (2013). The 100 participants (44 females and 56 males, average age = 18.84 years) were majoring in various fields, including business, engineering, and economics. The participants' essays on two topics were analyzed. Note that the essay topics were (a) *It is important for college students to have a part-time job* (PTJ) and (b) *Smoking should be completely banned at all the restaurants in the country* (SMK). The average lengths of the PTJ and SMK essays were 223 words (*SD* = 24.1) and 219 words (*SD* = 26.1), respectively.

5.1.2.2 Written Accuracy Measures

Study 3 used four accuracy measures (i.e., WCR) based on the results of Study 2. Study 2 showed that these accuracy measures reflect the same construct of written accuracy. The WCR rating scale consists of four categories, definitions, and scores (Table 5.1). After all sentences in an essay were divided into clauses, each clause was categorized by its gravity of error according to the definitions. It should be noted that 0 should not be awarded because even Level 3 clauses are linguistically accurate to a certain degree (Foster & Wigglesworth, 2016). However, the traditional accuracy measures such as EFCR do not consider the gravity of errors. Table 5.2 shows the descriptions of all accuracy measures used in Study 3.

Table 5.1

Rating Scale of a Weighted Clause Ratio

Category	Definition	Score
No error	The clause is accurately constructed.	1.0
Level 1	The clause has only minor errors (e.g., morphosyntax) that do not compromise meaning.	0.8
Level 2	The clause contains serious errors (e.g., verb tense, word choice, or word order), but the meaning is recoverable, though not always obvious.	0.5
Level 3	The clause has very serious errors that make the intended meaning far from obvious and only partly recoverable.	0.1

Table 5.2

Description of Written Accuracy Measures

Measures	Code
Error-free clauses per total of all clause	EFCR
Error-free T-units per total of all T-unit	EFTR
Error-free T-units per total of all word	EFT/W
Weighted clause ratio	WCR

5.1.2.3 Complexity Measures

The present study used three complexity measures based on the previous studies (Barrot & Agdeppa, 2021; Kato, 2019): dependent clauses per total of all clause (DC/C); clauses per total of all sentence (C/S); and verb phrases per total of all T-unit (VP/T). The Table 5.3 shows the descriptions of the complexity measures.

Table 5.3

Description of Written Complexity Measures

Measures	Code	Descriptions
Dependent clauses per total of all clause	DC/C	Number of dependent clauses per total of all clause
Clauses per total of all sentence	C/S	Number of clauses per total of all sentence
Verb phrases per total of all T-unit	VP/T	Number of verb phrases per total of all T-unit

5.1.2.4 Scoring

Weighted clause ratio

Study 3 used WCR as an accuracy measure obtained in Study 1. Before the scoring, the researcher divided each sentence into clauses that were checked by the raters. Subsequently, raters independently evaluated all essays using the final version of the rating scale. Following the same procedure used during the rater training, raters were required to find errors in each clause and score the severity of errors according to the extent to which the error affects readers' comprehension.

Notably, the same errors (e.g., word errors) could be categorized under different levels because the severity of these errors was often contextual. The raters had agreed to read the definitions of the rating scale upon finding such errors to categorize them. Moreover, when there were multiple errors (e.g., Level 1 and Level 3 errors) in the same clause, the clause was categorized according to its worst-level error, as suggested by Foster and Wigglesworth (2016). The final WCR score was calculated as follows: WCR = the number of accurate clauses × 1.0 + the number of Lv.1 clauses × 0.8 + the number of Lv.2 clauses × 0.5 + the number of Lv.3 clauses × 0.1 / all clauses in the essay.

Traditional written accuracy measures

The traditional written accuracy measures (e.g., EFCR) were calculated based on the WCR data. The accuracy measures with the linguistic units (e.g., clauses, T-units, and sentence) were calculated by dividing the error-free linguistic units (e.g., error-free clauses) by the linguistic units (e.g., all clauses). Therefore, after calculating the score of the WCR, these traditional accuracy measures were produced.

Complexity measures

The complexity measures were calculated by the automated tool L2SCA (Lu, 2010). The reliability and validity were investigated in the previous studies (Lu, 2010, Polio & Yoon, 2018). According to the study of Lu (2010), the reliability of the L2SCA ranged from 0.834 to 1.00. Moreover, recent studies (Polio & Yoon, 2018) confirmed the reliability and validity of the L2SCA.

5.1.2.5 Data Analysis

This study used a correlation analysis using the Maximum information coefficient (MIC, Reshef et al., 2011) with *minerva* packages for RQ3-1. MIC is a statistical method for investigating the relationships between two variables that are non-linear. The MIC score is close to 1 if the relationships between the two variables are strong. In contrast, the MIC score will be close to 0 if the relationships between the two variables are weak.

In addition, the MIC score is positive even if the relationships are downward to the right. Therefore, it is necessary to use not only the MIC score but also the shape of the relationships in the plot to interpret the MIC meaning. Moreover, the present study used the HSIC test (Hilbert-Schmidt Independence Criteria; Gretton et al., 2005) to investigate the significance of the correlation.

Considering RQ3-2, the present study compared the WCR scores among the four proficiency groups, and the normality of the CAF measures was not confirmed. Therefore, this study conducted a non-parametric statistical analysis: the Kruskal-Wallis test (K–W test). To interpret the effect size *r*, based on Field et al. (2012), the present study set .10 as a small effect size, .30 as a medium effect size, and .50 as a large effect size.

In addition, the Steel-Dwass's multiple comparison test was used to compare the WCR scores in each group. While the Bonferroni method can be used, the method requires two assumptions: the normality and homogeneity of variance. Moreover, the Bonferroni method could cause a Type-1 error, which is the probability of rejecting the null hypothesis even though it is true because the *p*-value is adjusted according to the number of the comparison. In contrast, the Steel-Dwass's multiple comparison test can be used when the data do not have the normality and homogeneity of variance.

5.1.3 Results

Table 5.4 shows the descriptive statistics of the number of mean words in the two essays. The number of words tends to increase according to the development of the CEFR levels in the INCALE corpus.

Table 5.4

Descriptive Statistics of the Number of Words

Levels	No. of essays	Number of words			
		M	*SD*	Min	Max
A2	25	213.2	11.7	196	252
B1_1	25	212.8	21.8	182	303
B1_2	32	227.7	22.6	185	291
B2+	18	234.1	29.6	196	294

Table 5.5 shows the descriptive statistics of all measures used in the present study. As for the written accuracy measures, the WCR score was higher than the written accuracy measures produced by the linguistic units (e.g., EFCR and EFTR) because, based on the rating scale of WCR, all clauses have scores higher than 0. Therefore, the score is higher than those of other written accuracy measures using linguistic units such as EFCR. The mean score of WCR was 0.84, indicating that essays in the present study were easy to understand.

Before conducting a correlation analysis, this study checked the linearity between the two types of measures (i.e., the complexity and accuracy) and the CEFR levels. As expected, they did not correlate linearity due to the EFCR levels (e.g., A2, B1_1). Therefore, this study conducted a correlation analysis based on the MIC.

The analysis showed that the complexity measures correlated with the CEFR levels. The correlation between the measure of DC_C and the CEFR levels was the weakest in the three complexity measures (MIC = .26). In contrast, the correlation between the measure of VP_T and the CEFR levels was the strongest in the three complexity measures (MIC = .28).

Table 5.5

Descriptive Statistics for Measures

Measures	Code	*M*	*SD*	Min	Max
Clauses per total of all sentence	C_S	1.8	0.35	0.83	2.95
Dependent clauses per total of all clause	DC_C	0.37	0.08	0.18	0.6
Verb phrases per total of all T-unit	VP_T	2.15	0.37	1.25	3.12
Error-free T-unis per total of all T-unit	EFTR	0.25	0.11	0.08	0.64
Error-free T-unit per total of all word	EFT/W	0.02	0.01	0.01	0.06
Error-free clauses per total of all clause	EFCR	0.41	0.12	0.11	0.79
Weighted clause ratio	WCR	0.84	0.05	0.62	0.95

In addition, the written accuracy measures also correlated with the CEFR levels (Table 5.6 and Figure 5.3). The correlation between the measure of CFER and the CEFR levels was the weakest in the three complexity measures (MIC = .21). In contrast, the correlation between the measure of WCR and the CEFR levels was the strongest in the three complexity measures (MIC = .33).

Figure 5.3

Graph of Correlations

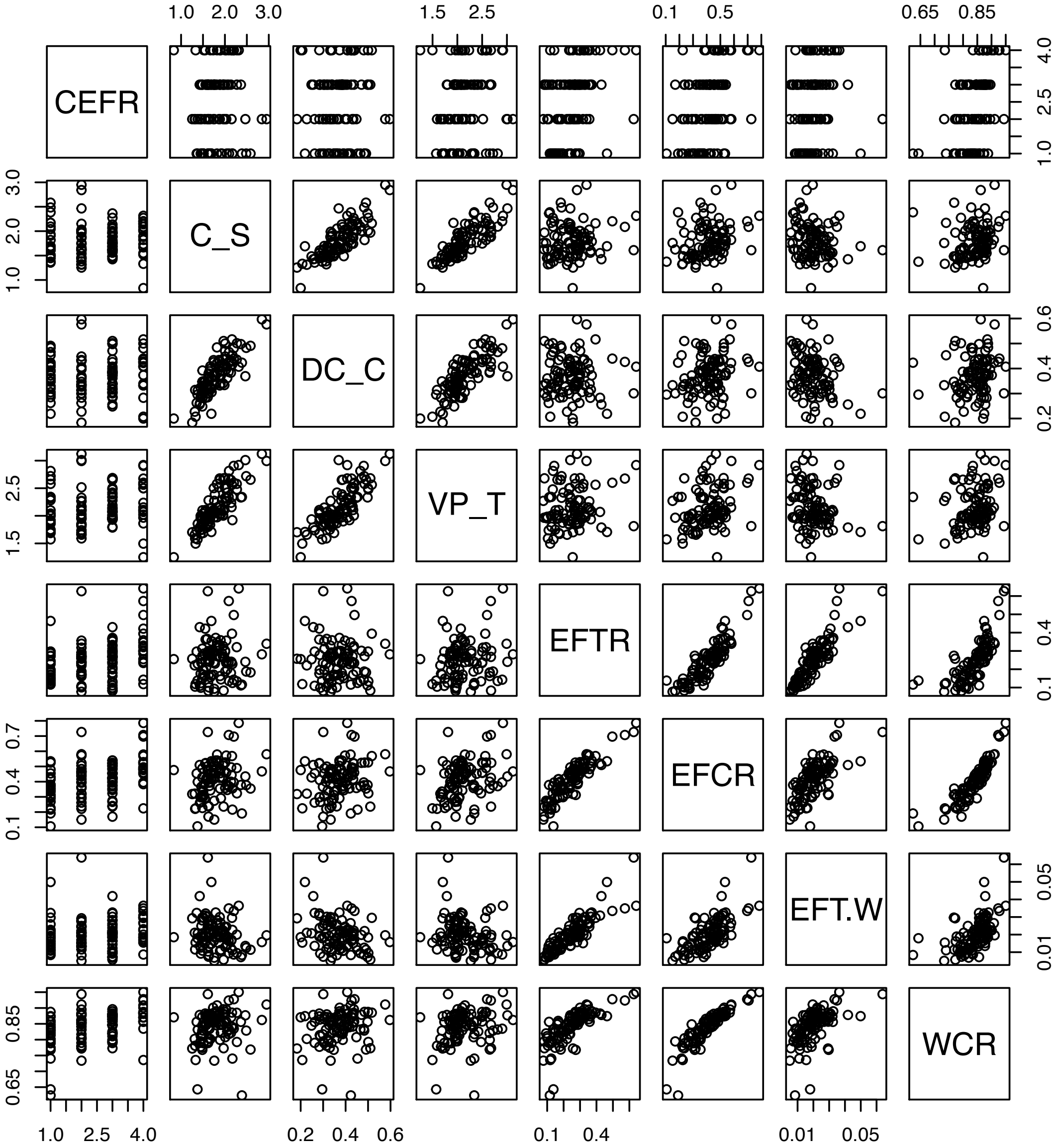


Table 5.6

Correlations Between Measures and CEFR Levels

Measures	Code	CEFR levels (MIC–R^2)
Clauses per total of all sentence	C_S	.26* (.26)
Dependent clauses per total of all clause	DC_C	.27 (.27)
Verb phrases per total of all T-unit	VP_T	.28 (.25)
Weighted clause ratio	WCR	.33** (.15)
Error-free clauses per total of all clause	EFCR	.21*** (.05)
Error-free T-unis per total of all T-unit	EFTR	.25*** (.13)
Error-free T-unit per total of all word	EFT/W	.23** (.20)

Note. $p^* < .05$, $p^{**} < .01$, $p^{***} < .001$ for the HSIC test.

Table 5.7 shows the descriptive statistics of the scores of four written accuracy measures in each English proficiency level in the ICNALE corpus. In addition, the four Figures (4 to 7) represent the graphs of each accuracy measure.

Table 5.7

Descriptive Statistics of Written Accuracy Measures

Measures	A2				B1_1				B1_2				B2+			
	M	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*
EFTR	0.21	0.08	0.12	0.46	0.24	0.11	0.08	0.63	0.24	0.09	0.08	0.43	0.34	0.13	0.12	0.64
EFCR	0.35	0.1	0.11	0.53	0.41	0.13	0.15	0.73	0.41	0.1	0.17	0.54	0.51	0.13	0.22	0.79
EFT/W	0.02	0.01	0.01	0.05	0.02	0.01	0.01	0.06	0.02	0.01	0.01	0.04	0.02	0.01	0.01	0.04
WCR	0.81	0.06	0.62	0.89	0.84	0.05	0.73	0.94	0.85	0.03	0.77	0.9	0.88	0.05	0.74	0.95

Note. EFTR = Error-free T-units per T-unit, EFCR = Error-free clauses per clause, EFT/W = Error -free T-units per word, WCR = Weighted clause ratio.

Figure 5.4

The Changes in EFTR Score

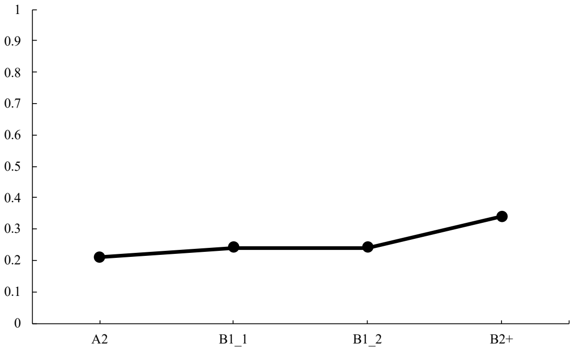


Figure 5.5

The Changes in EFCR Score

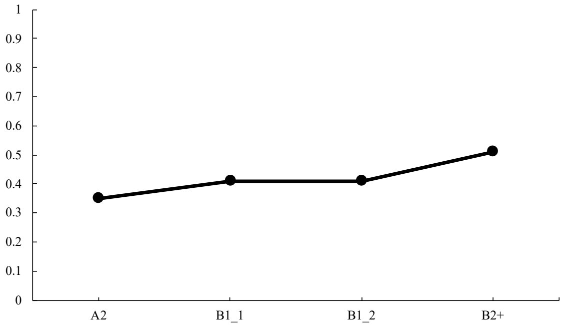
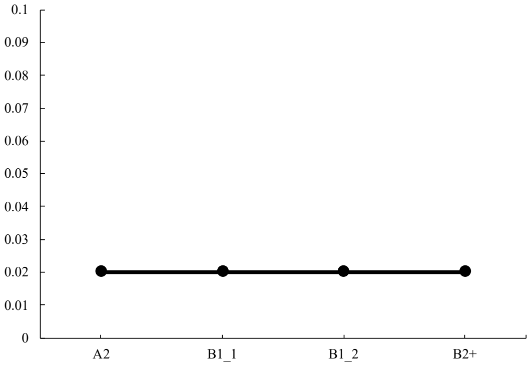


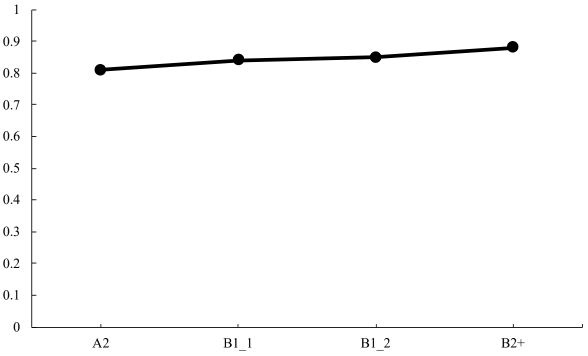
Figure 5.6

Figure 5.7

The Changes in EFT/W Score



The Changes in WCR Score



The mean scores of the EFTR and EFCR showed similar trends, and they tended to increase according to the increases in the CEFR levels. The K–W test revealed that the EFTR scores are different among four English levels, $H(3) = 14.56$, $p = .002$, $z = 3.05$, $r = .30$. Steel-Dwass's multiple comparison test was conducted and showed that the EFTR score between A2 and B2+ ($p = .002$, $z = 3.17$, $r = .32$) and B1_1 and B2+ ($p = .02$, z = 2.37, $r = .24$) were significantly different.

However, A2 and B1_1 was not significantly different, $p = .70$, $z = 0.38$, $r = .04$. In addition, A2 and B1_2 ($p = .37$, $z = 0.92$, $r = .09$), B1_1 and B1_2 ($p = .87$, $z = 0.16$, r

= .02), and B1_2 and B2+ (p = .08, z = 1.76, r = .18) were not significantly different. Table 5.8 summarizes the results of the changes in the EFTR.

Table 5.8

Summary of the Results of the K-W Test in the EFTR

Comparison	r (z-value)
A2 vs. B1_1	.04 (0.38)
A2 vs. B1_2	.09 (1.76)
A2 vs. B2+	.32 (3.17)**
B1_1 vs. B1_2	.02 (0.16)
B1_1 vs. B2+	.24 (2.37)*
B1_2 vs. B2+	.18 (1.76)

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

After examining the differences of the CEFR levels in EFTR scores, the EFCR scores were compared among the four CEFR levels. The K–W test revealed that the EFCR scores are different among four English levels, H(3) = 19.35, p < .001, z = 3.68, r = .37. The Steel-Dwass's multiple comparison test was conducted and showed that the EFCR score between A2 and B2+ (p < .000, z = 3.29, r = .35), B1_1 and B2+ (p = .002, z = 2.15, r = .26), and B1_2 and B2+ (p = .002, z = 2.31, r = .23) were significantly different.

In contrast, the results showed that the EFCR scores between A2 and B1_1 were not significantly different, p = .35, z = 0.95, r = .09. In addition, A2 and B1_2 (p = .06, z = 1.88, r = .19) and B1_1 and B1_2 (p = .97, z = 0.36, r = .00) were not significantly different. Table 5.9 summarizes the results of the changes in the EFCR.

Table 5.9

Summary of the Results of the K-W Test in the EFCR

Comparison	*r* (*z*-value)
A2 vs. B1_1	.09 (0.95)
A2 vs. B1_2	.19 (1.88)
A2 vs. B2+	.35 (3.29)***
B1_1 vs. B1_2	.00 (0.36)
B1_1 vs. B2+	.26 (2.15)**
B1_2 vs. B2+	.23 (2.31)**

Note. $p < .05$*, $p < .01$**, $p < .001$***.

Furthermore, the mean score of the EFT/W did not change at all, as shown by descriptive statistics. The K–W test revealed that the EFT/W scores were not different among four English levels, $H(3) = 19.35$, $p = .05$, $z = 1.93$, $r = .19$. The EFT/W scores between A2 and B2+ ($p = .04$, $z = 2.01$, $r = .21$) were significantly different.

In contrast, the EFT/W scores between A2 and B1_1 ($p = .78$, $z = 0.27$, $r = .03$), A2 and B1_2 ($p = .92$, $z = 0.10$, $r = .01$), B1_1 and B1_2 ($p = 1.00$, $z = 0.00$, $r = .00$), B1_1 and B2+ ($p = .12$, $z = 1.55$, $r = .16$), and B1_2 and B2+ ($p = .20$, $z = 1.26$, $r = .13$) were not significantly different. Table 5.10 summarizes the results of the changes in the EFT/W.

Table 5.10

Summary of the Results of the K-W Test in the EFT/W

Comparison	*r* (*z*-value)
A2 vs. B1_1	.03 (0.27)
A2 vs. B1_2	.01 (0.10)
A2 vs. B2+	.21 (2.01)*
B1_1 vs. B1_2	.00 (0.00)
B1_1 vs. B2+	.16 (1.55)
B1_2 vs. B2+	.13 (1.26)

Note. $p < .05$*.

