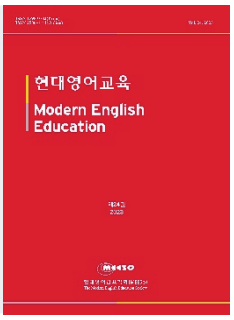
Modern English Education, Vol. 24, (2023) pp. 110-126
doi: https://doi.org/10.18095/meeso.2023.24.1.110

Modern English Education

ISSN: 1598-0782 / e-ISSN: 2586-6141
http://journal.meeso.or.kr



Exploring the Linguistic and Textual Factors That Affect the Quality of L2 Writing: The Coh-Metrix Study

Inam Ullah (Kwangwoon University)
Hyojung Lim (Kwangwoon University)

Received: 24 May 2023
Revised: 29 June 2023
Accepted: 26 July 2023

Ullah, Inam, & Lim, Hyojung. (2023). Exploring the linguistic and textual factors that affect the quality of L2 writing: The Coh-Metrix study. *Modern English Education*, *24*, 110-126.

Keywords
Coh-Metrix, L2 writing, text analysis
코메트릭스, 제2언어 작문, 문서 분석

Inam Ullah (First author)
PhD Candidate
Department of English Language and Literature
Kwangwoon University
mhnomi77@gmail.com

Hyojung Lim (Corresponding author)
Associate professor
Department of English Language and Literature
Kwangwoon University
lim@kw.ac.kr

Abstract
Efforts have been made to improve the output of automated essay evaluation systems to emulate human ratings in the field of applied linguistics. The current study aimed to explore the textual features that distinguish good from poor second-language (L2) writing using Coh-Metrix version 3.0 and to compare the output of an online text analysis tool to human rating. Sixty essays were collected from tertiary-level on-campus essay writing contests in 2017 - 2018. Three experienced English instructors rated the essays using a rubric from Jacobs et al. (1981). The Coh-Metrix output, including indices of descriptive, text easability, lexical diversity, connectives, latent semantic analysis (LSA), word information, and L2 readability, were evaluated by correlation analyses and independent t-tests. The results showed that higher scorers wrote longer, with a greater number of sentences and paragraphs. They were also likely to use more concrete words, more frequent words, and better cohesion, all of which contributed to readability. The pedagogical implications and limitations of using Coh-Metrix for L2 writing testing and instruction were further discussed.

INTRODUCTION

Second language (L2) writing has been particularly challenging in terms of instruction and evaluation (Rao, 2019). In the case of teaching and testing the writing skills of L2 learners, the method of assessment posed the biggest challenge of all (Lee, 2014). Lee further investigated the textual features that determine ESL learners' L2 writing proficiency. Results showed that ESL writers lacked the use of lexical cohesive devices in their essays. In addition, they also had difficulty in adequately justifying a claim statement and relating sub-points to a concluding remark.

The current evolvement of the automated writing evaluation (AWE) systems, such as Coh-Metrix, Criterion, and

Grammarly, have allowed teachers as well as researchers to analyze learners' writing in depth at multiple levels, ranging from simple writing mechanics to text coherence. Admittedly, AWE has been far from perfect; the feedback generated from AWE could be too vague for learners to understand (Ferster et al., 2012). Most importantly, AWE often fails to read writers' intention, idea, and mind (Lim & Kahng, 2012). Despite the limitations of AWE, however, a number of scholars have supported its promising affordances, such as motivating learners and easing teachers' workload (Hadjerrouit, 2020). Shermis (2014) also claimed that the natural language processing (NLP) tools are likely to help improve the language learning process. When it comes to Coh-Metrix, Crossley and his associates (e.g., Crossley & McNamara, 2011) have validated its use as a writing assessment tool. As language models in NLP develop rapidly, L2 researchers should be keen on what the recent AWE tools can do to improve the quality of writing instruction and assessment.

Taken together, the purpose of this study is twofold. Firstly, we aimed to identify the linguistic features of English essays written by Korean university students by using Coh-Metrix. By comparing groups (high scorers vs. low scorers), we hoped to reveal what linguistic and/or textual features contribute to higher EFL writing proficiency. Second, we also investigated the comparability between Coh-Metrix output and human rating. Human rating served as a baseline in a way of validating the performance of AWE—to what extent the automatic text analysis tool can emulate human raters. As stated in Blake (2009), writing is the area that continues to advance in conjunction with technological improvement. Given that online text analytic tools are making progress every day, L2 research on its pedagogical application needs to be updated on a regular basis.