The mean score of the WCR ranged from 0.81 to 0.88 and was the highest in the three written accuracy measures. The score tended to increase according to the increases in the CEFR levels as well as the EFCR and EFTR. In addition, as the increases between the levels of B1_1 and B1_2 were small in the EFCR and EFTR, the score of the WCR increased by only 0.1 point.

The K–W test revealed that the WCR scores are different among four English levels, $H(3) = 20.88$, $p < .001$, z = 3.86, $r = .39$. The Steel-Dwass's multiple comparison test was conducted and showed that the scores between A2 and B1_1 did not differ significantly, $p = .20$, z = 1.29, $r = .12$. In addition, the scores between B1_1 and B1_2 did also not differ significantly, $p = .92$, $z = 0.09$, $r = .01$.

However, there were significant differences between A2 and B1_2 ($p = .003$, z = 2.24, $r = .22$), A2 and B2+ ($p <.001$, z = 3.29, $r = .34$), B1_1 and B2+ ($p = .003$, z = 2.20, $r = .22$), and B1_2 and B2+ ($p = .002$, z = 2.28, $r = .23$). Table 5.11 summarizes the results of the changes in the WCR.

Table 5.11

Summary of the Results of the K-W Test in the WCR

Comparison	*r* (*z*-value)
A2 vs. B1_1	.12 (1.29)
A2 vs. B1_2	.22 (2.24)**
A2 vs. B2+	.34 (3.29)***
B1_1 vs. B1_2	.01 (0.09)
B1_1 vs. B2+	.22 (2.20)**
B1_2 vs. B2+	.23 (2.28)**

Note. $p < .05$*, $p < .01$**, $p < .001$***.

5.1.4 Discussion

RQ 3-1:To what extents does WCR correlate with CEFR levels? (See p.137)

The purpose of RQ3-1 was to reveal the relationships between the WCR as the written accuracy measure and the English proficiency levels in the ICNALE corpus. While the previous studies have primarily focused on the relationships between complexity and writing performance (e.g., Kyle & Crossley, 2018), few studies have investigated the relationships between written accuracy (e.g., WCR) and English proficiency. Educational institutions (e.g., universities) often use the CEFR, and there is a growing need to understand the differences between the English proficiency levels (Hawkings & Buttery, 2010). Therefore, the present study will play an important role in filling the research gap.

Moreover, confirming the correlation between the WCR and English proficiency will be essential for the writing research areas and the classes. The previous studies have investigated the correlation between the CAF measures and the writing performance and especially focused on how the complexity measures correlated with the writing and L2 performance (e.g., Kyle & Crossley, 2018). In contrast to these studies, Pallotti (2009) argued that many have claimed that a valid measure is not necessarily different among groups and does not necessarily correlate with other proficiency levels because, importantly, the measure reflects its underlying construct. It is important to recognize how adequate the measures reflect its theoretical construct.

However, it should also be noted that researchers could investigate further how the performances in each learner group are different because of the presence of a correlation. In addition, English teachers should know the features of performances produced by high-proficient learners and use their time on the areas that need improvement.

The present study conducted the correlation analysis with MIC, which can analyze the non-linear relationships between two variables, and investigated the relationships between the written accuracy measures, complexity measures, and CEFR levels. The results showed that the written accuracy and complexity measures correlated with the CEFR levels. Interestingly, the relationships between the WCR and the CEFR levels were the strongest (MIC = .33) in all measures.

The relationships between the written accuracy measures (e.g., WCR) and the English proficiency levels were higher than the complexity, which corresponds to the results in Kojima and Kaneda's (2020) study, although it focused on writing performance. Kojima and Kaneda's study investigated the relationships between CAF measures and writing performance. The results showed that the written accuracy measures correlated with the writing performance (r = .44) stronger than the syntactic complexity measures (r = .15). While it would be difficult to directly compare the results in the present study and Kojima and Kaneda's study because of the different statistical analyses, the fact that the written accuracy correlates with the proficiency was similar.

As for the reasons for the stronger correlation of the WCR with the CEFR levels when compared to other accuracy measures, the results obtained in the present study are natural given the advantages of the WCR and the characteristic of the writing performance. The traditional written accuracy measures (e.g., EFCR) have not considered the severity of linguistic errors. For example, the written accuracy measures have regarded subject-verb agreement errors and serious vocabulary choice errors as the same errors. However, Foster and Wigglesworth (2016) criticized this limitation and proposed the WCR to capture small changes in the written accuracy in the writing performance. Based on the evaluation with the WCR, any clause has a chance to obtain a score even if it has an error. In addition to the advantage of the WCR, it can be argued that the writing performance would be better according to the development of English proficiency. Therefore, the WCR would correlate with English proficiency because the more proficient learners produce understandable writing performance and can obtain a high WCR score.

Based on the results, the present study concluded that the extrapolation inference, which connects the construct and the target score, was confirmed positively. The present study clearly provided evidence that the WCR correlated with the CEFR levels for the warrant.

RQ 3-2: Does WCR provide more detailed information than the traditional accuracy measures in assessing the writing accuracy development? (See p.137)

RQ3-2 has the purpose of investigating whether the WCR can provide useful and detailed information about the development of written accuracy for researchers. It is critical to confirm the utilization inference not only to complete the validation studies but also to demonstrate the usefulness of the WCR in the corpus studies in writing studies.

There has been an increase in the number of writing studies investigating the development of writing performance based on the corpus (e.g., Barrot & Adgeppa, 2021). Hence, the amount of information is critical for researchers to deeply understand the characteristics of the development. They check the descriptive statistics of the scores to understand the trends among their participants. In teaching practices, teachers should interpret the meaning of the score easily and choose the appropriate instructions.

While the previous studies have focused on the construct and the correlations among other CAF measures, the present study offers suggestions about the ease of interpretations and the amount of information that the WCR can provide. In the discussion on RQ3-2, the present study demonstrated the usefulness of the WCR from the perspectives of academic studies and English teaching classes.

First, the present study used the basic statistical method (i.e., the descriptive statistics), which most studies have used, and showed the advantages of the WCR. Table 5.7 illustrates that, although the WCR has a higher score than other written, all measures would increase according to the development of the CEFR levels. However, the meaning of the final scores of the written accuracy is different. The EFCR is calculated by dividing the number of error-free clauses by the total number of clauses. The meaning of a score of 0.5 in an essay, for example, is interpreted as the essay having 50% accurate clauses. While the EFCR score is easy to understand, the EFCR cannot provide information about the other 50% parts of the essay, which have linguistic errors. This is because the EFCR regards any errors as the same and does not consider the influences of the errors (i.e., the error gravity). In contrast, the WCR shows how understandable the whole essay is. If the essay has a score of 0.5, it means that the entire essay is difficult to understand (Table 5.1).

Moreover, there is a substantial difference in the amount of information provided.

In particular, the descriptions of four categories in the WCR are the most informative among other written accuracy measures (e.g., EFCR and EFTR). As mentioned above, the EFCR and EFTR do not consider the error gravity. Foster and Wigglesworth (2016) criticized that "the fact that in a single clause one small error carries the same weight as several major errors means that finely grained differences in accuracy will remain below its radar (p.104)".

However, the WCR has four categories related to the clauses with a variety of errors. Evans et al. (2014) claimed that the WCR could detect small changes in the written accuracy in writing performances. Although Evans et al. (2014) reported that the data evaluated by one rater were not included because of the lower reliability and pointed out the difficulty of the evaluation, the high reliability in the present study was confirmed in Study 1.

The present study examined the differences of the four written accuracy scores among the four CEFR levels in the ICNALE corpus to determine how much the WCR can detect the differences in written accuracy. The K-W test and Steel-Dwass's multiple comparison test showed that the WCR was able to suggest the difference more than other written accuracy measures. In particular, the accuracy measures (e.g., EFCR and EFTR) did not show any difference between A2 and B1_2, but the WCR was able to detect the difference. As mentioned above, the WCR can detect small changes in the written accuracy in writing performances (Evans et al., 2014). Therefore, the present study succeeded in revealing the small differences in the written accuracy.

5.1.5 Conclusion

In conclusion, Study 3 is summarized at first. Subsequently, the summary of the whole results of the validation study (Study 1, Study 2, and Study 3) is presented.

5.1.5.1 The Summary of Study 3

Study 3 first examined the extrapolation inference, which links the construct of written accuracy to a target score. The confirmation of the extrapolation inference is not only necessary to confirm the validation but also for the studies of the performance development in L2 writing using learner corpora because numerous studies have investigated the relationships between the complexity in CAF and language proficiency (e.g., Barrot & Gabinete, 2019; Khushik & Huhta, 2019) but have not focused on written accuracy.

To verify the extrapolation inference, the present study set the warrant and the assumption. The warrant is that the construct of written accuracy as assessed by WCR relates to the development of CEFR levels in the ICNALE corpus. In addition, the assumption to be confirmed was that the score of WCR correlates with CEFR levels (RQ3-1).

As for RQ3-1, the present study conducted the correlation analysis between the written accuracy measures and the CFER levels set in the ICNALE corpus. The results showed that the written accuracy and complexity measures correlated with the CEFR levels. It should be noted that the relationships between the WCR and the CEFR levels were the strongest (MIC = .33) in all measures. The results corresponded to Kojima's (2020) suggestion that the accuracy measures considering the error gravity, such as the WCR, correlate with writing performance and English proficiency stronger than the written accuracy measures excluding the error gravity. Based on these results, the present study concluded that the extrapolation inference was confirmed positively.

After confirming the extrapolation inference, the present study investigated the usefulness of the WCR to verify the utilization inference (RQ3-2). The utilization inference connects the target score and test use. Given that the ICNALE was created to investigate the L2 writing developments of Asian learners, the WCR provides researchers with detailed information about written accuracy and the accuracy measure, allowing them to interpret the meaning of the score easily.

To confirm the utilization inference, the present study set the warrant (i.e., the evaluation of the written accuracy by WCR provides more detail information on small changes in written accuracy). The amount of information obtained from evaluations should be indispensable for researchers to determine even the slight changes in the development. In addition, the assumption was that the WCR provides insightful information about the characteristics and changes in writing performances more than the traditional accuracy measures.

The present study provided the descriptive statistics of the WCR and traditional accuracy measures and compared the amount of information to demonstrate how much information the WCR has (RQ3-2). The results showed that the WCR has more detailed information about the characteristics of the performance than other written accuracy measures. The K-W test and Steel-Dwass's multiple comparison test revealed that the WCR could detect the small changes in written accuracy in the four CEFR levels (e.g., A2–B1_2). Based on these results, the present study concluded that the utilization inference was confirmed positively.

5.1.5.2 The Summary of Validation Study

Table 5.12 shows the results of the whole validation study in the present study.

Table 5.12

Interpretive Argument in the Validation Study and the Results

Inference	Warrants	Assumptions	Results
Domain definition	The WCR represents the written accuracy of the writing performance obtained from the argumentative tests.	1. The WCR is representative of the written accuracy domain obtained in the argumentative writing.	Supported
Evaluation	Observation of writing performance evaluated by the WCR can be noted as the observed scores.	2. When raters evaluate the accuracy using the WCR, the reliability of the evaluation is appropriate.	Supported
Generalization	Observed scores of the WCR are estimates of expected scores over the parallel versions of tasks and across raters.	3. A sufficient number of tasks and raters are included to provide stable estimates of participants' performance.	Supported
Explanation	Expected scores of the WCR represent a construct of the written accuracy.	4. The relationship between the WCR and a construct corresponds to the theory. 5. There is little variation in the scores due to texts factors that can seriously affect the interpretation of the scores.	Supported

Interpretive Argument in the Validation Study and the Results (Continued)

Inference	Warrants	Assumptions	Results
Extrapolation	Expected scores of the construct are correlated with language performance over test situations.	6. WCR is correlated with English proficiency.	Supported
Utilization	The target scores provide the detail information about the language development.	7. The WCR provides a more detailed picture of the development of writing accuracy development in Japanese learners of English than traditional measures.	Supported

The validation study (Study 1 to Study 3) confirmed that the use of WCR would be valid when the measure is used for investigating the development of written accuracy. While previous studies using the WCR have used surveys to measure the effect of corrective feedback (e.g., Barrot, 2021) or studied its relationship to complexity (e.g., Fox, 2019), the present study focused on the validity of inferences of the WCR scores. In particular, this study tested the validity of the WCR when investigating the development of English language learners' accuracy using a learner corpus. Although validation of the use of WCR scores has been conducted in previous work (e.g., Evans et al., 2014), the evidence provided by these studies is insufficient for developmental studies using a corpus.

Recently, studies have examined the developmental process of CAF in L2 learners using the WCR (e.g., Barrot & Adgeppa, 2021). Because of the growing importance of the WCR in studies using learner corpora, investigating the validity of inferences made by the WCR scores is important for obtaining reliable research results. Based on the results, Study 4 and Study 5 investigated the development patterns of written accuracy in writing performance made by Japanese EFL learners.

Chapter 6

Study 4: Investigating the Development of Written Accuracy

6.1 The Relationships Among CAF Measures

6.1.1 Purposes and Research Questions

In Studies 1 through 3, this doctoral dissertation examined the validity of the WCR. The results showed that the inference from WCR scores to accuracy ability is valid for the purpose of using the corpus to capture changes in accuracy.

In Study 4, we used the WCR to assess the accuracy of English compositions produced by Japanese learners of English and investigate how their ability improves. Although some studies have used the WCR to explore the development of accuracy (e.g., Barrot & Adgeppa, 2021), they have captured changes in Asian learners of English as a whole. However, CAF also is influenced by factors such as learners' learning history and proficiency level. Therefore, it is difficult to know the characteristics of Japanese EFL learners' accuracy development from the overall changes in Asian learners of English.

In addition, accuracy change and development is also related to complexity and fluency. Polio and Shea (2014) argued that accuracy development may be less than expected due to the influence of complexity and fluency. If there is no significant difference in WCR scores between a given proficiency level, other factors (e.g., complexity) may also have an impact.

Therefore, this study used the numerical values measured by the WCR to elucidate the developmental patterns of accuracy in Japanese EFL learners. We also investigated how complexity and fluency play a role in this development. This present study set two research questions (RQs).

RQ 4-1: Do complexity and fluency measures change in accordance with the changes in EFCR scores?

RQ 4-2: Do complexity and fluency measures change as the WCR scores change? If so, is there any difference in relationships with complexity and fluency measures between the EFCR and WCR?

6.1.2 Method

6.1.2.1 Participants

The present study used data used in Study 1, which is extracted from the International Corpus Network of Asian Learners of English developed by Ishikawa (2013). The 100 participants (44 females and 56 males, average age = 18.84 years) were majoring in various fields, including business, engineering, and economics. The participants' essays on two topics were analyzed. Note that the essay topics were (a) *It is important for college students to have a part-time job* (PTJ) and (b) *Smoking should be completely banned at all the restaurants in the country* (SMK). The average lengths of the PTJ and SMK essays were 223 words (*SD* = 24.1) and 219 words (*SD* = 26.1), respectively.

6.1.2.2 Written Accuracy Measures

Study 4 used WCR and EFCR as the written accuracy measure. The WCR rating scale consists of four categories, definitions, and scores (Table 6.1). After all sentences in an essay were divided into clauses, each clause was categorized by its gravity of error according to the definitions. It should be noted that 0 should not be awarded because even Level 3 clauses are linguistically accurate to a certain degree (Foster & Wigglesworth, 2016). However, the traditional accuracy measures such as EFCR do not consider the gravity of errors. In addition, EFCR was calculated based on the WCR data. EFCR was calculated by dividing the error-free linguistic units (e.g., error-free clauses) by the linguistic units (e.g., all clauses).

Table 6.1

Rating Scale of a Weighted Clause Ratio

Category	Definition	Score
No error	The clause is accurately constructed.	1.0
Level 1	The clause has only minor errors (e.g., morphosyntax) that do not compromise meaning.	0.8
Level 2	The clause contains serious errors (e.g., verb tense, word choice, or word order), but the meaning is recoverable, though not always obvious.	0.5
Level 3	The clause has very serious errors that make the intended meaning far from obvious and only partly recoverable.	0.1

6.1.2.3 Complexity Measures

The present study used three complexity measures based on the previous studies (Barrot & Agdeppa, 2021; Kato, 2019); dependent clauses per total of all clause (DC/C), clauses per total of all sentence (C/S), and verb phrases per total of all T-unit (VP/T). Table 6.2 shows the descriptions of the complexity measures. In addition, the present study added three overall complexity measures (i.e., mean length of the clause, mean length of T-units, and mean length of sentence).

Table 6.2

Description of Written Complexity Measures

Measures	Code	Descriptions
Dependent clauses per total of all clause	DC/C	Number of dependent clauses per total of all clause
Clauses per total of all sentence	C/S	Number of clauses per total of all sentence
Verb phrases per total of all T-unit	VP/T	Number of verb phrases per total of all T-unit
Mean length of clause	MLC	Number of words per clause
Mean length of T-units	MLT	Number of words per T-unit
Mean length of sentence	MLS	Number of words per sentence

6.1.2.4 Fluency Measures

The present study used three fluency measures based on the previous studies (Barrot & Agdeppa, 2021); words per text (W/Tx), T-units per text (T/Tx), and Clauses per text (C/Tx). Table 6.3 shows the descriptions of the fluency measures.

Table 6.3

Description of the Fluency Measures

Measures	Code	Descriptions
Words per text	W/Tx	Number of words per text
T-units per text	T/Tx	Number of T-units per text
Clauses per text	C/Tx	Number of clauses per text

6.1.2.5 Scoring

Weighted clause ratio

Study 4 used WCR as an accuracy measure obtained in Study 1. Before the scoring, the researcher divided the sentences in each essay into clauses which was checked by the raters. Subsequently, raters independently evaluated all essays using the final version of the rating scale. Following the same procedure used during the rater training, raters were required to find errors in each clause and score the severity of errors according to the extent to which the error affects readers' comprehension.

Importantly, the same errors (e.g., word errors) could be categorized under different levels because the severity of these errors was often contextual. The raters had agreed to read the definitions of the rating scale upon finding such errors to categorize them. Moreover, when there were multiple errors (e.g., Level 1 and Level 3 errors) in the same clause, the clause was categorized according to its worst-level error, as suggested by Foster and Wigglesworth (2016). The final WCR score was calculated as follows: WCR = the number of accurate clauses × 1.0 + the number of Lv.1 clauses × 0.8 + the number of Lv.2 clauses × 0.5 + the number of Lv.3 clauses × 0.1 / all clauses in the essay.

Complexity measures

The complexity measures were calculated by the automated tool L2SCA (Lu, 2010). The reliability and validity were investigated in the previous studies (Lu, 2010, Polio & Yoon, 2018). According to the study of Lu (2010), the reliability of the L2SCA ranged from 0.834 to 1.00. Moreover, a recent study (Polio & Yoon, 2018) confirmed the reliability and validity of the L2SCA.

Fluency measures

The fluency measures were calculated using the data obtained in calculations of the written accuracy measures (e.g., EFCR, EFTR, and EFSR). The number of linguistic units, which are also necessary for the fluency measures, was already produced in Study 1. Study 4 used the data and calculated three fluency measures.

6.1.2.6 Data Analysis

As for RQ4-1 and 4-2, the present study compared the WCR scores among the four proficiency groups, and the normality of the CAF measures was not confirmed. Therefore, this study conducted a non-parametric statistical analysis: the Kruskal-Wallis test (K–W test). To interpret the effect size *r*, based on Field et al. (2012), the present study set .10 as a small effect size, .30 as a medium effect size, and .50 as a large effect size.

In addition, the Steel-Dwass's multiple comparison test was used to compare the WCR scores in each group. While the Bonferroni method can be used, the method requires two assumptions: the normality and homogeneity of variance. Moreover, the Bonferroni method could cause a Type-1 error, which is the probability of rejecting the null hypothesis even though it is true because the *p*-value is adjusted according to the number of the comparison. In contrast, the Steel-Dwass's multiple comparison test can be used when the data do not have the normality and homogeneity of variance.

6.1.3 Results

Table 6.4 shows the descriptive statistics of the number of mean words in two essays. The number of words tended to increase according to the development of the CEFR levels in the INCALE corpus.

Table 6.4

Descriptive Statistics of the Number of Words

Levels	No. of essays	No. of words			
		M	*SD*	Min	Max
A2	25	213.2	11.7	196	252
B1_1	25	212.8	21.8	182	303
B1_2	32	227.7	22.6	185	291
B2+	18	234.1	29.6	196	294

Table 6.5 shows the descriptive statistics of all measures used in the present study. As for the written accuracy measures, the WCR score was higher than the written accuracy measures produced by the linguistic units (e.g., EFCR, EFTR) because, based on the rating scale of WCR, all clauses had scores higher than 0. Therefore, the score was higher than those of other written accuracy measures using linguistic units such as EFCR. The mean score of WCR was 0.84, indicating that essays in the present study were easy to understand. In addition, Table 6.6 is the descriptive statistics of scores all measures in each CEFR level.

Table 6.5

Descriptive Statistics for Measures

Measures	Code	*M*	*SD*	*Min*	*Max*
Weighted clause ratio	WCR	0.84	0.05	0.62	0.95
Dependent clauses per total of all clause	DC/C	0.37	0.08	0.18	0.6
Clauses per total of all sentence	C/S	1.80	0.35	0.83	2.95
Verb phrases per total of all T-unit	VP/T	2.15	0.37	0.32	1.25
Mean length of clause	MLC	7.84	0.84	4.88	10.34
Mean length of T-unit	MLT	12.93	2.36	7.91	18.77
Mean length of sentence	MLS	14.20	3.14	7.91	22.82
Words per text	W/Tx	221.48	23.14	181.5	303
T-units per text	T/Tx	0.08	0.02	0.05	0.13
Clauses per text	C/Tx	0.14	0.02	0.09	0.23

Table 6.6

Descriptive Statistics for Measures in Each CEFR Level

Measure	A2				B1_1			
	M	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*
WCR	0.81	0.06	0.62	0.89	0.84	0.05	0.73	0.94
DC/C	0.37	0.07	0.22	0.49	0.36	0.09	0.18	0.60
C/S	1.79	0.36	1.36	2.58	1.80	0.44	1.25	2.95
VP/T	2.08	0.34	1.57	2.81	2.11	0.42	1.59	3.12
MLC	7.46	0.67	6.41	9.04	7.65	0.64	6.69	9.21
MLT	12.22	2.25	9.16	16.53	12.42	2.34	9.29	18.01
MLS	13.34	3.17	9.16	20.89	13.61	3.37	10.16	22.82
W/Tx	213.16	11.65	195.50	252	212.78	21.80	181.5	303
T/Tx	0.09	0.02	0.06	0.11	0.09	0.02	0.05	0.12
C/Tx	0.14	0.02	0.10	0.19	0.14	0.03	0.09	0.23

Descriptive Statistics for Measures in Each CEFR Level (Continued)

Measure	B1_2				B2+			
	M	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*
WCR	0.85	0.03	0.77	0.9	0.88	0.05	0.74	0.95
DC/C	0.37	0.07	0.25	0.51	0.39	0.09	0.20	0.52
C/S	1.79	0.24	1.43	2.36	1.87	0.38	0.83	2.31
VP/T	2.19	0.28	1.79	2.68	2.25	0.45	1.25	2.92
MLC	8.29	0.73	7.10	10.34	7.84	1.15	4.88	9.57
MLT	13.55	2.07	10.62	18.77	13.52	2.76	7.91	16.86
MLS	14.79	2.42	10.84	19.65	15.16	3.67	7.91	22.11
W/Tx	227.66	22.55	185	290.5	234.14	29.58	195.5	293.5
T/Tx	0.08	0.01	0.06	0.10	0.07	0.02	0.06	0.13
C/Tx	0.13	0.01	0.11	0.15	0.13	0.02	0.10	0.18

Note. WCR = Weighted clause ratio, DC/C = Dependent clauses per total of all clause, C/S = Clauses per total of all sentence, VP/T = Verb phrases per total of all T-unit, MLC = Mean length of clause, MLT = Mean length of T-unit, MLS = Mean length of sentence, W/Tx = Words per text, T/Tx = T-units per text, C/Tx = Clauses per text.

Although the WCR and EFCR scores were shown in Study 3, these results were reviewed again. Figure 6.1 shows the changes in the WCR scores. The K–W test revealed that the WCR scores were different among the four English levels, $H(3) = 20.88, p < .001$, z = 3.86, $r = .39$. The Steel-Dwass's multiple comparison test was conducted and showed that the scores between A2 and B1_1 did not differ significantly, $p = .20$, z = 1.29, $r = .12$. In addition, the scores between B1_1 and B1_2 did also not differ significantly, $p = .92$, $z = 0.09$, $r = .00$.

In contrast, there were significant differences between A2 and B1_2 ($p = .003$, z = 2.24, $r = .22$), A2 and B2+ ($p <.001$, z = 3.29, $r = .34$), B1_1 and B2+ ($p = .003$, z = 2.20, $r = .22$), and B1_2 and B2+ ($p = .002$, z = 2.28, $r = .23$). Table 6.7 summarizes the results of the changes in the WCR.

Figure 6.1

Changes in WCR Score

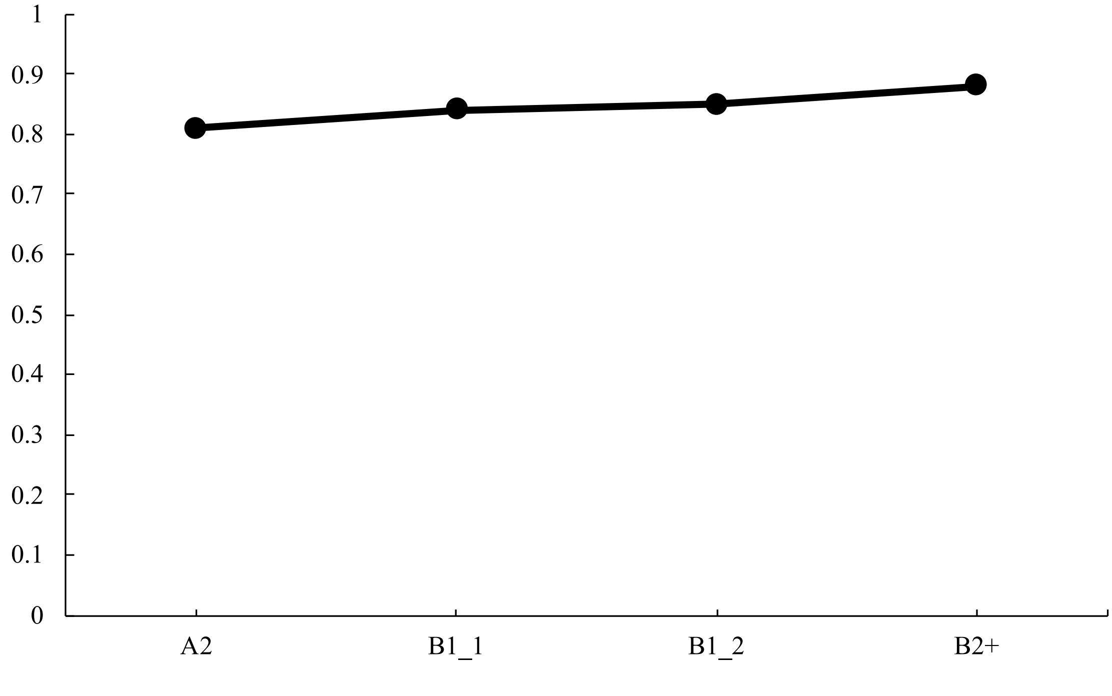


Table 6.7

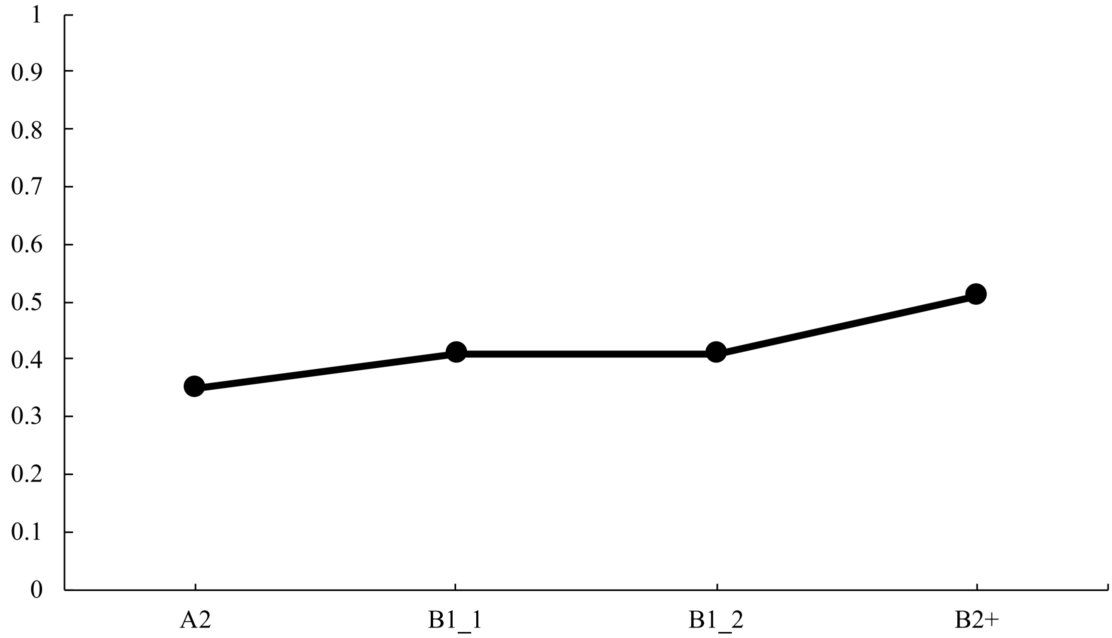
Summary of the Results of K-W Test in WCR

Comparison	*r* (*z*-value)
A2 vs. B1_1	.12 (1.29)
A2 vs. B1_2	.22 (2.24)**
A2 vs. B2+	.34 (3.29)***
B1_1 vs. B1_2	.92 (0.09)
B1_1 vs. B2+	.22 (2.20)**
B1_2 vs. B2+	.23 (2.28)**

Note. $p < .05$*, $p < .01$**, $p < .001$***.

Figure 6.2

Changes in EFCR Score



Moreover, Figure 6.2 describes the changes in the EFCR scores. The K–W test revealed that the EFCR scores are different among the four English levels, $H(3) = 19.35$, $p < .001$, z = 3.68, $r = .37$. The Steel-Dwass's multiple comparison test was conducted and showed that the EFCR score between A2 and B2+ ($p < .000$, z = 3.29, $r = .35$), B1_1 and B2+ ($p = .002$, z = 2.15, $r = .26$), and B1_2 and B2+ ($p = .002$, z = 2.31, $r = .23$) were significantly different.

In contrast, the results showed that the EFCR scores between A2 and B1_1 were not significantly different, $p = .35$, z = 0.95, $r = .09$. In addition, A2 and B1_2 ($p = .06$, z = 1.88, $r = .19$) and B1_1 and B1_2 ($p = .97$, z = 0.36, $r = .00$) were not significantly different. Table 6.8 summarizes the results of the changes in the EFCR.

Table 6.8

Summary of the Results of K-W Test in EFCR

Comparison	*r* (*z*-value)
A2 vs. B1_1	.09 (0.95)
A2 vs. B1_2	.19 (1.88)
A2 vs. B2+	.35 (3.29)***
B1_1 vs. B1_2	.00 (0.36)
B1_1 vs. B2+	.26 (2.15)**
B1_2 vs. B2+	.23 (2.31)**

Note. $p < .05$*, $p < .01$**, $p < .001$***.

After analyzing the WCR scores among four CEFR levels, the complexity measures were analyzed. Figure 6.3 shows the changes in the VP/T scores. The K–W test revealed that the VP/T scores were not different among the four English levels, $H(3) =$

4.85, *p* = .18, z = 1.33, *r* = .13.

Moreover, the VP/T scores between A2 and B1_1 (*p* = .99, z = 0.00, *r* = .00), A2 and B1_2 (*p* = .41, z = 0.82, *r* = .08), A2 and B2+ (*p* = .34, z = 0.95, *r* = .10), B1_1 and B1_2 (*p* = .54, z = 0.62, *r* = .06), B1_1 and B2+ (*p* = .50, z = 0.67, *r* = .07), and B1_2 and B2+ (*p* = .77, z = 0.29, *r* = .03) were not significantly different. Table 6.9 summarizes the results of the changes in the VP/T.

Figure 6.3

Changes in VP/T Score

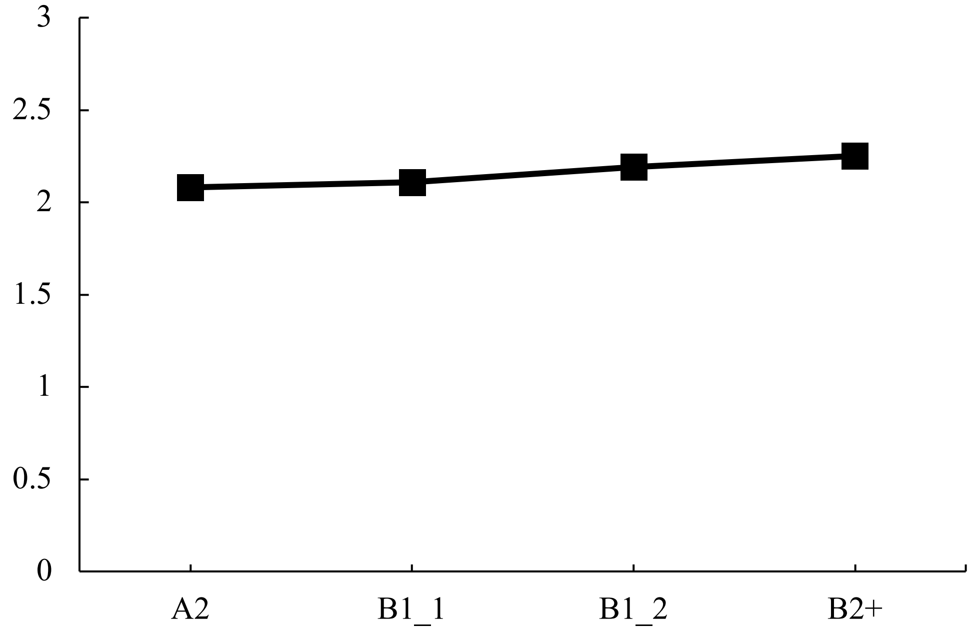


Table 6.9

Summary of the Results of K-W Test in VP/T

Comparison	*r* (*z*-value)
A2 vs. B1_1	.00 (0.00)
A2 vs. B1_2	.08 (0.82)
A2 vs. B2+	.10 (0.95)
B1_1 vs. B1_2	.06 (0.62)
B1_1 vs. B2+	.07 (0.67)
B1_2 vs. B2+	.03 (0.29)

Figure 6.4 shows the changes in the DC/C scores. The K–W test revealed that the DC/C scores were not different among the four English levels, *H*(3) = 2.51, *p* = .47, z = 0.72, *r* = .07.

Moreover, the DC/C scores between A2 and B1_1 (*p* = .94, z = 0.08, *r* = .01), A2 and B1_2 (*p* = .98, z = 0.03, *r* = .00), A2 and B2+ (*p* = .76, z = 0.30, *r* = .03), B1_1 and B1_2 (*p* = 1.00, z = 0.01, *r* = .00), B1_1 and B2+ (*p* = .60, z = 0.53, *r* = .05), and B1_2 and B2+ (*p* = .46, z = 0.74, *r* = .07) were not significantly different. Table 6.10 summarizes the results of the changes in the DC/C.

Figure 6.4

Changes in DC/C Score

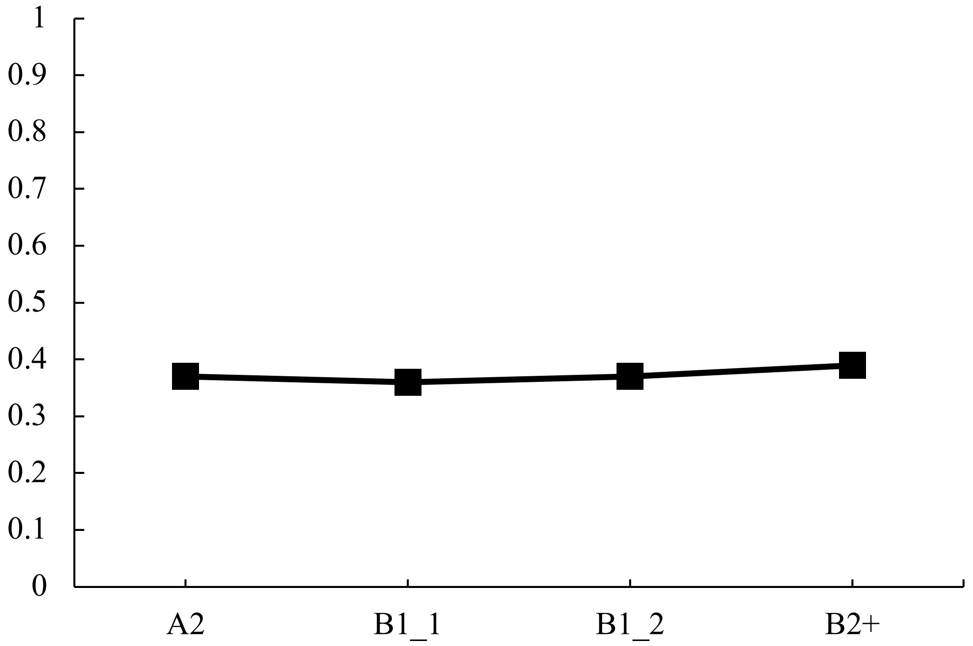


Table 6.10

Summary of the Results of K-W Test in DC/C

Comparison	*r* (*z*-value)
A2 vs. B1_1	.01 (0.08)
A2 vs. B1_2	.00 (0.03)
A2 vs. B2+	.03 (0.08)
B1_1 vs. B1_2	.00 (0.01)
B1_1 vs. B2+	.05 (0.53)
B1_2 vs. B2+	.07 (0.74)

Figure 6.5 shows the changes in the C/S scores. The K–W test revealed that the C/S scores were not different among the four English levels, $H(3) = 3.11$, $p = .37$, z = 0.89, *r*

= .08.

Moreover, the C/S scores between A2 and B1_1 (*p* = 1.00, z = 0.01, *r* = .00), A2 and B1_2 (*p* = .96, z = 0.05, *r* = .00), A2 and B2+ (*p* = .60, z = 0.54, *r* = .05), B1_1 and B1_2 (*p* = .92, z = 0.10, *r* = .01), B1_1 and B2+ (*p* = .49, z = 0.69, *r* = .07), and B1_2 and B2+ (*p* = .45, z = 0.75, *r* = .08) were not significantly different. Table 6.11 summarizes the results of the changes in the C/S.

Figure 6.5

Changes in C/S Score

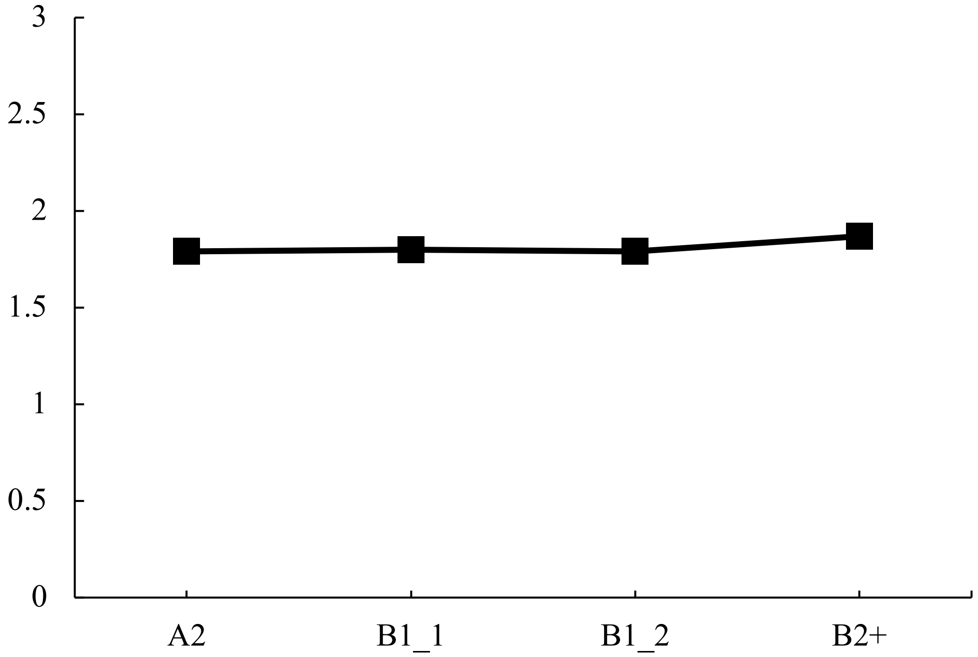


Table 6.11

Summary of the Results of the K-W Test in C/S

Comparison	*r* (*z*-value)
A2 vs. B1_1	.00 (0.01)
A2 vs. B1_2	.00 (0.05)
A2 vs. B2+	.05 (0.54)
B1_1 vs. B1_2	.01 (0.10)
B1_1 vs. B2+	.07 (0.69)
B1_2 vs. B2+	.08 (0.75)

Figure 6.6 describes the changes in the MLC scores. The K–W test revealed that the MLC scores were different among four English levels, $H(3) = 17.63$, $p < .001$, z = 3.47, $r = .35$. The Steel-Dwass's multiple comparison test was conducted and showed that the scores between A2 and B1_2 differed significantly, $p < .001$, z = 3.62, $r = .36$. In addition, the scores between B1_1 and B1_2 also differed significantly, $p = .01$, $z = 2.80$, $r = .28$.