This study can be of great importance in providing insight into the characteristics of advanced Korean EFL learners' essay writing, especially when technological resources (e.g., online dictionaries) are available to learners, and in further analyzing the linguistic and/or textual factors that affect the perceived quality of L2 writing. Findings would therefore benefit the development of L2 writing curriculum at the institutional level, L2 writing instruction at the individual level, and L2 writing assessment at all levels.

LITERATURE REVIEW

L2 Writing Ability

From a cognitive perspective, L2 writing is viewed as a product of L1 writing ability and L2 language proficiency (Weigle, 2007). As in other language skills, writing requires both bottom-up and top-down processing; the former refers to the use of L2 vocabulary and grammar knowledge, while the latter the higher-order thinking ability, which is language independent, such as synthesizing and organizing information. Aryadoust and Liu (2015) also stated that with a proper level of L2 writing ability, a person should be able to encode lexical and syntactic information, form sentence structures, and eventually express his ideas by deploying his/her metacognition. When it comes to writing instruction, Weigle made a claim that less proficient learners need to work on L2 language, while more proficient learners on the composition process. This may echo the threshold hypothesis (e.g., Goo, 2012; Lee & Schallert, 1997), in which L2 learners are likely to transfer their L1 literacy skills to L2 only if they meet a certain level of L2 proficiency. In this vein, Goo (2012) reported that among the 9th grade Korean learners of English, more advanced learners were able to transfer rhetorical patterns in their L2 essays.

From a socio-cultural perspective, however, writing is a purposeful behavior for the sake of communication with community members, thereby being often defined in a social and cultural context (Vygotsky & Cole, 1978). In other words, the ability to write allows individuals to be able to translate their ideas into words and then publish them for others to understand (Plakans & Gebril, 2017). L2 writing has become a dominant tool for communication online (e.g., Instagram), as Web 2.0 technologies have transcended the limitations of time and space. In this regard, developing L2 writing skills is often associated with the learning of genres and discourses. In reality, for EFL university students to succeed in their academic careers and job markets, strong writing skills in both L1 and L2 are required. Having said that, language learners are likely to perceive writing as the most difficult language skill to master (Sukandi & Syafar, 2018), lacking either the bottom-up or top-down processing skills.

Writing Assessment

Human Rating

The relationship between linguistic features and human evaluation of written texts has been extensively studied by Crossley and McNamara (2011). Various empirical studies have investigated human raters' decision-making and text evaluation mechanisms to examine L2 writing ability. According to Cumming et al. (2002), expert raters have developed a more comprehensive mental representation of the essay evaluation problem and used criteria, such as self-control strategies, logic and knowledge sources to read and evaluate the written texts. Crossley and McNamara (2011) reported that human raters were likely to focus on content, structure, coherence, and other linguistic factors that contribute to effective writing skills. Focusing on human raters' cognitive processing, Winke and Lim (2015) investigated raters' attention to a rubric and its relation to rater agreement; their eye-tracking study found that raters tended to pay the most attention to *organization* and *content*, and the least to *mechanics*. The criteria that received the most attention also showed the highest inter-rater reliability, which the authors speculated was due to the effect of rubric design. *Organization* and *content* were placed on the left and were therefore read first.

Rater bias has been another popular research topic in L2 writing. In Schaefer's (2008) study, where native English-speakers rated Japanese EFL learners' writing, raters who were harsh on *content* and *organization* were likely to be lenient on *language use* and *mechanics*. Some raters were harsher on more proficient L2 writers than on less proficient writers. In the context of Korean learners of English, Park (2012) reported that non-native raters tended to be harsher than native raters, although the difference was not statistically significant. Raters were likely to be harsh on *grammar*, but lenient on *content*. Taken together, human raters may read a given rubric, either analytic or holistic, differently, depending on raters' background, the rubric design, the learner group, and the context.

Automated Writing Evaluation (AWE)

Recent developments in numerous fields, including computational linguistics, discourse analysis, and data retrieval, have made it possible to automatically examine multiple surface and deep-level variables of lexical refinement, syntactic difficulty, and text harmony, thus providing specific and comprehensive language analysis (Lu, 2010). Coh-Metrix analysis is not only used by teachers to evaluate various L2 essays written by ESL students, but it is also frequently used by students and ESL learners in order to improve their writing skills. Coh-Metrix evaluates the texts based on cohesion, coherence, syntactic correctness, vocabulary variety, and narrative (Graesser et al., 2004). Coh-Metrix analyzes several sophistication-related lexical indices, such as psycholinguistic word knowledge, semantic word characteristics along with semantic word relations, word frequency indices, and several lexical indices. Coh-Metrix also uses a parser to incorporate syntax-related indices (McNamara et al., 2010).

Factors Influencing Essay Quality

Essay length appears to be a determinant of writing content (Intaraprawat & Steffensen, 1995). Thirty-nine percent of the variation in estimating the total word count of the essay explained the essay scores. Researchers worked on comparing rhetorical research of English essays written by Chinese and American students and discovered that American students wrote more extended essays than Chinese students (Lentz & De Jong, 1997). In awarding student essay scores, automatic essay scoring algorithms were found to prioritize the essay length (Machicao, 2019).

Vocabulary knowledge is another important factor in L2 language acquisition (Schmitt, 2008). Lexical diversity and syntactic complexity are also relevant indicators of English writing proficiency level. Measures of lexical diversity, sophistication, density, cohesion, and fluency are potentially of great value. McNamara and Graesser (2012) also reported that lexical diversity, word frequency, word meaningfulness, aspect repetition, and word familiarity influence the overall writing quality. This suggests that highly rated essays were not more linguistically coherent but rather more linguistically sophisticated. Independent clauses should be demarcated, or associated with connections, with complete stops (Hongwei & Liqin, 2013). According to Mody and Silliman (2008), readability is a measure of the difficulty or ease of understanding a written text.

Malmcrona (2020) mentioned that easy reads tend to be more cohesive but linguistically less sophisticated. The readability indices involve the usage of a readability technique to compute an essay's readability scores in order to measure and determine the comprehensibility of text writings. Traditional formulas for assessing readability indices include the Flesch

Reading Ease Ranking, the Flesch-Kincaid Grade Level, and the Automatic Readability Index. These typical indexes typically use non-semantic characteristics such as sentence length, syllable count, and multi-syllable word percentage (Goh et al., 2020).

Lexical diversity refers to the variety of vocabulary words used in a written text. It was identified as a salient feature of essay content in Zaytseva et al. (2019). Theoretically, essays with more diverse vocabulary were expected to gain higher scores than those with a less diverse vocabulary. Empirical studies have shown that a primary measure of future academic success is vocabulary competence (Crossley et al., 2019). Language sophistication refers to the development of unique and more complex linguistic characteristics. For example, essays written with fewer common words and a greater lexical diversity are considered more linguistically sophisticated.

Previous Studies on Coh-Metrix

Coh-Metrix has been used by several scholars to estimate various language and speech measures, such as a curriculum model, a status model, or a rhetorical framework. A number of textbooks developed by eminent writers; Blass and Pike-Baky (2007) emphasize English learners to follow certain rules to produce a good piece of writing. In the writing process, the role of correct use of grammar or spelling is trivial compared to sentence construction, structure organization of structure, or development of ideas, which have been identified as crucial features of writing skills (Ferris & Roberts, 2001; Wang et al., 2009).

According to Ferris and Roberts (2001), in English writing, the importance of accurate grammar and spelling is dismal in comparison to sentence building, development of ideas, or structure organization. However, despite all of the features mentioned earlier, the role of user feedback has still been emphasized (Ferris & Roberts, 2001; Kong & Yoon, 2013). An examination of actual writing processes such as high-level goals: idea generation, and discourse structure have been stressed more important than monitoring low-level processes like spelling and grammar (Ferrari et al., 1998). However, due to their complex nature the previous empirical research only examined high-level goals in writing. Ahn (2020) represented different proficiency levels of Korean L2 learners to identify linguistic features of written text using an automated assessment tool Coh-Metrix. From the research, Ahn shared the findings that, highly-rated text had more adverbial phrases, varied vocabulary words, and highly-rated essays consisted of less repetition and overlapped words. Ahn also discovered that a low-level text used less meaningful word, less lexical diversity, referential cohesion, and complex connectives. The study provides evidence for several aspects of language that may be used to evaluate the quality of L2 written text.