However, the MLC scores between A2 and B1_1 ($p = .77$, z = 0.29, $r = .03$), A2 and B2+ ($p = .36$, z = 0.92, $r = .09$), B1_2 and B2+ ($p = .62$, z = 0.49, $r = .05$), and B1_2 and B2+ ($p = .68$, z = 0.42, $r = .04$) were not significantly different. Table 6.12 summarizes the results of the changes in the MLC.

Figure 6.6

Changes in MLC Score

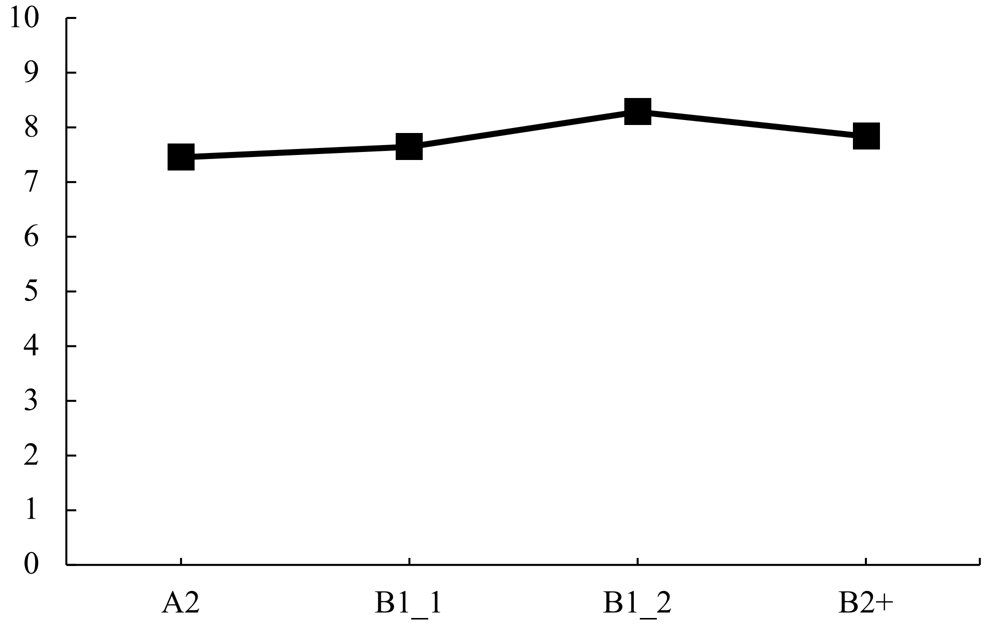


Table 6.12

Summary of the Results of the K-W Test in MLC

Comparison	*r* (*z*-value)
A2 vs. B1_1	.03 (.029)
A2 vs. B1_2	.36 (3.62)***
A2 vs. B2+	.09 (0.92)
B1_1 vs. B1_2	.28 (2.80)**
B1_1 vs. B2+	.05 (0.49)
B1_2 vs. B2+	.04 (0.42)

Note. *p* < .05*, *p* < .01**, *p* < .001***.

Figure 6.7 describes the changes in the MLT scores. The K–W test revealed that the MLT scores were different among the four English levels, *H*(3) = 8.38, *p* = .04, z = 2.01, *r* = .21.

However, the MLT scores between A2 and B1_1 (*p* = .99, z = 0.01, *r* = .00), A2 and B1_2 (*p* = .11, z = 1.59, *r* = .16), A2 and B2+ (*p* = .21, z = 1.25, *r* = .13), B1_1 and B1_2 (*p* = .13, z = 1.54, *r* = .15), B1_1 and B2+ (*p* = .34, z = 0.90, *r* = .09), and B1_2 and B2+ (*p* = .97, z = 0.02, *r* = .00) were not significantly different. Table 6.13 summarizes the results of the changes in the MLT.

Figure 6.7

Changes in MLT Score

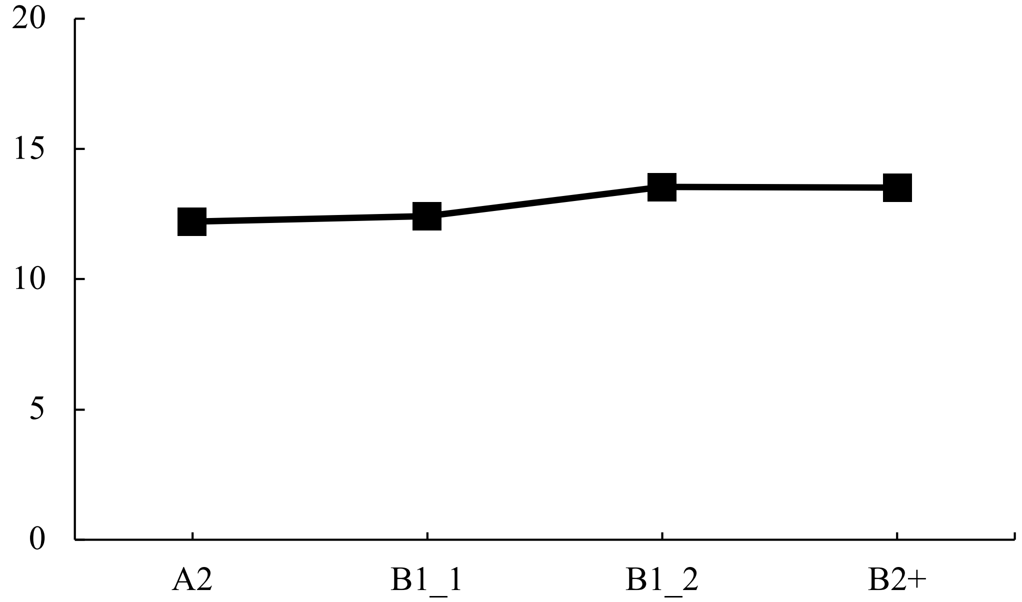


Table 6.13

Summary of the Results of K-W Test in MLT

Comparison	*r* (*z*-value)
A2 vs. B1_1	.00 (0.01)
A2 vs. B1_2	.16 (1.59)
A2 vs. B2+	.13 (1.25)
B1_1 vs. B1_2	.15 (1.54)
B1_1 vs. B2+	.09 (0.90)
B1_2 vs. B2+	.00 (0.02)

Figure 6.8 shows the changes in the MLS scores. The K–W test revealed that the MLT scores were different among the four English levels, *H*(3) = 8.05, *p* = .04, z = 2.01, *r* = .20. However, the MLS scores between A2 and B1_1 (*p* = 1.00, z = 0.01, *r* = .00), A2 and B1_2 (*p* = .15, z = 1.43, *r* = .14), A2 and B2+ (*p* = .25, z = 1.15, *r* = .12), B1_1 and B1_2 (*p* = .16, z = 1.41, *r* = .14), B1_1 and B2+ (*p* = .31, z = 1.01, *r* = .10), and B1_2 and B2+ (*p* = .91, z = 0.11, *r* = .01) were not significantly different. Table 6.14 summarizes the results of the changes in the MLS.

Figure 6.8

Changes in MLS Score

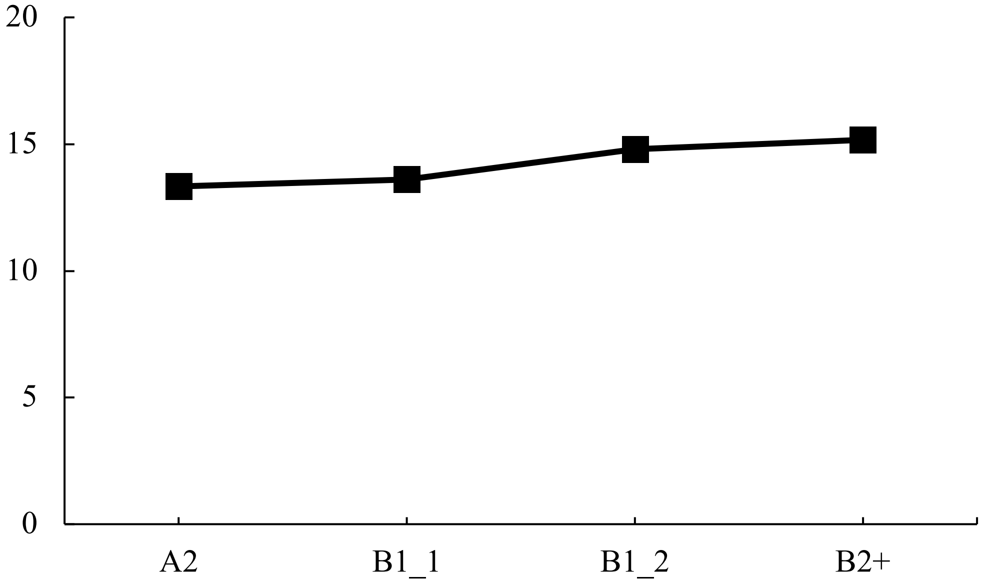


Table 6.14

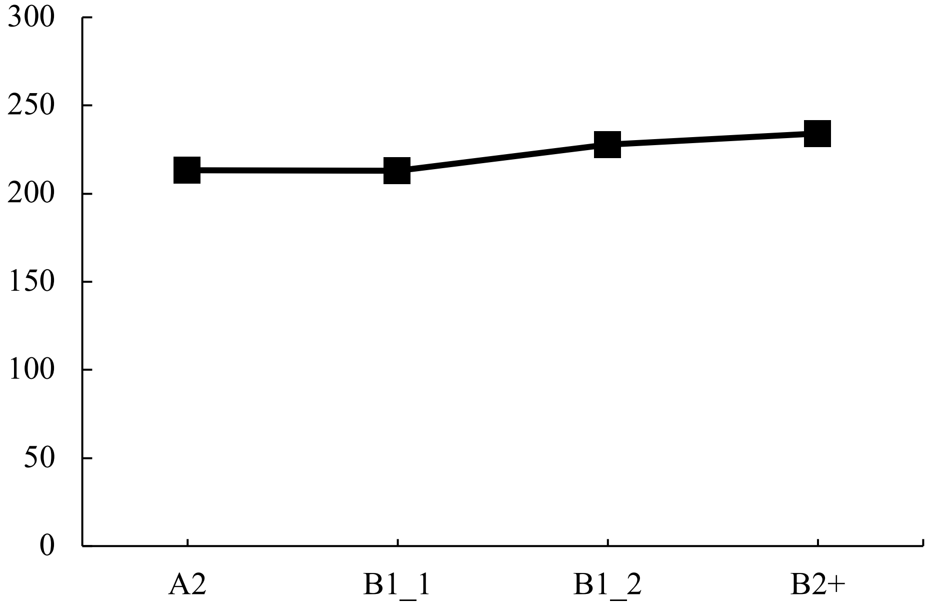
Summary of the Results of K-W Test in MLS

Comparison	*r* (*z*-value)
A2 vs. B1_1	.00 (0.01)
A2 vs. B1_2	.14 (1.43)
A2 vs. B2+	.12 (1.15)
B1_1 vs. B1_2	.14 (1.41)
B1_1 vs. B2+	.10 (1.01)
B1_2 vs. B2+	.01 (0.11)

After analyzing the complexity measures' scores among the four CEFR levels, the fluency measures were analyzed. Figure 6.9 shows the changes in the W/Tx scores.

Figure 6.9

Changes in W/Tx Score



The K–W test revealed that the W/Tx scores were different among the four English levels, $H(3) = 18.17$, $p < .001$, z = 3.54, $r = .35$. Moreover, the W/Tx scores between A2 and B1_2 ($p < .05$, z = 2.89, $r = .29$), B1_1 and B1_2 ($p < .05$, z = 3.16, $r = .34$), and B1_1 and B2+ ($p < .05$, z = 2.03, $r = .20$) were significantly different.

In contrast, the W/Tx scores between A2 and B1_1 ($p = .89$, z = 0.14, $r = .01$), A2 and B2+ ($p = .08$, z = 1.74, $r = .17$), and B1_2 and B2+ ($p = .99$, z = 0.01, $r = .00$) were not significantly different. Table 6.15 summarizes the results of the changes in the W/Tx.

Table 6.15

Summary of the Results of K-W Test in W/Tx

Comparison	*r* (*z*-value)
A2 vs. B1_1	.01 (0.14)
A2 vs. B1_2	29 (2.89)*
A2 vs. B2+	.17 (1.74)
B1_1 vs. B1_2	.34 (3.16)*
B1_1 vs. B2+	.20 (1.74)*
B1_2 vs. B2+	.00 (0.01)

Note. *p* < .05*.

Figure 6.10 shows the changes in the C/Tx scores. The K–W test revealed that the C/Tx scores were not different among four English levels, *H*(3) = 6.24, *p* = .10, z = 1.64, *r* = .16. In addition, the C/Tx scores between A2 and B1_1 (*p* = .87, z = 0.16, *r* = .01), A2 and B1_2 (*p* = .13, z = 1.51, *r* = .15), A2 and B2+ (*p* = 1.00, z = 0.0, *r* = .02), B1_1 and B1_2 (*p* = .46, z = 0.74, *r* = .07), B1_1 and B2+ (*p* = .86, z = 0.17, *r* = .02), and B1_2 and B2+ (*p* = .24, z = 1.17, *r* = .12) were not significantly different. 6.16 summarizes the results of the changes in the C/Tx.

Figure 6.10

Changes in C/Tx Score

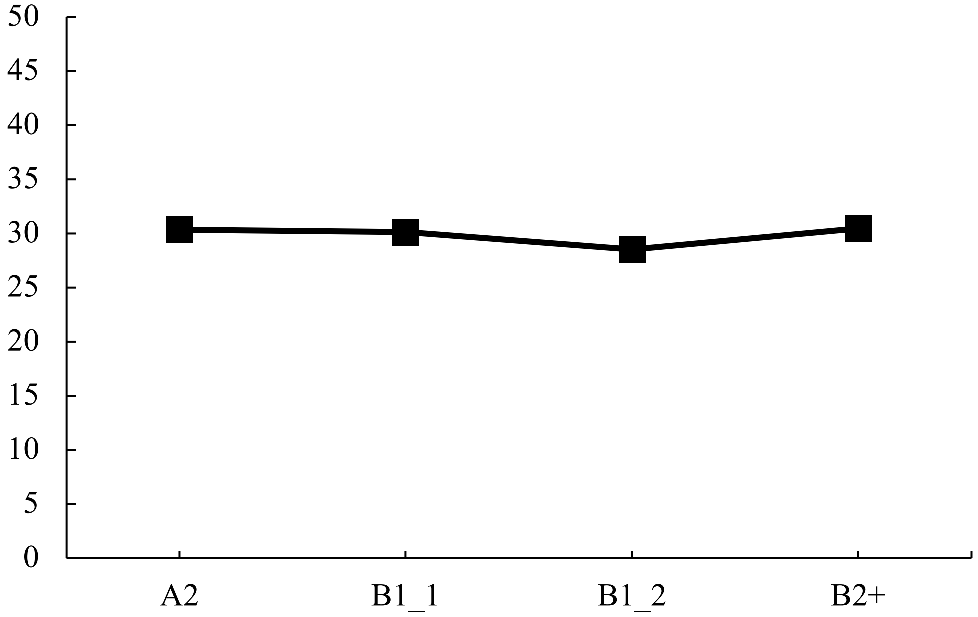


Table 6.16

Summary of the Results of K-W Test in C/Tx

Comparison	*r* (*z*-value)
A2 vs. B1_1	.01 (0.16)
A2 vs. B1_2	.15 (1.51)
A2 vs. B2+	.02 (0.0)
B1_1 vs. B1_2	.07 (0.74)
B1_1 vs. B2+	.02 (0.17)
B1_2 vs. B2+	.12 (1.17)

Figure 6.11 shows the changes in the T/Tx scores. The K–W test revealed that the T/Tx scores were not different among the four English levels, *H*(3) = 2.36, *p* = .50, z =

0.67, *r* = .07. In addition, the T/Tx scores between A2 and B1_1 (*p* = .99, z = 0.01, *r* = .00), A2 and B1_2 (*p* = .78, z = 0.28, *r* = .03), A2 and B2+ (*p* = .48, z = 0.70, *r* = .07), B1_1 and B1_2 (*p* = .94, z = 0.08, *r* = .01), B1_1 and B2+ (*p* = .79, z = 0.27, *r* = .03), and B1_2 and B2+ (*p* = .90, z = 0.13, *r* = .01) were not significantly different. Table 6.17 summarizes the results of the changes in the T/Tx.

Figure 6.11

Changes in T/Tx Score

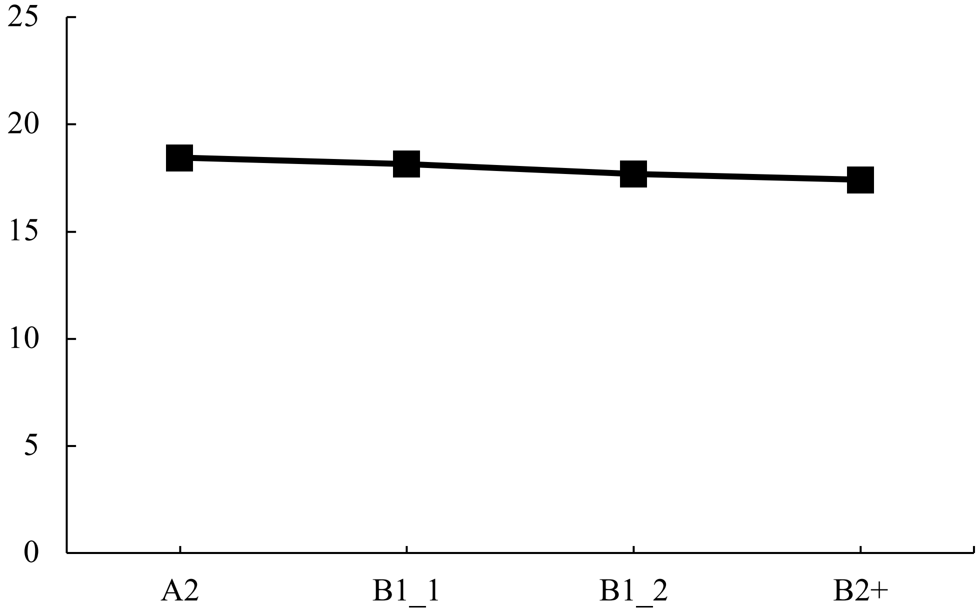


Table 6.17

Summary of the Results of K-W Test in T/Tx

Comparison	*r* (*z*-value)
A2 vs. B1_1	.00 (0.01)
A2 vs. B1_2	.03 (0.28)
A2 vs. B2+	.07 (0.70)
B1_1 vs. B1_2	.01 (0.08)
B1_1 vs. B2+	.03 (0.27)
B1_2 vs. B2+	.01 (0.13)

Note. $p < .05$*, $p < .01$**, $p < .001$***.

6.1.4 Discussion

RQ4-1: Do complexity and fluency measures change in accordance with the changes in EFCR scores? (See p.162)

Considering RQ 4-1, the present study aimed to demonstrate how complexity and fluency measures change as the EFCR scores change. The investigations that examined the development patterns of the written accuracy are critical for understanding not only what learners can and cannot do but also for demonstrating how teachers should effectively allocate their time.

Although written accuracy increases as the English proficiency level develops, some studies claimed that written accuracy should be interpreted with complexity and fluency measures (e.g., Michel, 2017). One possible reason is that written accuracy development may be masked by other domains such as complexity (Bulté & Housen, 2014). For instance, students increased their mean length of clauses. This change makes English learners susceptible to an increase in errors in clauses. Polio and Shea (2014) also suggested that if complexity is developing and accuracy is not decreasing, it may indicate the development of written accuracy.

Initially, the present study investigated the development of written accuracy using the EFCR, which has been used in past studies. The analysis suggested that there were both significant and insignificant differences. More specifically, the results showed that the EFCR score was significantly different between A2 and B2+, B1_1 and B2+, and B1_2 and B2+ levels.

Furthermore, although there was no difference in the phrasal complexity measures (e.g., VP_T), the syntactic complexity measures (e.g., MLC) were significantly different among some parts in the English levels. The analysis showed that the MLC score differed between A2 and B1_2, and B1_1 and B1_2 levels. Finally, the fluency measures (e.g., W/Tx) were examined, and the analysis showed that the W/Tx score differed between A2 and B1_2, B1_1 and B1_2, and B1_1 and B2+ levels.