Sawaki et al. (2013) discovered that human raters put more emphasis on grammatical correction in L2 students' written text. In contrast, few scholars prioritize the study of the role of cohesion in L2 writing assessment. Lee then conducted research on cohesion in written text, concentrating on how Korean L2 learners used cohesion in a sentence and paragraph development. For this purpose, Lee conducted research on seventy-six university freshmen who were learning English.

In her study, Lee (2014) analyzed the use of cohesion, and found a significant difference between native and non-native English writers in terms of written text cohesion. However, there was no significant difference between the lower-level and higher-level L2 writers based on cohesion in their writing. This research study provides useful guidance for the English teachers on the paragraph organization, structure, and cohesion in essay writing.

Petchprasert (2021) utilized Coh-Metrix to evaluate the essay writing of ESL students. Forty students participated in an assessment of their English essay writing ability using Coh-Metrix. The students were assigned tasks to write essays on two different topics, and submit them after a revision to be evaluated using Coh-Metrix. The research results showed that various language features such as deep cohesion, referential cohesion, and word concreteness played a significant role in the students' writing performance. The findings suggest that the Coh-Metrix tool has significant potential for automated assessment to identify students' writing skill development.

Snow et al. (2015) investigated a variety of language skills applied by learners in writing prompts using Coh-Metrix to determine language characteristics and their relationship. Snow et al. estimated that 45 students published essays over eight days. Results indicated that students had a good connection with their context and prior-writing abilities in their plot versatility. Among others, Aryadoust and Liu (2015) utilized Coh-Metrix to develop a theoretical model by analyzing the relationships between text complexity and quality of written texts. The authors called for a future study to determine if the Coh-Metrix analysis tool can replace human evaluators by analyzing textual features of various writings.

Crossley and McNamara (2009) investigated the effects of text cohesion on reading comprehension. The researchers used a corpus of 100 texts that had been rated for their cohesion. The results showed that texts with higher levels of cohesion were easier to read and comprehend. Crossley et al. (2016) also examined the relationship between writing quality, text elaboration, and text cohesion. The authors found that the essays with the increased cohesion were rated as being of higher

quality. Larsen-Freeman (2006) explored the complex approach to the development of L2 speaking and writing. She examined five Chinese students who met once a week as participants in her study. For analysis, she evaluated their writing using Coh-Metrix consistently for six months. The findings showed that the Chinese students' writing skills had improved, although not in a linear fashion, in terms of complexity, fluency, and accuracy.

Coh-Metrix Output

Indices from the given variables used in the current study are as follows.

Descriptive

The descriptive indices measure the quality of a text-based on the number of words, the number of sentences, number of paragraphs, words per sentence, sentences per paragraph, syllables per word, and word frequency (Halliday & Hasan, 2014). These indices enable individuals in determining the readability of their written texts. Most second-language essay writers use longer sentences that might affect overall readability of their writings.

Studies have demonstrated that using more sentences within paragraphs has an effect on essay grades. The results show that first language users tend to score higher marks compared to second language learners (Crossley & McNamara, 2011). Lengthy sentence could result in an adverse impact on the overall quality of the essay. According to the research, second language learners tend to use more words in their sentences and rigorous and lengthy words than first language writers. The tendency of using sophisticated vocabulary and long paragraphs adversely influence the readability of their essays.

Text Easability

Coh-Metrix provides text information at various levels of linguistic research, including word descriptions, expression features, and discussion connections between ideas in a text. In comparison to conventional text easability, Coh-Metrix can provide a complete description of the potential obstacles a reader can encounter, as well as the potential trusses that the document can provide (Graesser et al., 2014).

Referential Cohesion

In research on cohesion, the distinction between cohesion and consistency is significant. Cohesion is often defined as the presence or absence of linguistic characteristics in the text that allow the reader to connect across the ideas in a piece of writing. McNamara's studies reported a positive relationship between human raters' judgments and cohesion in written text, suggesting that strong textual cohesion has a significant impact on human raters of written text. After analyzing one hundred and twenty essays, McNamara found that the essays written with strong cohesion managed to receive higher scores than those that exhibited a lack of cohesion and coherence in their essays.