Based on the results, it can be argued that written accuracy develops as English proficiency improves, implying that the linguistic errors would decrease, whereas the number of error-free clauses would increase. Some studies have reported that linguistic errors such as lexical choices and morphosyntactic errors decreased in accordance with English proficiency development (e.g., Abe, 2019; Thewissen, 2013).

Thewissen (2013) examined the L2 accuracy development patterns using error-tagged EFL learner corpus. She focused on 46 error categories (e.g., article errors and verb choice errors) and investigated how these errors changed as the CEFR levels developed. The results indicated that 30 error categories decreased as the CEFR levels increased. In addition, Abe (2019) investigated the linguistic errors in the writing performance made by Japanese EFL learners and showed that the number of linguistic errors such as subject-verb agreement, articles, and lexical errors decreased across

English proficiency. The high-proficiency participants in the present study showed a decrease in the number of linguistic errors. As the EFCR in the present study scores increased (although some parts did not change statistically), the number of linguistic errors decreased as the CEFR levels increased.

However, it should be noted that the development of the written accuracy might have been influenced by complexity and fluency while the EFCR scores increased. The results indicated that the EFCR scores were not statistically different in these CEFR sections, where the complexity and fluency measures differed. More specifically, the results suggested that the MLC scores differed between A2 and B1_2 and B1_1 and B1_2. In addition, the W/Tx as the fluency measure also differed between A2 and B1_2, B1_1 and B1_2, and B1_1 and B2+ levels. In particular, the MLC and W/Tx scores significantly increased between A2 and B1_2 levels, whereas the EFCR score was not significantly different. The MLC is the number of clauses per clause, and W/Tx is the number of words per text. As the clausal complexity (e.g., C/S) also did not change, it could be argued that the phrasal complexity (e.g., prepositions per nominal) might increase.

As for this, Crossely and McNamara (2014) showed an example and reported that high-quality essays tend to have longer noun phrases. According to their example, in the sentence *The boy eats the pepperoni pizza under the tree*, the noun phrases are *The boy* and *the pepperoni pizza*. As the essay quality increases, the noun phrases tend to be longer. Therefore, while it can be argued that the written accuracy would increase according to the CEFR levels, the development might be alterable because of the other domains.

RQ4-2: Do complexity and fluency measures change as the WCR scores change? If so, is there any difference in relationships with complexity and fluency measures between the EFCR and WCR? (See p.162)

The present study used not only the EFCR, which has been used in the past studies but also the WCR as the new written accuracy measure because the WCR could detect small changes in written accuracy and development. The results in RQ4-1 revealed that accuracy measuring EFCR may develop as English proficiency increases. However, because of the influence of complexity and fluency, the results indicated that the accuracy development may be masked. The possible reason is that the number of errors increased because Japanese EFL learners in the present study produced a more complex and fluent performance. However, some studies have agreed that the WCR considers the error gravity and can capture small changes in written accuracy (Evans et al., 2014; Foster & Wigglesworth, 2016). Therefore, the WCR can provide the researchers with insightful knowledge about written accuracy development.

The results showed that the WCR score significantly differed between not only A2 and B2+, B1_1 and B2+, and B1_2 and B2+ levels but also A2 and B1_2 levels. The difference between A2 and B1_2 levels was not revealed when using the EFCR, implying that the WCR captured the detailed changes which the EFCR could not. These results correspond to the claims of the past studies (Evans et al., 2014; Foster & Wigglesworth, 2016).

However, the results of the WCR scores in the present study do not completely correspond to the previous studies (Barrot & Agdeppa, 2021). Barrot and Agdeppa included 5,236 essays in the ICNALE corpus and used the WCR as a written accuracy measure. The results showed that the differences between A2 (M = 0.84, SD = 0.12) and B1_1 (M = 0.86, SD = 0.07) were not significantly different (p = .05), which is similar with the present study. However, Barrot and Agdeppas' study showed that the WCR scores between B1_1 and B1_2 (*M* = 0.88, *SD* = 0.09) were significantly different (*p* < .001).

The possible reasons are the influences of internal factors such as the learners' backgrounds. Barrot and Gabinete (2019) used the CAF scores in the writing performance in the ICNALE corpus and compared the scores between EFL learners (e.g., China and Japan) and ESL learners (e.g., Hong Kong and the Philippines). The findings indicated that the written accuracy scores (EFT/T and EFCR) were influenced by the L1 background. As the mean score of the B1_1 and B1_2 in the present study was lower than the Barrot, and Agdeppas' study showed, the Japanese EFL learners in B1_1 and B1_2 might have produced more serious errors such as Lv.2 and Lv.3 categories.

Moreover, according to the results, the WCR can provide new knowledge about relationships with complexity and fluency. It is a fact that the number of clauses having serious errors (e.g., Lv.2 category) may decrease as the English proficiency levels develop. In particular, the MLC and W/Tx significantly increased between A2 and B1_2 levels, but the WCR score increased in the section. This implies that learners in B1_2 groups obtained a high score because they tended to produce more local errors such as subject-verb agreement errors. Learners can obtain scores even if they produce linguistic errors. Therefore, the WCR score significantly increased despite an increase in complexity and fluency in the writing performance.

6.1.5 Conclusion

Study 4 aimed to investigate how written accuracy develops as the CEFR levels develop. In RQ4-1, the present study aimed to show how complexity and fluency measures change as the EFCR scores change. The analysis indicated that there were both significant and insignificant differences. The results showed that the EFCR score was significantly different between A2 and B2+, B1_1 and B2+, and B1_2 and B2+ levels. In addition, the analysis showed that the MLC score and W/Tx score differed (e.g., between A2 and B1_2). Therefore, while it can be argued that written accuracy increases in accordance with the CEFR levels, the development may be varied because of the other domains.

In RQ4-2, the present study examined the written accuracy development using the WCR not only to compare the EFCR but also to provide new knowledge about the written accuracy development. The results showed that the WCR score significantly differed between not only A2 and B2+, B1_1 and B2+, and B1_2 and B2+ levels but also A2 and B1_2 levels. Although the MLC and W/Tx significantly increased between A2 and B1_2 levels, the WCR score increased in the section, indicating that learners in B1_2 groups obtained the high score because they tended to produce more local errors such as subject-verb agreement errors.

While the present study showed the small changes in the written accuracy using the whole WCR score, the changes in the categories of clauses in the WCR (e.g., Lv.1) were not investigated. It can be argued that the categories in the WCR decrease as the CEFR levels develop. Therefore, Study 5 aimed to examine how the number of categories would change.

Chapter 7

Study 5: Examining the Development Patterns of Categories in the WCR Rating Scale

7.1 Overview of Study 5

7.1.1 The Summary of Study 4

According to Study 4, the WCR scores tended to increase gradually, although there were no significant differences in some sections (e.g., A2 and B2+). However, Study 4 focused on the changes of the whole WCR scores and did not examine how clause categories in the WCR rating scale (e.g., Lv.1 and Lv.2) changed as the CEFR levels increased.

In addition, Study 4 indicated that the syntactic complexity (e.g., MLC) significantly increased as the CEFR levels developed (e.g., A2 and B1_2). Therefore, Study 4 concluded that some clause categories in the WCR would increase which would lead to a decrease in the WCR score.

7.1.2 Purposes and Research Questions

Study 5 aimed to examine how the clause categories in the WCR rating scale would change as the CEFR levels increased. This point has also not been investigated in the previous studies (e.g., Barrot & Agdeppa, 2021); hence, the present study will provide a deeper understanding of the written accuracy development.

In addition, Study 5 aimed to discuss the written accuracy development with the complexity and fluency changes obtained in Study 4. Study 5 sought to show the changes of the number of clause categories. In sum, this study addressed the following research question (RQ):

RQ5: Are there differences in the number of accurate Lv.1, Lv.2, and Lv.3 clauses in the WCR among CEFR levels?

7.2.1 Method

7.2.1.1 Participants

The present study used data used in Study 1, which is extracted from the International Corpus Network of Asian Learners of English developed by Ishikawa (2013). The 100 participants (44 females and 56 males, average age = 18.84 years) were majoring in various fields, including business, engineering, and economics. The participants' essays on two topics were analyzed. Note that the essay topics were (a) *It is important for college students to have a part-time job* (PTJ) and (b) *Smoking should be completely banned at all the restaurants in the country* (SMK). The average lengths of the PTJ and SMK essays were 223 words (*SD* = 24.1) and 219 words (*SD* = 26.1), respectively.

7.2.1.2 WCR As a Written Accuracy Measure

The WCR rating scale consists of four categories, definitions, and scores (Table 7.1). After all sentences in an essay were divided into clauses, each clause was categorized by its gravity of error according to the definitions. It should be noted that 0 should not be awarded because even Level 3 clauses are linguistically accurate to a certain degree (Foster & Wigglesworth, 2016).

Table 7.1

Rating Scale of a Weighted Clause Ratio

Category	Definition	Score
No error	The clause is accurately constructed.	1.0
Level 1	The clause has only minor errors (e.g., morphosyntax) that do not compromise meaning.	0.8
Level 2	The clause contains serious errors (e.g., verb tense, word choice, or word order), but the meaning is recoverable, though not always obvious.	0.5
Level 3	The clause has very serious errors that make the intended meaning far from obvious and only partly recoverable.	0.1

7.2.1.3 Scoring

Weighted clause ratio

Study 5 used WCR as an accuracy measure, which was obtained in Study 1. Before the scoring, the researcher divided each sentence in the essay into clauses which was checked by the raters. Subsequently, raters independently evaluated all essays using the final version of the rating scale. Following the same procedure used during the rater training, raters were required to find errors in each clause and score the severity of errors according to the extent to which the error affects readers' comprehension.

It should be noted that the same errors (e.g., word errors) can be categorized under different levels because the severity of these errors was often contextual. The raters had agreed to read the definitions of the rating scale upon finding such errors to categorize them. Moreover, when there were multiple errors (e.g., Level 1 and Level 3 errors) in the same clause, the clause was categorized according to its worst-level error, as suggested by Foster and Wigglesworth (2016). The final WCR score was calculated as follows: WCR = the number of accurate clauses × 1.0 + the number of Lv.1 clauses × 0.8 + the number of Lv.2 clauses × 0.5 + the number of Lv.3 clauses × 0.1 / all clauses in the essay.

7.2.1.4 Data Analysis

As for RQ5, the present study compared the WCR scores among the four proficiency groups, and the normality of the CAF measures was not confirmed. Therefore, this study conducted a non-parametric statistical analysis: the Kruskal-Wallis test (K–W test). To interpret the effect size *r*, based on Field et al. (2013), the present study set .10 as a small effect size, .30 as a medium effect size, and .50 as a large effect size.

In addition, Steel-Dwass's multiple comparison test was used to compare the WCR scores in each group. While the Bonferroni method can be used, the method requires two assumptions: the normality and homogeneity of variance. Moreover, the Bonferroni method can cause a Type-1 error, which is the probability of rejecting the null hypothesis even though it is true because the *p*-value is adjusted according to the number of the comparison. In contrast, the Steel-Dwass's multiple comparison test can be used when the data do not have the normality and homogeneity of variance.

7.2.3 Results

Table 7.2 shows the descriptive statistics of the number of mean words in two essays. The number of words tends to increase according to the development of the CEFR levels in the INCALE corpus. In addition, Table 7.3 presents the descriptive statistics of the WCR score in each CEFR level.

Table 7.2

Descriptive Statistics of the Number of Words

Levels	No. of essays	No. of words			
		M	*SD*	*Min*	*Max*
A2	25	213.2	11.7	196	252
B1_1	25	212.8	21.8	182	303
B1_2	32	227.7	22.6	185	291
B2+	18	234.1	29.6	196	294

Table 7.3 shows that the WCR score increased as the CEFR levels developed. As shown in Study 4, the Steel-Dwass's multiple comparison revealed that the scores between A2 and B1_1 did not differ significantly, $p = .20$, $z = 1.29$, $r = .12$. In addition, the scores between B1_1 and B1_2 did also not differ significantly, $p = .92$, $z = 0.09$, $r = .00$. However, there were significant differences between A2 and B1_2 ($p = .003$, $z = 2.24$, $r = .22$), A2 and B2+ ($p <.001$, $z = 3.29$, $r = .34$), B1_1 and B2+ ($p = .003$, $z = 2.20$, $r = .22$), and B1_2 and B2+ ($p = .002$, $z = 2.28$, $r = .23$).

Table 7.3

Descriptive Statistics for the WCR Score in Each CEFR Level

Levels	*M*	*SD*	*Min*	*Max*
A2	0.81	0.06	0.62	0.89
B1_1	0.84	0.05	0.73	0.94
B1_2	0.85	0.03	0.77	0.90
B2+	0.88	0.05	0.74	0.95

Table 7.4 and Figures (7.1–7.4) shows the descriptive statistics of the changes in the number of clauses in the WCR rating scale. *No error* means the accurate clause in writing performances. Lv.1 indicates the clauses having local errors such as subject-verb agreements. Lv.2 means the clauses have serious errors (e.g., verb tense), which are difficult to understand sometimes. Finally, Lv.3 indicates the clauses containing serious errors that make the intended meaning far from obvious. The results indicated different trends comparing the three CEFR levels (i.e., A2, B1_1, and B1_2). The number of accurate clauses is the highest in the four categories in the WCR and in the four CEFR levels. The number of clauses having errors then decreased according to the gravity of errors.

Table 7.4

Descriptive Statistics of WCR

Category	A2				B1_1				B1_2				B2+			
	M	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*
No error	10.6	3.1	4.25	17	12.3	4.2	3.6	19.9	11.9	3.6	4.5	18	15.6	4.0	8.4	24.0
Lv.1	15.6	2.7	10.6	21.6	15.1	3.4	7.1	23.8	14.2	2.5	10.8	19.5	12.9	3.6	5.3	20.1
Lv.2	2.5	1.3	0.8	5.9	1.75	1.2	0.1	3.9	1.8	0.9	0.5	4.1	1.5	1.1	0.3	4.9
Lv.3	1.6	1.6	0.1	7.0	1.01	0.7	0	2.5	0.6	0.4	0.1	2.0	0.5	0.8	0	3.6

Figure 7.1

Changes in the A2 Level Group

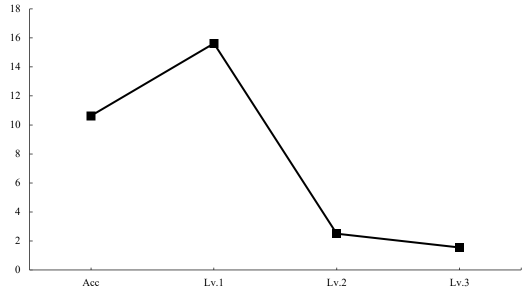


Figure 7.2

Changes in the B1_1 Level Group

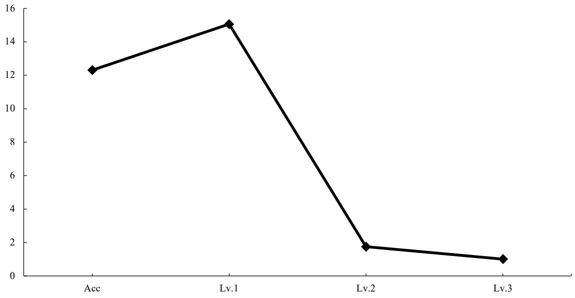


Figure 7.3

Changes in the B1_2 Level Group

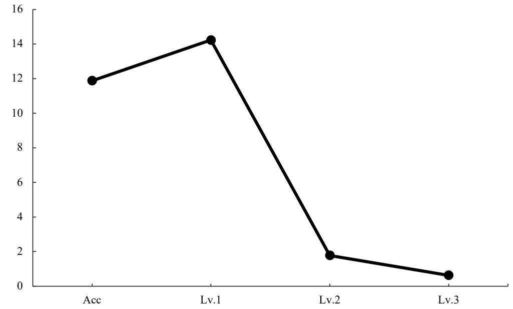


Figure 7.4

Changes in the B2+ Level Group

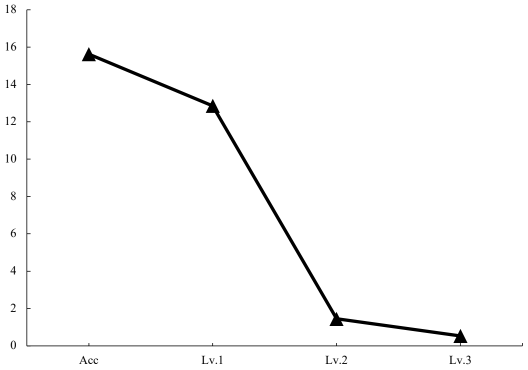


Figure 7.1 illustrates the changes in the number of the accurate clauses (i.e. Acc) in the writing performances. The K–W test revealed that the number of accurate clauses were different among the four English levels, $H(3) = 14.78, p = .002, z = 3.09, r = .31$. In addition, the number of the accurate clauses between A2 and B2+ ($p < .001, z = 3.43, r = .34$) and B1_2 and B2+ ($p = .02, z = 2.37, r = .24$) was significantly different.

In contrast, the number of accurate clauses between A2 and B1_1 ($p = .49, z = 0.70, r = .07$), A2 and B1_2 ($p = .60, z = 0.52, r = .05$), B1_1 and B1_2 ($p = .98, z = 0.02, r = .00$), B1_1 and B2+ ($p = .90, z = 1.70, r = .17$) were not significantly different. Table

7.5 summarizes the results of the changes in the number of the accurate clauses.

Table 7.5

Summary of the Results of the K-W Test for the Accurate Clauses

Comparison	*r* (*z*-value)
A2 vs. B1_1	.07 (0.70)
A2 vs. B1_2	.05 (0.52)
A2 vs. B2+	.34 (3.43)***
B1_1 vs. B1_2	.00 (0.02)
B1_1 vs. B2+	.17 (1.70)
B1_2 vs. B2+	.24 (2.37)**

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

Figure 7.2 illustrates the changes in the number of the Lv.1 clauses in the writing performances. The K–W test revealed that the number of the Lv.1 clauses was different among the four English levels, $H(3) = 9.64$, $p = .02$, $z = 2.29$, $r = .23$. In addition, the number of the Lv.1 clauses between A2 and B2+ ($p = .04$, $z = 2.06$, $r = .21$) was significantly different.

However, the number of the Lv.1 clause between A2 and B1_1 ($p = .65$, $z = 0.45$, $r = .04$), A2 and B1_2 ($p = .19$, $z = 1.31$, $r = .13$), B1_1 and B1_2 ($p = .52$, $z = 0.65$, $r = .07$), B1_1 and B2+ ($p = .19$, $z = 1.28$, $r = .13$), and B1_2 and B2+ ($p = .53$, $z = 0.63$, $r = .06$) was not significantly different. Table 7.6 summarizes the results of the changes in the Lv.1 clauses.

Table 7.6

Summary of the Results of K-W Test for the Lv.1 Clauses

Comparison	*r* (*z*-value)
A2 vs. B1_1	.04 (0.45)
A2 vs. B1_2	.13 (1.31)
A2 vs. B2+	.21 (2.06)*
B1_1 vs. B1_2	.07 (0.65)
B1_1 vs. B2+	.13 (1.28)
B1_2 vs. B2+	.06 (0.63)

Note. $p < .05$*.

Figure 7.3 illustrates the changes in the number of the Lv.2 clauses in the writing performances. The K–W test revealed that the number of the Lv.2 clauses was different among the four English levels, $H(3) = 11.54$, $p = .01$, $z = 2.61$, $r = .26$. In addition, the number of the Lv.2 clauses between A2 and B2+ ($p = .004$, $z = 2.89$, $r = .29$) was significantly different.

However, between A2 and B1_1 ($p = .26$, $z = 1.14$, $r = .11$), A2 and B1_2 ($p = .06$, $z = 1.85$, $r = .19$), B1_1 and B1_2 ($p = 1.00$, $z = 0.00$, $r = .00$), B1_1 and B2+ ($p = .85$, $z = 0.19$, $r = .02$), and B1_2 and B2+ ($p = .29$, $z = 1.06$, $r = .12$), it was not significantly different. Table 7.7 summarizes the results of the changes in the Lv.2 clauses.

Table 7.7

Summary of the Results of K-W Test for the Lv.2 Clauses

Comparison	*r* (*z*-value)
A2 vs. B1_1	.11 (1.14)
A2 vs. B1_2	.19 (1.85)
A2 vs. B2+	.29 (2.89)**
B1_1 vs. B1_2	.00 (0.00)
B1_1 vs. B2+	.02 (0.19)
B1_2 vs. B2+	.12 (1.06)

Note. $p < .05$*, $p < .01$**.

Figure 7.4 illustrates the changes in the number of the Lv.3 clauses in the writing performances. The K–W test revealed that the number of the Lv.3 clauses was different among the four English levels, $H(3) = 19.30$, $p < .001$, $z = 3.68$, $r = .37$. In addition, the number of the Lv.3 clauses between A2 and B1_2 ($p = .007$, $z = 2.71$, $r = .27$), A2 and B2+ ($p < .001$, $z = 3.72$, $r = .37$), and B1_1 and B2+ ($p = .02$, $z = 2.32$, $r = .23$), was significantly different.

However, between A2 and B1_1 ($p = .79$, $z = 0.27$, $r = .03$), B1_1 and B1_2 ($p = .16$, $z = 1.42$, $r = .14$), and B1_2 and B2+ ($p = .27$, $z = 1.12$, $r = .11$), it was not significantly different. Table 7.8 summarizes the results of the changes in the Lv.3 clauses.

Table 7.8

Summary of the Results of K-W Test for the Lv.3 Clauses

Comparison	*r* (*z*-value)
A2 vs. B1_1	.03 (0.27)
A2 vs. B1_2	.27 (2.71)*
A2 vs. B2+	.37 (3.72)***
B1_1 vs. B1_2	.14 (1.12)
B1_1 vs. B2+	.23 (2.32)**
B1_2 vs. B2+	.11 (1.12)

Note. $p < .05$*, $p < .01$**, $p < .001$***.

7.2.4 Discussion

RQ5: Are there differences in the number of each category in the WCR among each CEFR level? (See p.194)

In RQ5, the present study aimed to examine the development patterns of each category in the WCR rating scale. The WCR score is calculated using the categories; hence, it is possible to examine which categories not only decrease but also lead to an increase in the WCR score.

The present study used the K-W test and showed that each category tended to decrease at some sections in the CEFR levels. As for the accurate clauses, A2–B2+ and B1_2–B2+ groups differed significantly. In the Lv.1 clauses, A2–B2+ groups were significantly different. In addition, the Lv.2 categories differed between A2–B2+ groups, whereas the Lv.3 categories were different between A2–B1_2, A2–B2+, and B1_1–B2+ groups. The participants in the A2 groups tended to produce more inaccurate clauses than participants in the other groups, especially B2+ groups. It should be natural to think that

English learners in the beginner levels sometimes produce numerous linguistic errors (e.g., Abe, 2017; Thewissen, 2013). Given that the number of all categories in the WCR rating scale tended to decrease as the CEFR levels developed, Japanese EFL learners were able to develop their linguistic knowledge and decrease the number of linguistic errors in accordance with the development of English proficiency.

However, interesting patterns were identified in the present study. The participants in the B1_2 group tended to produce inaccurate clauses more than those in B1_1, which is the lower-proficiency group although there was not significantly difference. As a result, the number of the accurate clause between B1_2 and B2+ groups statistically differed. One possible reason is an increase in complexity in the writing performance in the B1_2 levels. Barrot and Agdeppa (2021) reported that the MLC score increased as the CEFR levels developed. Given that Study 4 also indicated that the MLC scores between A2 and B1_2, and B1_1 and B1_2, the participants in the B1_2 groups produced more complex performances and inaccurate clauses than those in A2 and B1_1 group.

This trend is shown in the number of the Lv.1 and Lv.2 category. The results of the present study indicated that the Lv.1 and Lv.2 categories differed only between A2–B2+ group, implying that the participants in the B1_2 group produced as many linguistic errors as the A2 and B1_1 group. In the Lv.2 category, the participants in the B1_2 group seemed to produce more errors than the B1_1 group (although there was no significant difference). While the trend indicating that the learners having more linguistic knowledge than lower-proficiency groups tend to produce many linguistic errors may correspond to the *U-shaped pattern* proposed in the second language acquisition (Gass et al., 2020), the longitudinal studies investigating the development patterns of accuracy are still needed.

Furthermore, the present study obtained an interesting suggestion about the development patterns of the WCR and the categories. It is a fact that the number of the

Lv.3 category leads to the difference between A2 and B1_2 levels. While Study 4 indicated the significant differences in WCR score between A2 and B1_2 groups ($p = .003$, $z = 2.24$, $r = .22$), the present study indicated that the number of categories in accurate clauses, Lv.1, and Lv.2 clauses, did not differ between them. The only statistical difference between A2 and B1_2 was the number of the Lv.3 category.

This result indicated that the participants in the B1_2 group tended not to produce very serious errors influencing the readers' understanding while still producing a variety of linguistic errors. Kudo (2009) used writing performance in a standardized test (GTEC for STUDENTS) developed by Benesse Corporation and examined how many global and local errors Japanese EFL participants produced. Kudo's study showed that the lower-proficiency participants tended to produce more global errors than high-proficiency participants, while all proficiency levels produced global errors. Although writing performances were obtained from the standardized test and may be different from the performance in the present study, a similar tendency was identified.

7.2.5 Conclusion of Study 5

The present study aimed to examine the development patterns of each category in the WCR rating scale. The results indicated that each category tended to decrease at some sections in the CEFR levels. While the whole WCR scores have been used to investigate the written accuracy developments (e.g., Barrot & Adgeppa, 2021), the investigation focusing on the clause categories in the WCR rating scale has not been conducted. Therefore, the present study can provide researchers and EFL teachers with a deep understanding of accuracy development.

In particular, the analysis showed that the number of Lv. 3 clauses significantly differed between the A2 and B2+, A2 and B1_2, and B1_1 and B2+ groups, suggesting that the number of clauses which are completely incomprehensible decreases as the English proficiency levels increase because these learners possess more sophisticated L2 knowledge and interlanguage skills (Housen et al., 2012). Consequently, they tend not to produce serious errors. These results can explain the increase of the WCR scores between the A2 and B1_2, and B1_1 and B2+ groups.

Chapter 8

General Discussion

8.1 Overview of Findings and General Discussion

The present dissertation outlined two main studies relating to the WCR: validation and written accuracy development. The validation study consisted of three studies (Studies 1–3) to confirm the validity of the WCR in the context of learner corpus studies. Then, further investigation was conducted (Study 4–5), which examined the development patterns of written accuracy in Japanese EFL learners. The validation study was a base on which the accuracy development study was conducted. In this chapter, the findings are summarized and discussed from theoretical and experimental perspectives.

Study 1 investigated the scoring and generalization inferences of the WCR. These inferences are strongly related to assessment reliability and the fundamental requirements to progress further studies such as Study 2. Study 1 sets three research questions (RQs):

RQ 1-1: If raters assess written accuracy using accuracy measures, to what extent could inter-rater reliability be obtained?

RQ 1-2: If raters assess written accuracy using accuracy measures, to what extent could the score variances be explained by the factors?

RQ 1-3: If raters assess written accuracy using accuracy measures, what is the degree of reliability (G coefficient) obtained?

To answer RQ 1-1, the Study 1 used and adjusted Cronbach's α. Generalizability theory (G theory) was used to answer RQ 1-2 and 1-3.

The analysis for RQ1-1 showed that the Cronbach's α coefficient of the WCR was high in two texts with different topics (PTJ: α = 0.91, SMK: α = 0.93). This result suggests that the scores from the four raters were similar and consistent. In addition, for other accuracy measures, such as EFCR and EFTR, the Cronbach's α coefficient was over 0.8. This result indicates that the WCR assessment is as reliable as traditional accuracy measures. Moreover, the adjusted Cronbach's α coefficient of both the WCR and traditional written accuracy measures were also over 0.8, suggesting that the raters themselves did not have an influence on the consistency of the accuracy assessment.

The tendency highlighted by this study corresponds to that found in previous studies that have investigated the reliability of the WCR. For example, Evans et al. (2014) examined the reliability of the three accuracy measures (WCR, EFCR, and EFTR) using the multi-faceted Rasch model, and reported that the rater severity of the WCR was quite similar to the two written accuracy measures. Additionally, Polio and Shea (2014) reported on the inter-rater reliability using Pearson's *r*, although they used weighted error-free T-units instead of the WCR; however, these were similar to the WCR. Their results yielded a Pearson's *r* of 0.84 and they concluded that the reliability was similar to the other measures, such as EFCR and holistic rating, for language use. Given these results from previous studies, the results from Study 1 appear reasonable and consistent.

While these previous studies have investigated the reliability of the WCR, it is worth noting that the reliability scores used were not original WCR scores. Evans et al.'s study transformed the WCR scores into the whole numbers (0–10), because the multi-faceted Rasch model cannot analyze decimal values, and Polio and Shea's study did not use the WCR. Because the scores used in subsequent experimental studies would be the original WCR scores with decimal values (e.g., Suzuki et al., 2021), the evidence from the aforementioned studies would be insufficient for writing and accuracy development studies, even though they provided evidence of the high reliability of the WCR. On the other hand, the present study, Study 1, demonstrated the high reliability of the WCR while using the original scores. Hence, researchers could reliably evaluate written accuracy using our WCR scores.

Study 1 used G theory to address RQ 1-2; G theory can determine not only variances derived from a person's ability, but also the degree of variation due to measurement errors (e.g., rater error). The results highlighted the influential errors in the WCR scores. The degree of person-factor (*p*) in the WCR was the largest among the eight written accuracy measures (38%), suggesting that the WCR can reflect the degree of Japanese EFL learners' accurate writing ability. Raters have to judge the degree of error gravity in every clause in a text when they use the WCR; by considering the error gravity, the WCR could capture small differences in accuracy better than traditional written accuracy measures.

In addition, the degree of the person-factor in the WCR would be better than analytic ratings. Schoonen (2005) investigated and compared the generalizability of writing scores between holistic and analytic ratings. Schoonen reported on the degree of the measurement errors in the *language use*, which focused on language errors, and results yielded a person-factor value of 29.4%. In addition, Barkaoui (2007) investigated how holistic and multi-trait scales affected L2 writing scores. Barkaoui found that the degree of the person-factor for *grammar* was 9.5%. Compared to the analytic or multi-traits scales, the WCR better reflects a participants' ability.

A possible reason for the above findings is that the WCR requires raters to evaluate each clause in the writing performance. In analytical ratings, raters usually read the whole text and give one point in the text; raters do not pay attention to every detail. Therefore, this assessment method provides only a rough measurement of accuracy. On the other hand, raters using the WCR have to judge the accuracy of every clause in a text; therefore, the WCR can evaluate parts that the analytical evaluation misses. A similar tendency can be inferred from traditional written accuracy measures, which require raters to judge the accuracy of every clause.

An interesting result is that the degree of rater-factor (*r*) in the WCR was the greatest (18.3%) among the eight written accuracy measures, meaning that these ratings were most influenced by the raters' characteristics; the degree of rater-factor of the EFCR was 7.2%, approximately 2.5 times smaller than the WCR. This result is due to the accuracy assessment method of the WCR, which has three steps: (1) divide the written texts into clauses; (2) identify the errors in every clause; and (3) judge the gravity of the error. Steps 1 and 2 are the same procedures as in traditional written accuracy measures (e.g., EFCR). Step 3 differs from the traditional written accuracy methods and is the unique aspect of the WCR. In WCR accuracy assessment, raters are required to judge the gravity of the errors, which indicates the influence of the errors on readers' comprehension. When raters judge the error gravity and categorize the clause, referring to the rating scale of the WCR, subjectivity is included in the rating process. Although Study 1 anticipated the raters' subjectivity and, to minimize the impact of their subjectivity, detailed error types in each clause for each four raters, the influence of rater-factor was still present.

Experimental studies focusing on the judgment of the error gravity (e.g., Hyland & Anan, 2006; Rao & Li, 2017) play an important role in explaining this result. These studies focused on the errors in English texts and compared the error perceptions between native speakers of English and non-native speakers of English teachers. The main finding of these studies is that error perceptions are affected by raters' backgrounds. Hyland and Anan (2006) compared the error perceptions of native English speaking and Japanese speaking English teachers and educated native English speaking non-teachers. The results reported that Japanese speaking English teachers tended to rate the gravity of errors more severely than native English teachers. In addition, Japanese speaking English teachers identified the highest number of serious errors (66 errors) compared to both native speakers of English (41 errors) and educated native English speaking non-teachers (41 errors).

More recently, Rao and Li (2017) compared error perceptions between native and non-native English-speaking teachers in a Chinese context. They used 10 kinds of errors (e.g., subject/verb agreement, tense consistency, and pronoun use) and compared the perceptions of each error. The results showed that native English-speaking teachers tolerated errors more than non-native English-speaking teachers. Furthermore, native English-speaking teachers regarded subject/verb agreement errors as the most serious error, while non-native English-speaking teachers regarded tense errors as the most serious error. The ranking of the gravity of other errors also differed between native and non-native English-speaking teachers. Rao and Li suggested that the causes of these disparities are due to raters' backgrounds (e.g., cultural belief, educational background, teaching style, and English proficiency).

The four raters in our present study (Study 1) had different backgrounds; hence, these factors may have affected our score variances. However, the degree of influence of the rater-factor in the WCR was not critical compared to the analytic ratings. Barkaoui (2007) compared the measurement errors between holistic and multi-trait rating, and reported that the rater-factor for *grammar* was 65.2% and the second biggest factor in six dimensions (e.g., *organization*). The influence of the rater-factor is inevitable because raters are required to judge the error gravity; however, efforts should still be made to reduce the disparity (e.g., rater training and reconstructing the rating scale).

In addition, D-study indicated that the G coefficient of the WCR was 0.91; this is similar to that traditional written accuracy measures (e.g., EFCR: G = .90). This result suggests that WCR scores in the present study are generalizable. Because there were no studies investigating the G coefficient of the WCR, our finding is new and has the potential to expand the theoretical background.

The detailed rating scale was developed by the four raters, as the rating scale proposed by Foster and Wigglesworth (2016) would not promote consistent ratings. In particular, the kinds of errors in each clause (e.g., Lv.1) are not described in detail. As the rating scale is fundamental for reliable assessment, we revised and amended the scale (see Appendix). It is easy for raters to categorize clauses by referring to the revised rating scale.

The detailed rating scale of the WCR is helpful for the evaluation of accuracy. Evans et al. (2014) study, which investigated the validity of the WCR, originally used three raters; however, one rater was excluded because they were found to be less reliable. Hence, only two raters were used in their study; however, these raters were highly experienced and well trained, so they were sufficient to rate written accuracy using the WCR. Although rater training should be necessary for reliable evaluation, a high-quality rating scale is also necessary to obtain consistent scoring. Therefore, our revised rating scale is informative for reliable evaluation and future research.

The results of Study 1 detail the scoring and generalization inferences. Being able to reliably produce the score is crucial for inferring a learners' ability and progressing their further studies (e.g., Study 2). The data obtained in Study 1 were used in the subsequent studies.

Study 2 investigated the explanation inference, which identifies which factors the WCR reflects. In addition, Study 2 investigated how textural features, such as clauses, correlated with written accuracy measures. Study 2 set the three RQs, which were answered using exploratory factor analysis (EFA) and correlation analysis:

RQ2-1: To what extent does the WCR reflect the factor that traditional accuracy measures do?

RQ2-2: To what extent do extracted factors and measures correlate with each other?

RQ2-3: To what extent do the accuracy measures correlate with textual features in essays?

The EFA showed that the WCR reflects the same construct of accuracy as that of traditional written accuracy measures, suggesting that the WCR is a sufficient accuracy measure in L2 writing. This result corresponds to the ideas of Evans et al. (2014) and Foster and Wigglesworth (2016). Evans et al. claimed that the WCR measures written accuracy based on adequacy. Although some studies expected the WCR to measure complexity (e.g., Fox, 2019), this is not possible because the correlation between the constructs of complexity and accuracy were quite low. Barrot and Agdeppa (2021) investigated the correlations between CAF measures and reported that accuracy measures, including the WCR's, did not highly correlate with complexity or fluency measures.

However, it should be noted that Pallotti (2009) claimed that accuracy measures can reflect spurious constructs if errors are classified according to their gravity. Pallotti further argued that "a 100-word production with 10 errors not compromising communication is not more 'accurate' than a text of the same length with 10 errors hindering comprehension but merely more 'understandable' or 'communicative effective'" (Pallotti, 2009, p.3). Pallotti's example is similar to the assessment of accuracy with the WCR.

These differences could be due to how researchers see the construct of accuracy and errors. Pallotti (2009) described accuracy as "the degree of conformity to certain norms" (p.3); based on this definition, how much a sentence conforms to certain norm is important, and the degree of error gravity, which reflects the degree of readers' comprehension, is an unnecessary measure. On the other hand, Evans et al. (2014) defined accuracy as "the ability to be free from errors while using language to communicate" (Wolfe-Quintero et al., 2998, p.33), and expected communicative situations. Our present study (Study 2) defined accuracy as "the ability to produce target-like and error-free language" (Housen et al., 2012). Although the definition does not use the word 'communication', Housen et al. (2012) assumed that accuracy relates to interlanguage and L2 knowledge. This kind of knowledge (e.g., grammatical knowledge) would not only be linguistically correct, but also meaningful (Purpura, 2004; Spinner, 2016). Based on this definition, the construct of accuracy can capture the extent to which meaning is conveyed to the reader because the scope of accuracy includes not only linguistic features, but also meaning. Language tests usually evaluate performance for the purpose of communication; the CAF perspective evaluates whether it is appropriate for communication (Pallotti, 2016). The CAF measures should not be interpreted simply as an increase or decrease in complexity, but instead as whether or not a text is appropriate for communication.

Furthermore, Study 2 investigated how textural features correlated with the WCR (RQ 2-3). The correlation analysis showed that the correlation between the accuracy measures and textural features were low (0.18–0.26), suggesting that textural features do not affect score interpretations. For example, if written accuracy measures correlate with textural features, an increase of scores would not always indicate the development of a learners' accuracy. This result is reflected in the severe control of the experiment in the ICNALE corpus project. When gathering text productions from Asian English learners, participants were required to write texts from 200 to 300 words. The descriptive statistics of the number of words in our present study showed small SD variations (Min: SD = 11.7, Max: SD = 29.6). If the number of words was not limited, the correlation between measures and textural features would be stronger. As a result, the interpretation of these scores might be difficult. Based on these results, the explanation inference is confirmed positively.

Study 3 focused on the extrapolation inference and investigated how the WCR correlated with English proficiency (i.e., CEFR). This study used the CEFR levels from the ICNALE corpus (A1, B1_1, B1_2, and B2+). Moreover, Study 3 investigated the utilization inference, which is related to the use of test scores. Two research questions were set:

RQ 3-1: To what extent does the WCR correlate with CEFR levels?

RQ 3-2: Does the WCR provide more detailed information than the traditional accuracy measures in assessing the writing accuracy development?

Correlation analysis was used to answer RQ 3-1, and non-parametric analysis and descriptive statistics were used to answer RQ 3-2. Correlation analysis showed that the WCR correlated with English proficiency, with a correlation score of 0.33. This result indicates that the WCR scores, and hence the comprehensibility of the writing, tend to increase as English proficiency increases. This tendency is reasonable because high-proficient learners have more L2 knowledge than low-proficient learners. Accuracy is assumed to relate to the interlanguage system (Housen et al., 2012), and the closer a learners' interlanguage is to the target language, the more accurate their writing performance is. As a result, the number of errors which hinder a readers' comprehension decreases gradually, and the WCR scores increase.

Moreover, the correlation between the WCR and English proficiency was the strongest amongst the traditional accuracy measures; the correlation values of three other accuracy measures (EFCR, EFTR, and EFT/W) did not reach 0.30. When using traditional accuracy measures (e.g., EFCR), all clauses with errors are given zero points, even if the errors are minor. Even high-proficient learners tend to produce errors (e.g., subject/verb agreement) in their writing; therefore, traditional accuracy measures cannot effectively capture progress of written accuracy. In contrast, the accuracy assessment using the WCR provides clauses with scores according to the influence of the errors. Therefore, English texts from high-proficient learners would be given high scores, and English texts from low-proficient learners would be given low scores. Hence, the WCR correlates with English proficiency stronger than other written accuracy measures.

This result is informative for writing research that investigates how well CAF predicts English or writing proficiency. Kojima and Kaneda (2020) used a meta-analysis and investigated the correlation between CAF measures and writing proficiency. They showed that the correlation between the accuracy measures (e.g., EFCR) and writing proficiency was $r = 0.44$, and was stronger than syntactic complexity measures ($r = .15$). Based on the results, Kojima and Kaneda suggested that accuracy assessments that consider the error gravity correlate with writing proficiency stronger than those that count the number of errors. The rating scale of the WCR could be applied to assess the error gravity, although automated assessment would be difficult. Based on these results, Study 3 confirmed the extrapolation inference.

Another aim of Study 3 was to investigate whether or not the utilization inference was confirmed. Descriptive statistics and non-parametric tests were used to answer RQ 3-2; these methods can provide information about the degree of accuracy and accuracy development that can be determined by the WCR can. The descriptive statistics of the WCR can inform researchers on how much accuracy a learner has. Moreover, the non-parametric tests can indicate in how much detail the WCR can capture accuracy development. Confirming these points is important to promote the use of the WCR scores in corpus studies because the main purpose of corpus studies is to understand how ability and linguistic features develop in detail.