Latent Semantic Analysis (LSA)

Latent semantic analysis (LSA) uses a set context window, such as the paragraph stage to perform a corpus-wide co-occurrence analysis (Evangelopoulos et al., 2012). A factor analytic approach is then extended to Coh-Metrix's co-occurrence resulting in a limited set of dimensions. LSA has been adopted by several organizations, ranging from broader samples of discourse for a particular corporation on specific domains, such as actual course themes that children and adults might have learned about.

Lexical Diversity

Lexical diversity refers to a writer's efficient use of a variety of words within a single piece of writing. Writers must use different words or synonyms to enhance writing quality (Crossley & McNamara, 2011). Similar words used in a document can reduce the overall quality of a text. A writer must use diverse word items in the same text in order to enhance its quality, but he must remember how he utilizes different words to describe the same idea in a given context. Most second language learners, for example, would use home and house interchangeably without contemplating any differences between these two words.

Connectives

Connectives serve as adhesives that hold clauses and sentences together. In a sentence, connectives link clauses, and in a piece of prose, they connect sentences (Crossley & McNamara, 2011). Connectives can consist of more than one word. Phrases like 'in other words' or 'besides' can establish connections in a text. Text linkages play an essential role in determining the overall content of an essay since they strengthen the connectivity between multiple paragraphs. ESL students who use more connectives in their essays tend to score higher than students who use fewer connectives.

Word Information

Word information provides density scores for multiple parts of speech. The given indices include nouns, verbs, adjectives, and adverbs. Grammar has always been a major challenge for second language (L2) learners (Graesser et al., 2004). The writers' fair and persuasive abilities require familiarity and application of grammatical concepts such as nouns, pronouns, and adverbs. Most L2 learners are found to overuse nouns, pronouns, adjectives, and adverbs, which reduces the quality of their essays.

Readability

The readability index goes beyond the level of reading comprehension. It also measures how quickly a person can differentiate between words or letters (Graesser et al., 2004). Improved readability enhances and speeds up the reading process, which is especially important for those with poor comprehension skills. Standard readability models anticipate accurate and consistent reported issues in the range = 0.8 or higher, according to a variety of first-language testing experiments (Dale & Chall, 1948). Sentence structure can also negatively influence the readability of the text (Crossley & McNamara, 2011). The most inappropriate behavior in writing is the use of tempting fonts and inefficient line spacing. The majority of L2 writers attempt to grab the attention of readers by using appealing fonts. Additionally, improper line spacing can make the text more difficult to read (Pitler & Nenkova, 2008).

The Present Study

The aims of the study are twofold. First, the study set out to investigate the comparability of the Coh-Metrix indices and human ratings — which and to what extent linguistic/textual indices are aligned with human judgments. To achieve this goal, the correlation coefficients were calculated between the learners' essay scores and the Coh-Metrix output. Secondly, the study explored the linguistic and textual features that differentiated good L2 writing from poor L2 writing. For data analyses, the Coh-Metrix output underwent independent t-tests for group comparisons. The research questions guiding the study are as follows:

Research Question 1. Which lexical and textual features of Coh-Metrix were correlated with human raters' essay scoring?

Research Question 2. What makes good L2 writing different from poor L2 writing?

Methodology

Data Collection

Seventy eight essays were collected from two essay writing contests held on campus during the 2017-2018 academic year (Table 1). However, only 60 essays were submitted to the current analysis, as they were evaluated by the same raters. The participants were Korean undergraduate EFL learners from various academic backgrounds. The purpose of the essay writing contest was to encourage learners to keep working on writing in English. Since they volunteered to take part in the essay contest outside of the classroom, the participants can be considered relatively motivated and advanced learners. They were asked to respond to one of the writing prompts in a form of MS Word for 75 minutes without a word limit. Participants were allowed to use an online dictionary.

TABLE 1
Information of Corpora

No. of students	Total Essays	Max. Words	Min. Words	Mean (M)	Standard Deviation (SD)
60	60	867	198	606.87	104.89

The writing topics dealt with either current events or controversial issues on campus in order to stimulate participants' critical thinking. The writing prompts in 2017 were as follows: 1) it is stated that the government's policy of raising the minimum wage has both pros and cons. Using particular examples, provide your viewpoint on the major impact of this policy on college students' substantial economic activity and part-time work, 2) recently, several Korean universities, including ours, have seen an enormous rise in foreign undergraduate students. Using concrete examples, describe your ideas, on the overall impact of recent changes on students' classroom activities, club activities, and school life. The writing prompts in 2018 were: 1) a boycott of Japanese items is currently taking place in Korean society due to the deterioration of Korea-Japan relations. Make a choice on whether to say yes or no, and explain why, 2) currently, there is an increased demand for the regular admission based on the Korean SAT exam, instead of the early admission. What is your opinion of the college admissions system?

Data Scoring

Three human raters, two Korean instructors and one native speaker of English, analytically assessed the 60 essays. All three raters had ten to twelve years of English teaching experience at the post-secondary level. The rubric used was adopted from Jacobs et al. (1981). The initial version, which has been widely examined in previous studies (Weigle 2007; Winke & Lim, 2015), comprises five sub-categories, including content, organization, vocabulary, language use, and mechanics. For the sake of practicality, however, the five categories were reduced to three for the current scoring: *content, organization*, and *language use*. *Content* assesses to what extent ideas in each essay are relevant, accurate, and thoroughly developed. *Organization* looks into how effectively sentences and paragraphs are structured in a coherent manner.

Finally, *language use* evaluates the appropriateness of vocabulary and grammar in the essay. *Mechanics* was excluded because the MS Word program corrects errors in punctuation on its own. Each criterion was scaled from limited, adequate, proficient, to exceptional. By using the analytic rubric, raters could assess English essays in a standardized and objective manner, which ensures consistency and fairness in the evaluation process. Learners can also benefit from the rubric-based scoring; scores from each sub-category provide them with clear feedback on their writing strengths and weaknesses. The human raters gave numerical grading from 1 to 10 for each criterion. The inter-rater reliability among the three human raters was calculated by using the Spearman-Brown prophecy formula:

$$\text{Spearman brown prophecy formula } rx'x' = \frac{nr}{1+(n-1)r}$$

As shown in Table 2, raters show low to moderate correlation (r = .26 between rater 1 and rater 2, r = .52 between rater 1 and rater 3, r = .27 between rater 2 and 3). Hence, inter-rater reliability among the three raters was .50.

TABLE 2
Correlation Coefficient Results

	Rater 1	Rater 2	Rater 3
Rater 1	1		
Rater 2	.26	1	
Rater 3	.52	.27	1

The student essays were further analyzed by Coh-Metrix. As illustrated earlier, Coh-Metrix has 106 indices that quantify the lexical, syntactic, and textual characteristics of a text. For analyses, we took an exploratory approach, focusing on the variables that made a significant distinction between less proficient and more proficient L2 writing.

To answer the first research question, Pearson correlation tests were conducted between the human ratings and the Coh-Metrix output. First, we calculated the correlation coefficients between the average scores of human raters. Taking individual

differences into consideration, however, we subsequently performed correlation analyses between three individual human raters and the Coh-Metrix output. To address the second research question, all sixty essays were divided into two groups, low and high scorers, and compared them. The low-scoring group (n = 30) comprised the bottom half of 60 essays, while the high-scoring group represented the upper half (n =30). For the group comparisons, the collected data were submitted to the independent t-test. To check the normality assumption, we used the Kolmogorov-Smirnov test.

The given are the statistics of Kolmogorov-Smirnov test. The total essay count was 60. The mean of the dataset was 61.53, which indicated that the average value of the data points was slightly above 61. The median, which is the middle value of the dataset, was 61, suggesting that the data was evenly distributed around the mean. The standard deviation was 12.08, implying that the data points were somewhat spread out from the mean. The skewness value of .08 indicated that the data was nearly symmetrical, with a slight positive skew. The kurtosis value of -1.31 indicated that the dataset has a platykurtic distribution, which means it has thinner tails and fewer extreme values than a normal distribution. The Kolmogorov-Smirnov test yielded a value of .208 (p =.137), indicating that the dataset did not significantly deviate from a normal distribution. As such, we concluded that the given dataset was likely to be normally distributed.

Results

According to the data presented in Table 3, there is a notable difference between Korean L2 writers in terms of the number of words, sentences, and paragraphs. The high scoring group (H) exhibited a statistically significant higher average of words (606) compared to the low scoring group (464), and (t = -5.21, p =.000). In addition, group (H) produced an average of 38 sentences, while group (L) wrote an average of 30 sentences, indicating a significant difference at (t = -3.55, p = .000). Furthermore, the difference in the average number of paragraphs produced by the two groups was also found to be significant (t = -2.72, p = .008), with group (H) producing a greater average number of paragraphs compared to group (L).