The descriptive statistics were conducted on the scores of four accuracy measures in each English proficiency level. For example, the WCR score in the A2 group was 0.81, with a standard deviation of 0.06. The rating scale of the WCR suggests that these scores indicate that the writing performance in the A2 group is comprehensible.

Using the WCR score enables us to identify whether or not the writing performance is not only linguistically accurate, but also easy to understand. On the other hand, the EFCR score in the A2 group was 0.35, meaning that 35% of clauses in the group's writing was linguistically accurate. The EFTR score in the A2 group was 0.21, meaning that 21% of all T-units in the group's writing performance was linguistically accurate. However, these measures only indicate whether or not learners can produce linguistically accurate writing, and not whether their writing is easy to understand.

Furthermore, the WCR can provide detailed information about clauses with errors. Table 7.3 shows the number of each clause (e.g., Lv.1) in each English proficiency level. For example, the writing task for the A2 group included an average of 15.6 clauses of Lv. 1, 2.5 clauses of Lv. 2, and 1.6 clauses of Lv. 3. By using the WCR, researchers can identify what levels of clauses appear within the text. In contrast, traditional accuracy measures only indicate that there are clauses with errors. In this regard, the WCR is useful for researchers who want to know detailed information regarding a learner's accuracy development.

The non-parametric test revealed that the WCR can better capture differences in accuracy among four English proficiency levels compared with traditional accuracy measures. The Kruskal-Wallis (KW) test showed that the WCR scores significantly differed in the four proficiency intervals (e.g., A2 to B1_2) for six comparisons. On the other hand, the EFCR, EFTR and EFT/W scores significantly differed in only three, two, and one proficiency intervals, respectively. Therefore, we can conclude that the WCR can capture small differences in accuracy better than other accuracy measures. The ease of interpretation and informative nature of WCR scores make the WCR a very useful accuracy measure for investigating learner accuracy development with a learner corpus. Based on these results, the utilization inference was confirmed positively.

The validation study outlined here confirmed all inferences in an argument-based approach (Chapelle et al., 2008). Therefore, we can conclude that the WCR score can be used for investigating accuracy development using a learner corpus. This validation study is fundamental because scores with low validity are not useful for research, and therefore lead to unreliable experimental results. In particular, Pallotti (2016) claimed that the discussion about the constructs and operational definition in CAF studies. The present study (Study 4) shed light on accuracy assessments and overcame this limitation.

Study 4 and Study 5 both focused on written accuracy development in Japanese EFL learners. Understanding accuracy development patterns is important when deciding appropriate instructions. By understanding the characteristics and differences of written accuracy in different proficiency groups, it is possible to identify the optimal areas to spend instructional time on. Therefore, English teachers will be able to provide instruction for the learner's specific accuracy and proficiency level.

Study 4 used the WCR and investigated development patterns of written accuracy in Japanese EFL learners. EFCR was used to compare the development patterns. Complexity and fluency measures were also used because these dimensions can mask the development of written accuracy. For example, Polio and Shea (2014) highlighted that complexity can mask changes in accuracy because learners with higher proficiencies tend to produce longer sentences, and therefore produce more errors. An increase in a complexity score without an increase in an accuracy score could still indicate development (Polio & Shea, 2014). Studies 4 and 5 were based on three research questions.

RQ 4-1: Do complexity and fluency measures change in accordance with the changes in EFCR scores?

RQ 4-2: Do complexity and fluency measures change as the WCR scores change? If so, is there any difference in relationships with complexity and fluency measures between the EFCR and WCR?

Before discussing RQ4-1, The differences in accuracy development between the WCR and EFCR will be discussed first. The KW test showed that the WCR scores significantly differed in four proficiency intervals in six comparisons (i.e., A2–B1_2, A2–B2+, B1_1–B2+, and B1_2–B2+). This suggests that the written accuracy of Japanese EFL learners tends to increase as English proficiency levels increase. As previously outlined, this tendency is reasonable because high-proficient learners would have more L2 knowledge than low-proficient learners. Accuracy is assumed to relate to the interlanguage system (Housen et al., 2012), and the closer the learners' interlanguage is to the target language, the more accurate the writing performance is. As a result, the number of errors that hinder a readers' comprehension gradually decreases, and the WCR scores increase.

However, the results of the WCR scores did not completely agree with previous studies (Barrot & Agdeppa, 2021). Barrot and Agdeppa studied 5236 essays from the ICNALE corpus and used the WCR to measure the accuracy of the writing. Their results showed a difference between A2 (*M* = 0.84, *SD* = 0.12) and B1_1 (*M* = 0.86, *SD* = 0.07) that was not significant (*p* = .05). This result is similar to the present study. However, Barrot and Agdeppas' study showed a difference between the WCR scores of B1_1 and B1_2 (*M* = 0.88, *SD* = 0.09), which was significant (*p* < .001).

Possible reasons for these results include influences of internal factors, such as the learners' backgrounds. Barrot and Gabinete (2019) used CAF scores to assess the writing performance of the ICNALE corpus and compared the scores between EFL learners (e.g., China and Japan) and ESL learners (e.g., Hong Kong and Philippines). Their findings suggested that written accuracy scores (EFT/T and EFCR) are influenced by the L1 background. The mean accuracy scores of the B1_1 and B1_2 groups in our study were lower than in Barrot, and Agdeppas' study; hence, Japanese EFL learners in the B1_1 and B1_2 groups might produce more serious errors, such as errors in the Lv. 2 or Lv. 3 categories.

Study 4 used EFCR and investigated accuracy development. The analysis showed that the EFCR scores significantly differed in three proficiency intervals in six comparisons (i.e., A2–B2+, B1_1–B2+, and B1_2–B2+). The same explanation as in the WCR case can be applied; as English proficiency increases, Japanese EFL learners store large amounts of L2 knowledge, and therefore the EFCR scores tend to increase.

Interestingly, there was a difference in the accuracy development patterns captured by the WCR and EFCR. The WCR indicated a significant difference between A2 and B1_2, however this was not indicated by the EFCR scores. As Evans et al. (2014) suggested, the WCR can capture small differences in accuracy because the WCR considers the gravity of each error. Because EFCR does not consider the gravity of errors, it therefore cannot capture possible differences in writing accuracy.

Study 5 investigated how complexity and fluency measures change as the EFCR scores change (RQs 4-1 and 4-2). These research questions can be informative for questions which Polio and Shea (2014) claimed. Initially, the present study investigated the development of written accuracy using the EFCR, which has been used in previous studies. The analysis showed significant differences; the EFCR score was significantly different between the A2 and B2+, B1_1 and B2+, and B1_2 and B2+ groups. Additionally, although there was no difference in the phrasal complexity measures (e.g., VP_T), the syntactic complexity measures (e.g., MLC) were significantly different between some proficiency levels; the analysis showed that the MLC score differed between the A2 and B1_2, and B1_1 and B1_2 groups. Finally, the fluency measures (e.g., W/Tx) were examined, and the results showed that the W/Tx score differed between the A2 and B1_2, B1_1 and B1_2, and B1_1 and B2+ groups.

These results suggest that written accuracy improves as English proficiency improves because the EFCR scores tend to increase as these two dimensions increase. This result also suggests that the number of linguistic errors and error-free clauses would decrease and increase, respectively. Some studies reported that the number of linguistic errors, such as lexical choices or morphosyntactic errors, decreased according to English proficiency development (e.g., Abe, 2019; Thewissen, 2013). Thewissen (2013) examined L2 accuracy development patterns using an error-tagged EFL learner corpus. She focused on 46 error categories (e.g., article errors and verb choice errors) and investigated how these errors changed as the CEFR levels developed. She found that the occurrence of 30 types of error decreased as the CEFR levels increased.

Additionally, Abe (2019) investigated linguistic errors in the writing of Japanese EFL learners and showed that the number of linguistic errors, such as subject-verb agreement, articles, and lexical errors, decreases as English proficiency increases. In our present study (Study 5), high-proficiency participants had fewer linguistic errors than low-proficiency participants. As the EFCR scores increased (although some parts did not change significantly), the number of linguistic errors decreased as the CEFR levels increased.

However, it should be noted that the development of written accuracy can be influenced by complexity and fluency, even when the EFCR scores increase. Our results indicated that the EFCR scores were not statistically different between CEFR levels, even though the complexity and fluency measures differed; however, there was one exception: the W/Tx score between the B1_1 and B2+ groups. More specifically, our results showed that the MLC scores differed between the A2 and B1_2, and B1_1 and B1_2 groups. In addition, the W/Tx (the fluency measure) also differed between the A2 and B1_2, B1_1 and B1_2, and B1_1 and B2+ groups. In particular, the MLC and W/Tx scores significantly increased from the A2 group to the B1_2 group, however the EFCR score was not significantly different. The MLC is the number of clauses per clause, and W/Tx is the number of words per text. As the clausal complexity (e.g., C/S) did not change, the phrasal complexity (e.g., prepositions per nominal) might have increased. Crossely and McNamara (2014) reported that high-quality essays tended to have longer noun phrases. For example, in the sentence *The boy eats the pepperoni pizza under the tree*, the noun phrases are *The boy* and *the pepperoni pizza*. Therefore, while it would be possible that written accuracy would increase with the CEFR levels, this development is actually changeable because of these other domains.

In summary, our study used not only the EFCR, but also the WCR to measure written accuracy because the WCR can detect small changes in written accuracy and development. The results from RQ6-1 and RQ6-2 indicated that accuracy increases as English proficiency increases. However, because of the influence of complexity and fluency, accuracy development might be masked. A possible reason for this increase in the number of errors is that the high-proficiency Japanese EFL learners in our study produced a more complex and fluent performance. However, some studies have also indicated that because the WCR considers error gravity, it can capture small changes in written accuracy (Evans et al., 2014; Foster & Wigglesworth, 2016). Therefore, we can conclude that the WCR can provide researchers with insightful knowledge about written accuracy development.

Our results showed a significant difference in the WCR scores between not only the A2 and B2+, B1_1 and B2+, and B1_2 and B2+ groups, but also the A2 and B1_2 group. A significant difference between the A2 and B1_2 groups was not found when using the EFCR, suggesting that the WCR was able to capture detailed changes that the EFCR could not. At first glance, the EFCR results do not suggest that accuracy has improved, however the WCR demonstrated that accuracy did, in fact, improve. These results correspond to claims made by previous studies (Evans et al., 2014; Foster & Wigglesworth, 2016).

Although Study 5 demonstrated that the WCR can capture differences in accuracy better than the EFCR, it did not investigate which types of clauses (e.g., Lv.1) decreased in number as the WCR scores increased. Moreover, the analysis indicated no significant difference in the WCR scores (or EFCR scores) between the A2 and B1_1, and B1_1 and B1_2 groups. This could be related to an increase or decrease in the number of different clause types. Therefore, in Study 6, we focused on the changes in the number of types of clauses in the WCR rating scale. Specifically, Study 6 investigated how the number of each clause type (e.g., Lv.1) changes as the WCR score increases; this study addressed the following research question:

RQ5: Are there differences in the number of each clause type in the WCR between each CEFR level?

We will now discuss the results of Study 6 in terms of the changes of the number of each clause type; these results are based on the results from Study 5. We recall that study 5 showed that the WCR score increased as English proficiency levels increased, and suggested that different clause types (as rated by the WCR) could further impact these changes.

The overview of the analysis in Study 6 is now summarized. The KW test showed that the number of accurate clauses was significantly different between the A2 and B2+, and B1_2 and B2+ groups. Moreover, further analysis showed that the number of both Lv. 1 and Lv. 2 clauses significantly differed between the A2 and B2+ groups. Finally, the analysis showed a significant difference in the number of Lv.3 clauses between the A2 and B1_2, and A2 and B2+, and B1_1 and B2+ groups. The number of all clause types significantly differed between the A2 and B2+ groups; this is reasonable because learners with higher-proficiency have more L2 knowledge than lower-proficiency learners. Japanese EFL learners in the B2+ group produced more linguistically accurate clauses and errors that had a lower impact compared to that of the A2 group. Given these differences, it is not surprising that there are significant differences in the WCR score between the two groups.

The number of accurate clauses was also significantly different between the B1_2 and B2+ groups, which could explain the significantly different WCR scores between these groups in Study 5. Descriptive statistics showed that the average number of accurate clauses in the B1_2 group (11.9) was lower than that of the B1_1 group (12.3), although the difference was not significant. A possible reason for this is the U-shaped learning pattern. Although English learners acquire new L2 knowledge and produce accurate writing in their early stages of learning, learners at a certain level tend to produce inaccurate clauses which gradually become correct as learning progresses (Gass et al., 2020). Our descriptive statistics showed that the number of Lv.2 clauses was slightly higher in the B1_2 group than that of the B1_1 group, although this difference was not significant.

The results showed that the number of Lv.1 clauses did not significantly differ between groups, except between A2 and B2+. All groups produced, on average, more than 10 Lv.1 clauses, suggesting that all groups tended to produce minor errors, such as subject/verb agreement, article, and lexical errors. Previous studies suggested that some minor errors would remain, even when the English proficiency levels increased (e.g., Thewissen, 2013). Abe (2019) investigated error rates across proficiency levels and found that the accuracy rate of the third-person singular *-s* usage did not increase. This tendency has also been identified in the results of other previous studies.

The number of Lv. 1 clauses was significantly different between some proficiency intervals, however the number of Lv. 2 clauses did not significantly differ in any groups, except between the A2 and B2+ groups. This result indicates that most learners can produce Lv. 2 clauses which contain serious, but recoverable, errors. Moreover, the analysis showed that the number of Lv. 3 clauses significantly differed between the A2 and B2+, A2 and B1_2, and B1_1 and B2+ groups, suggesting that the number of clauses which are completely incomprehensible decreases as the English proficiency levels increase because these learners possess more sophisticated L2 knowledge and interlanguage skills (Housen et al., 2012). Consequently, they tend not to produce serious errors. These results can explain the increase of the WCR scores between the A2 and B1_2, and B1_1 and B2+ groups.

Interestingly, there was no significant difference in the number of clause types between the A2 and B1_1, and B1_1 and B1_2 groups. Hence, the WCR scores also did not exhibit a significant difference. This could be because Japanese EFL learners in these proficiency levels obtain new L2 knowledge and build structurally accurate clauses.

Chapter 9
Conclusion

9.1 Summary of Findings in the Present Study

This study focused on the WCR, a new measure of accuracy in L2 writing. While previous studies using the WCR have used surveys to measure the effect of corrective feedback (e.g., Barrot, 2021) or studied its relationship to complexity (e.g., Fox, 2019), our study focused on the validity of inferences of the WCR scores. In particular, this study tested the validity of the WCR when investigating the development of English language learners' accuracy using a learner corpus. Although validation of the use of WCR scores has been conducted in previous work (e.g., Evans et al., 2014), the evidence provided by these studies is insufficient for developmental studies using a corpus. Recently, studies have examined the developmental process of CAF in L2 learners using the WCR (e.g., Barrot & Adgeppa, 2021). Because of the growing importance of the WCR in studies using learner corpora, investigating the validity of inferences made by the WCR scores is important for obtaining reliable research results.

The present study investigated the validity of the WCR using an argument-based approach (Studies 1 to 3). The results confirmed the validity of the WCR by demonstrating the validity of using WCR scores to infer the degree of written accuracy when investigating the developmental process of L2 learners' accuracy using a learner corpus.

In addition, we examined the developmental process of learners' accuracy using English texts (N = 100) produced by Japanese EFL learners. We used the ICNALE corpus (e.g., Ishikawa, 2013), which collects English texts written by Asian learners, including Japanese learners. Analysis revealed that accuracy, as measured by the WCR, generally improved as English proficiency increased (Study 5). However, we also found that there were some proficiency intervals (A2–B1_1 and B1_1–B1_2) between which written accuracy did not improve significantly. Furthermore, the detailed developmental process of learners' accuracy was investigated in terms of the types of clauses (e.g., Lv. 1) within the WCR rating scale. Our analysis revealed that the number of all clause types significantly differed between several proficiency levels.

9.2 Implications for Researchers and Teachers

Our results have three theoretical implications. First, the WCR can be used to study the development of English language learners' accuracy using a learner corpus. Previous studies have not limited the contexts in which WCR is used. Moreover, the evidence provided from previous studies is both partial and insufficient. Therefore, the reliability of the results of studies using scores obtained from the WCR is unclear. The study outlined in this dissertation overcame these limitations and demonstrated the reliability and validity of WCR-based assessments for the purpose of developmental research in a learner corpus. We provided reliable results for L2 writing researchers who wish to use the WCR to understand the development of accuracy and its characteristics at different levels of English proficiency.

It is worth noting, however, that some caution should be taken regarding accuracy assessment using the WCR. For example, there is a great deal of potential for subjectivity in the WCR assessment, and the raters' ratings may not agree with each other. This is a consequence of the results from the G study in Study 1. Hence, more time is needed for inter-rater training.

The second research implication relates to the finding that the WCR can identify differences in learner performance better than traditional accuracy measures. The results of Studies 3 and 5 showed that the WCR identified differences in accuracy better than the EFCR, which is frequently used in accuracy measures. The ability to finely distinguish between small differences in learner performance plays a crucial role in understanding differences in learner's accuracy. The WCR's ability to finely categorize clause types suggests that it can identify differences in accuracy performance better than traditional measures.

The third research implication is that the development of accuracy among Japanese EFL learners is strongly associated with a decrease in the number of clauses containing errors. In particular, a decrease in the number of clauses in Lv. 3 may be related to an improvement of written accuracy. The analysis in Study 5 revealed that there are significant differences in the number of Lv. 3 clauses across multiple proficiency levels, however differences in the number of Lv. 1 and Lv. 2 clauses were only significant between the A2 and B2+ groups. This result could not be identified by conventional measures that evaluate accuracy solely by the presence or absence of errors.

Furthermore, this study has two pedagogical implications. First, it allows teachers to recognize that the English texts produced by Japanese EFL learners are composed of clauses that are essentially comprehensible, albeit with a variety of errors. The results of the descriptive statistics show that the WCR score is above 0.8 for all proficiency groups, indicating little impact on the reader's comprehension. Traditional accuracy measures focus on the fact that there are many errors in English texts produced by EFL learners, however these errors can have little or no impact on the reader's understanding.

Second, if English teachers only want to focus on a learner's accuracy, they should devote more time to reducing Lv. 3 clauses. In this case, the ultimate goal is to reduce the number of errors and to develop learners who can write linguistically accurate English texts. It is impossible to address all errors in a piece of writing, and doing so may decrease learners' motivation to learn English (e.g., Lee, 2019). Focusing on clauses that have a significant impact on the reader's comprehension allows for communicative instruction. In Study 5, the A2 and B1_1 groups were found to have more Lv. 3 clauses than B2+. Hence, instructions may need to focus on the number of Lv. 3 clauses, particularly for less proficient learners.

9.3 Limitations

Three limitations are highlighted in this study for the purpose of validation. First, there is room for reexamination of the rating scale. The purpose of the present study was to validate the WCR; however, there is a possibility that the rating scale of the WCR was underdeveloped. The description of the errors that fall under each category in previous methods was ambiguous; therefore, in consultation with four raters, we developed a new rating scale with detailed error descriptions. While the results of the reliability analysis indicated that the ratings were very reliable, further investigation is needed to determine whether the content of the rating scale and its categorization of errors are valid. Some studies (Thewissen, 2013) have classified errors in more detail than in our study (Polio & Shea, 2014). It is necessary to create a rating scale that is easy to evaluate, while taking into account the practicality of the evaluation.

The second limitation is that a confirmatory factor analysis could not be performed. An exploratory factor analysis was conducted first because there was no consensus regarding whether the WCR reflects a construct reflected by traditional accuracy measures. The results showed that the WCR reflects the construct of accuracy. However, because the study was conducted on a learner corpus, it was not possible to test whether the model of its factor structure would fit other data. One way to solve this problem would be to conduct a confirmatory factor analysis using English texts written by learners with different L1 knowledge. This investigation would provide stronger evidence that the WCR measures accuracy.

Third, the present study was unable to conduct an investigation of the relationship between writing proficiency and the WCR. Our study was based on the CEFR, which is used by many educational institutions (e.g., universities), to investigate the relationship between the WCR and CEFR. By investigating the relationship between the WCR and English proficiency, suggestions could be made to improve teaching according to learners' proficiency level. Given that accuracy measures are also being sought for research into the predictive accuracy of writing ability (Kojima & Kaneda, 2020), research on the relationship between the WCR and writing ability is of great importance.

The WCR is a relatively new accuracy measure (Foster & Wigglesworth, 2016), and its emergence has allowed writing researchers to understand the accuracy of English compositions in greater detail than traditional accuracy measures, and to gain a more accurate picture of the accuracy levels that learners possess. While there is much appeal for the topic, questions about constructs and debates about adequacy remain unresolved. Depending on the purpose of a study, the use of the WCR may not be appropriate. Theoretical discussions, as well as validation studies, are important to enable the selection of appropriate measures for different purposes. Theoretical development of accuracy measures, including the WCR, is an important topic for future writing research.

References

Abe, M. (2007). Grammatical errors across proficiency levels in L2 spoken and written English. *The Economic Journal of Takasaki City University of Economics*, *49*(3), 117–129.

Aijmer, K. (2009). *Corpora and language teaching*. John Benjamins.

Anastasi, A. (1950). The concept of validity in the interpretation of test scores. *Educational and Psychological Measurement*, *10*(1), 67–78. https://doi.org/10.1177/001316445001000105

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, *51*, 1–38.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1966). Standards for educational and psychological tests and manuals. American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). Standards for educational and psychological testing. American Psychological Association.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly: An International Journal*, *2*(1), 1–34. https://doi.org/10.1207/s15434311laq0201_1

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

Bachman, L., & Palmer, A. (2010). *Language assessment in practice: Developing* *language assessments and justifying their use in the real world*. Oxford University Press.

Baddeley, A. D. (1986). *Working memory*. Oxford University Press.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, *7*, 54–74. https://doi.org/10.1080/15434300903464418

Barrot, J. S. (2021). Using automated written corrective feedback in the writing classrooms: effects on L2 writing accuracy. *Computer Assisted Language Learning*, 1–24. https://doi.org/10.1080/09588221.2021.1936071

Barrot, J. S., & Agdeppa, J. Y. (2021). Complexity, accuracy, and fluency as indices of college-level L2 writers' proficiency. *Assessing Writing*, *47*, 1–11. https://doi.org/10.1016/j.asw.2020.100510

Barrot, J., & Gabinete, M. K. (2019). Complexity, accuracy, and fluency in the argumentative writing of ESL and EFL learners. *International review of applied linguistics in language teaching*, *59*(2), 1–24. https://doi.org/10.1515/iral-2017-0012

Biber, D., Gray, B., & Staples, S. (2016). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, *37*, 639–668. https://doi.org/10.1093/applin/amu059

Brennan, R.L. (2001). *Generalizability theory*. Springer.

Brennan, R. L. (2013). Commentary on "validating the interpretations and uses of test scores". *Journal of Educational Measurement*, *50*(1), 74–83. https://doi.org/10.1111/jedm.12001

Brown, J. D. (2011). What do the L2 generalizability studies tell us? *Asian Journal of Assessment in Teaching and Learning*, *1*, 1–37. Retrieved from http://ojs.upsi.edu.my/index.php/AJATeL/article/view/1895

Brown, J. D. (2017). Classical test theory. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 323–335). Taylor & Francis.

Brumfit, C. (1984). Communicative Methodology in Language Teaching. Cambridge University Press.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 23–46). John Benjamins.

Burt, M. K. (1975). Error analysis in the adult EFL classroom. *TESOL quarterly*, *9*(1) 53–63. https://doi.org/10.2307/3586012

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied linguistics*, *1*(1), 1–47.

Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing*, *12*(3), 267–296. https://doi.org/10.1016/S1060-3743(03)00038-9

Chang, J. Y. (2014). The use of general and specialized corpora as reference sources for academic English writing: A case study. *ReCALL*, *26*, 243–259. http://dx.doi.org/10.1017/S0958344014000056.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2008). *Building a validity argument for the test of English as a foreign language*. Routledge.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational* *measurement* (2^{nd} ed.) (pp.443-507). American Council on Education.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? New directions for testing and measurement: Measuring achievement over a decade. Proceedings of the 1979 ETS Invitational Conference (pp. 99–108). San Francisco: Jossey-Bass.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Lawrence Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281‑302. https://doi.org/10.1037/h0040957

Cumming, A. (1990). Metalinguistic and Ideational thinking in second language composing. *Written Communication*, *7*(4), 482–511. https://doi.org/10.1177/0741088390007004003

Cureton, E. E. (1950). Validity. In E. F. Lingquist (Ed.), *Educational measurement* (pp. 621–694). American Council on Education.

Davies, M. (2008). The Corpus of Contemporary American English (COCA): 520 million words, 1990–present. Available online at http://corpus.byu.edu/coca/.

Ebel, R. (1961). Must all tests be valid? *American Psychologist*, *16*(10), 640–647. https://doi.org/10.1037/h0045478

Ellis, R., & Barkhuizen, G. (2005). *Analyzing learner language*. Oxford University Press.

Evans, N. W., Hartshorn, K. J., McCollum, R. M., & Wolfersberger, M. (2010). Contextualizing corrective feedback in second language writing pedagogy. *Language Teaching Research*, *14*(4), 445–463. https://doi.org/10.1177/1362168810375367

Evans, N. W., Hartshorn, K. J., Cox, T. L., & de Jel, T. M. (2014). Measuring written linguistic accuracy with weighted clause ratios: A question of validity. *Journal of Second Language Writing*, *24*, 33–50. https://doi.org/10.1016/j.jslw.2014.02.005

Ferris, D. R. (2010). Second language writing research and written corrective feedback in SLA: Intersections and practical applications. *Studies in Second Language Acquisition*, *32*(2), 181–201. https://doi.org/10.1017/S0272263109990490

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Sage publications.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage publications.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32*(4), 365–387.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, *18*(3), 299–323. https://doi.org/10.1017/S0272263100015047

Foster, P., & Wigglesworth, G. (2016). Capturing accuracy in second language performance: The case for a weighted clause ratio. *Annual Review of Applied Linguistics*, *36*, 98–116. https://doi.org/10.1017/S0267190515000082

Gaillat, T., Simpkin, A., Ballier, N., Stearns, B., Sousa, A., Bouyé, M., & Zarrouk, M. (2021). Predicting CEFR levels in learners of English: the use of microsystem criterial features in a machine learning approach. *ReCALL*, 1–17. https://doi.org/10.1017/S095834402100029X

Gaskell, D., & Cobb, T. (2004). Can learners use concordance feedback for writing errors? *System*, *32*(3), 301–319. http://dx.doi.org/10.1016/j.system.2004.04.001.

Gass, S. M., Behney, J., & Plonsky, L. (2020). *Second language acquisition: An introductory course*. Routledge.

Gebril, A. (2010). Bringing reading-to-write and writing-only assessment tasks together: A generalizability analysis. *Assessing Writing*, *15*(2), 100–117. https://doi.org/10.1016/j.asw.2010.05.002

Gilquin, G., De Cock, S., & Granger, S. (2010). The Louvain international database of apoken English interlanguage. Presses Universitaires de Louvain.

Goulden, N. R. (1992). Theory and vocabulary for communication assessments. *Communication Education*, *41* (3), 258–269. https://doi.org/10.1080/03634529209378887

Goulden, N. R. (1994). Relationship of analytic and holistic methods to rater's scores for speeches. *The Journal of Research & Development in Education*, *27*(2), 73–82.

Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing: an applied linguistic perspective*. Routledge.

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2009). The International Corpus of Learner English. Presses universitaires de Louvain.

Gretton, A., Bousquet, O., Smola, A., & Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. *International conference on algorithmic learning theory*, 63–77. Springer. https://doi.org/10.1007/1156408

Guion, R. (1977). Content validity: The source of my discontent. *Applied Psychological Measurement*, 1–10. https://doi.org/10.1177/014662167700100103

Hawkins, J. A. & Buttery, P. (2010) Criterial features in learner corpora: Theory and illustrations. *English Profile Journal*, *1*(1), 1–23. https://doi.org/10.1017/S2041536210000103

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing process. In C. M. Levy & S. Ransdell (Eds.), *The sciences of writing: Perspectives on writing: Research, theory, and practice*. (pp. 1–27) Lawrence Erlbaum Associates.

Hayes, J. R. (2011). Kinds of knowledge-telling: Modeling early writing development. *Journal of Writing Research*, *3*(2), 73–92. https://doi.org/10.17239/jowr-2011.03.02.1

House, E. R. (1977). *The Logic of Evaluative Argument*. Center for the Study of Evaluation.

House, E. R. (1980). *Evaluating with validity*. Sage.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 461–473. https://doi.org/10.1093/applin/amp048

Housen, A., Kuiken, F., & Vedder, I. (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy, and fluency in SLA*. John Benjamins.

Huebner, A., & Lucht, M. (2019). Generalizability theory in R. *Practical Assessment, Research, and Evaluation*, *24*(5), 1–12. https://doi.org/10.7275/5065-gc10

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

Hyland, K., & Hyland, F. (Eds.). (2019). Feedback in second language writing: Contexts and issues. Cambridge university press.

Ishikawa, S. (2011). A new horizon in learner corpus studies: The aim of the ICNALE project. In G. Weir, S. Ishikawa, & K. Poonpon (Eds.), *Corpora and language technologies in teaching, learning and research*. (pp. 3–11). University of Strathclyde Press.

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner Corpus Studies in Asia and The World*, *1*, 91–118.

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

Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin*, 112, 527–535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of educational Measurement*, *38*(4), 319–342. https://doi.org/10.1111/j.1745-3984.2001.tb01130.x

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed, pp. 17–64). American Council on Education and Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18*(2), 5–17.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*(1), 31–36. https://doi.org/10.1007/BF02291575.

Kato, T. (2019). Constructing Measurement Models of L2 Linguistic Complexity: A Structural Equation Modeling Approach. *JLTA Journal*, *22*, 23–43. https://doi.org/10.20622/jltajournal.22.0_23

Kaufer, D., Hayes, J. R., & Flower, L. S. (1986). Composing written sentences. *Research in the Teaching of English*, *20*, 121–140. https://www.jstor.org/stable/40171073

Kellogg, R. T. (1996). A model of working memory in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 57–71). Lawrence Erlbaum Associates.

Kitamura, M. (2011). Influence of Japanese EFL learner errors on essay evaluation. *ARELE: Annual Review of English Language Education in Japan*, *22*, 169–184. https://doi.org/10.20581/arele.22.0_169

Koizumi, R. (2018). How to choose and use English 4-skills tests: A validity perspective. [Eigo yonginoutesuto no erabikata to tukaikata: Datousei no kantenkara]. ALC.

Koizumi, R., & In'nami, Y. (2014). Modeling complexity, accuracy, and fluency of Japanese learners of English: A structural equation modeling approach. *JALT journal*, *36*(1), 25–46. https://doi.org/10.37546/JALTJJ36.1-2

Kojima, M. & Kaneda, T. (2020). The relationship between writing assessment and linguistic indicators: A meta-analysis for integrate research findings. [Raithinnguhyouka to genngotekisihyou no kannkei: Metabunnseki niyoru kennkyuuseika no tougou]. In Y. Ishi & Y. Kondo (Eds.), *Automated scoring in English language education: Its current situations and issues* (pp. 33–72). Hitsuji Shobou.

Kormos J (2012) The role of individual differences in L2 writing. *Journal of Second Language Writing*, *21*(4), 390–403. https://doi.org/10.1016/j.jslw.2012.09.003

Kudo, Y. (2009). A study of the characteristics of global errors of learners with different levels of English writing ability. [Eigo raithingunouryoku no reberuga kotonaru gakusyusya no gurobaru era no tokutyou ni kannsuru kennkyuu]. *ARCLE Review*, *3*, 110–121.

Kuiken, F., & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing*, *17*(1), 48–60. https://doi.org/10.1016/j.jslw.2007.08.003

Kyle, K., & Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *Modern Language Journal*, *102*(2), 333–349. https://doi.org/10.1111/modl.12468

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist*, *41*(11), 1183–1192. https://doi.org/10.1037/0003-066X.41.11.1183

Larsen-Freeman, D., & Strom, V. (1977). The construction of a second language acquisition index of development. *Language Learning*, *27*(1), 123–134. https://doi.org/10.1111/j.1467-1770.1977.tb00296.x

Lee, I. (2019). Teacher written corrective feedback: Less is more. *Language Teaching*, *52*(4), 524–536. https://doi.org/10.1017/S0261444819000247

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Input log to analyze and visualize writing processes. *Written Communication*, *30*(3), 358–392. https://doi.org/10.1177/0741088313491692

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*(3), 387–417. https://doi.org/10.1111/j.1467-1770.1990.tb00669.x

Lennon, P. (1991). Error: Some problems of definition, identification, and distinction. *Applied Linguistics*, *12*(2), 180–196. https://doi.org/10.1093/applin/12.2.180

Loevinger, J. (1966). The meaning and measurement of ego development. *American Psychologist*, *21*(3), 195–206. https://doi.org/10.1037/h0023376.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*, 474–496. https://doi.org/10.1075/ijcl.15. 4.02lu

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL quarterly*, *45*(1), 36–62. https://doi.org/10.5054/tq.2011.240859.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 33–45). Lawrence Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). American Council on Education.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241–256. https://doi.org/10.1177/026553229601300302

Michel, M. (2017). Complexity, accuracy, and fluency in L2 production. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 50–68). Taylor & Francis.

Michel, M., Révész, A., Shi, D., & Li, Y. (2019). The effects of task demands on linguistic complexity and accuracy across task types and L1/L2 speakers. In Z. Wen & M. J. Ahmadian (Eds.), *Researching L2 task performance and pedagogy: In honour of Peter Skehan* (pp. 133–151) John Benjamins.

Michel, M., Révész, A., Lu, X., Kourtali, N. E., Lee, M., & Borges, L. (2020). Investigating L2 writing processes across independent and integrated tasks: A mixed-methods study. *Second Language Research*, *36*(3), 307–334. https://doi.org/10.1177/0267658320915501

Mislevy, R. J. (2003). Substance and structure in assessment arguments. *Law, Probability and Risk*, *2*(4), 237–258. https://doi.org/10.1093/lpr/2.4.237

Mizumoto, A., & Chujo, K. (2016). Who is data-driven learning for? Challenging the monolithic view of its relationship with learning styles. *System*, *61*, 55–64. https://doi.org/10.1016/j.system.2016.07.010

Moon, S., & Oh, S. Y. (2018). Unlearning overgenerated be through data-driven learning in the secondary EFL classroom. *ReCALL*, *30*(1), 48–67. https://doi.org/10.1017/S0958344017000246

村山航. (2012). Validity: Historical and psychometric perspectives. *The Annual Report of Educational Psychology in Japan*, *51*, 118–130.

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

Oka, H. (2019). Effect of explicitness of indirect feedback on accuracy in essay writing.

KATE Journal, *33*, 27–40 https://doi.org/10.20806/katejournal.33.0_27

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*(4), 492–518. https://doi.org/10.1093/applin/24.4.492

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*(4), 590–601. https://doi.org/10.1093/applin/amp045

Pienemann, M. (1998). *Language processing and second language development: Processability theory*. John Benjamins.

Polio, C. G. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning*, *47*(1), 101–143. https://doi.org/10.1111/0023-8333.31997003

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

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

Purpura, J. E. (2004). *Assessing grammar*. Cambridge University Press.

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

R Core Team. (2018). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. https://www.R-project.org/.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., ... & Sabeti, P. C. (2011). Detecting novel associations in large data sets. *Science*, *334*, 1518–1524.

Rifkin, B., & Roberts, F. D. (1995). Error gravity: A critical review of research design.

Language Learning, *45*(3), 511–537. https://doi.org/10.1111/j.1467-1770.1995.tb00450.x

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, *22*(1), 1–30. https://doi.org/10.1191/0265532205lt295oa

Schoonen, R. (2013). The generalizability of scores from language tests. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing*. (pp. 377–391). Routledge.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage Publications.

Sheen, Y. (2007). The effect of focused written corrective feedback and language aptitude on ESL Learners' acquisition of articles. *TESOL Quarterly*, *41*(2), 255–283. https://doi.org/10.1002/j.1545-7249.2007.tb00059.x

Shimizu, Y . (2004). For understanding validity in measurement. [Sokutei niokeru datousei no rikai no tameni]. Departmental Bulletin Paper in Ritsumeikan University, *16*(4), 241–254.

Skehan, P. (1989). *Individual Differences in Second Language Learning*. Edward Arnold.

Skehan, P. (2003). Task-based instruction. *Language teaching*, *36*(1), 1–14. https://doi.org/10.1017/S026144480200188X

Spinner, P (2021). Measuring grammar. Winke, P., & Brunfaut, T. (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 233–242) Routledge.

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual review of clinical psychology*, *5*, 1–25. https://doi.org/10.1146/annurev.clinpsy.032408.153639

Stubbs, M., & Halbe, D. 2012. Corpus linguistics: Overview. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 1377–1379). Blackwell.

Suppe, P. (1977). *The structure of scientific theories*. University of Illinois Press.

Suzuki, S., Kormos, J., & Uchihara, T. (2021). The Relationship between utterance and perceived fluency: A Meta-analysis of correlational studies. *The Modern Language Journal*, *105*(2), 435–463. https://doi.org/10.1111/modl.12706.

Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal*, *97*(S1), 77–101. https://doi.org/10.1111/j.1540-4781.2012.01422.x

Tono, Y. (2007).. Nihonjin chukosei 1 man nin no eigo kopasu: JEFLL corpus. [JEFLL Corpus: English corpus of 10,000 Japanese junior and senior high school students]. Shogakukan.

Toulmin, S. E. (2003). *The uses of argument*. Cambridge University Press.

Toyoda, H. (2012). *Introduction to Factor Analysis*. [*Inshibunnseki nyuumonn*]. Tokyo tosho.

Truscott, J., & Hsu, A. Y. P. (2008). Error correction, revision, and learning. *Journal of second language writing*, *17*(4), 292–305. https://doi.org/10.1016/j.jslw.2008.05.003Get

Vann, R. J., Meyer, D. E., & Lorenz, F. O. (1984). Error Gravity: A Study of Faculty Opinion of ESL Errors. *TESOL Quarterly*, *18*(3), 427–440. https://doi.org/10.2307/3586713

Van Beuningen, C. G., 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.1467-9922.2011.00674.x

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

Weir, C. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.

Wen, Q., Wang, L., & Liang, M. (2005). *Spoken and written English corpus of Chinese learners*. Foreign Language Teaching and Research Press.

Wigglesworth, G., & Foster, P. (2008). *Measuring accuracy in second language performance*. Paper presented at the TESOL Convention.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity*. University of Hawaii Press.

Yoon & Burton (2021). Measuring L2 writing. Winke, P., & Brunfaut, T. (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 295–304). Routledge.

Appendix

Appendix A: The Rating Scale of the WCR Developed by the Present Study

Level	Code	Kinds of errors
Lv. 1 (0.8)	vhp	動詞の時制や分詞に関する誤り
	prep	前置詞の欠如・余剰・間違いに関する誤り
	rc	関係詞の欠如・余剰・間違いに関する誤り
	sv_agree	主語－動詞の一致
	pro/pos	代名詞/所有格に関する誤り
	sg/plu	名詞の単数形/複数形に関する誤り
	neg	否定の欠如・余剰・間違いに関する誤り
	art	冠詞の欠如・余剰・間違いに関する誤り
	wlex	語彙の選択に関する誤り
	wwf	品詞の誤り
	m/ex_w	語彙の欠如・余剰に関する誤り
	punc	修辞法に関する誤り
	ge/inf	動名詞/不定詞の欠如・余剰・間違いに関する誤り
	sf	断片文に関する誤り
	sp	スペリングの誤り
	wop	語順に関する誤り
	conj	接続詞の欠如・過剰・間違いに関する誤り
	qn_agree	数詞に関する誤り
	com/sup	比較級・最上級の欠如・余剰・間違いに関する誤り
	run-o	Run-on に関する誤り
Lv. 2 (0.5)	mv	be 動詞の欠如に関する誤り
	mv_cop	動詞の目的語の欠如・余剰に関する誤り
	art	冠詞の欠如・余剰・間違いに関する誤り
	prep	前置詞の欠如・余剰・間違いに関する誤り
	wlex	語彙の選択に関する誤り
	wwf	品詞の誤り
	wop	語順に関する誤り
	m/ex_w	語彙の欠如・余剰に関する誤り
	ge/inf	動名詞/不定詞の欠如・余剰・間違いに関する
	rc	関係詞の欠如・余剰・間違いに関する誤り
	wm	法助動詞の欠如・余剰・間違いに関する誤り
Lv. 3 (0.1)	nc	節の意味が読み取れない
	mv	一般動詞の欠如に関する誤り
	ms	主語の欠如に関する誤り
	mv_cop	動詞の目的語の欠如・余剰・間違い関する誤り
	rc	関係詞の欠如・余剰・間違いに関する誤り
	pv	受動態の欠如・余剰・間違いに関する誤り
	wlex	語彙の選択に関する誤り
	m/ex_w	語彙の欠如・余剰に関する誤り
	wop	語順に関する誤り