TABLE 3
Descriptive Statistics

Variables	High-Scoring Group M (SD)	Low-Scoring Group M (SD)	*t*	*df*	*p*
No. of paragraphs	5.53 (1.83)	4.33 (1.56)	-2.72	28	.008
No. of sentences	38.97 (9.42)	30.23 (9.62)	-3.55	28	.000
No. of words	606.87 (104.89)	464.30 (107.02)	-5.21	28	.000
No. of sentences in a paragraph	7.62 (2.75)	8.11 (6.37)	-0.38	28	.700
No. of words in a sentence	16.36 (4.48)	16.13 (3.37)	-0.23	28	.812

Correlations Between Coh-Metrix Output and Human Judgements

The results in Table 4 present correlation between various indices and human rater scores, where the first five indices show a significant correlation at $p < .05$, while the last two indices have a significant correlation at $p < .01$. The results indicate a negative correlation between the Z score for word concreteness and human rater scores (r = -.31, p = .025), as well as a positive correlation between the Z score for text easability PC connectivity and human rater scores (r = .30, p = .021).

Additionally, a negative correlation was observed between the incidence of additive connectives and human rater scores (r = -.26, p = .043). The indices of lexical diversity, type-token ratio, and content words were seen to be significantly and negatively correlated with the human rater's average scores (r = -.32, p = .019), indicating that learners who used fewer concrete words, fewer conjunctions, and had lower type-token ratios received higher scores from human raters.

Furthermore, there was a positive correlation between LSA overlap of adjacent paragraphs mean and human rater scores (r = .27, p = .047), suggesting that essays written coherently received higher scores. Lastly, the average acquisition level for

content words was found to be positively correlated with human rater scores (r = .31, p = .012), indicating that learners who used words that were expected to be acquired later were given higher scores by human raters.

TABLE 4
Correlation of Human Rater's Scores and Coh-Metrix Results

Variables	*r*	*t*	*p*
Text easability, PC word concreteness, z score	-.31	-2.38	.025
Text easability, PC connectivity, z score	.30	2.32	.021
LSA overlap, adjacent paragraphs, mean	.27	2.09	.047
Additive connectives incidence	-.26	-2.03	.043
Lexical diversity, type-token ratio, content word	-.32	-2.43	.019
Age of acquisition for content words, mean	.31	-2.42	.012

The results in Table 5 highlight the individual differences among the three human raters. Each rater seemed to have their own focus and preferences when evaluating the essays. Rater 2's scores showed a significant positive correlation with LSA given/new sentences (p = .017), type-token ratio (p = .010), polysemy for content words (p = .012), and hypernymy for verbs (p = .025).

This suggests that rater 2 was more likely to give higher scores to the essays that were well-organized, used a lower ratio of types to tokens, contained fewer polysemous words, and included verbs with fewer subordinate or superordinate words. In contrast, rater 3's scores showed a significant correlation with additive connectives (p = .003), age of acquisition for content words (p = .042), and concreteness of content words (p = .032).

Rater 3 tended to give lower scores to those who used a greater number of coordinating conjunctions, included words that are acquired earlier in life, and used more concrete words. On the other hand, Rater 1 did not show any significant correlation with the Coh-Metrix output, indicating that this rater may have a different approach or focus when evaluating the essays.

TABLE 5
Correlation of Human Rater's Scores and Coh-Metrix Indices

Variables	Rater 1	Rater 2	Rater 3
Text easability, PC word concreteness	-.20	-.17	-.35**
Text easability, PC connectivity, z score	.21	.10	.40**
LSA given/new, sentences, mean	.00	.32**	.21
Lexical diversity, type-token ratio, content words	-.37	-.34**	-.24
Lexical diversity, type-token ratio, all words	-.04	-.35**	-.20
Additive connectives incidence	-0.17	-.17	.41**
Age of acquisition for content words, mean	.22	.22	.27*
Concreteness for content words, mean	-.17	-.11	-.28*
Polysemy for content words, mean	-.10	-.34**	-.17
Hypernymy for verbs, mean	-.18	-.30*	-.06

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

What Makes L2 Writing Good

Text Easability

The indices of text easability, as presented in Table 6, indicate a clear distinction between high and low scoring groups of Korean ESL learners in their essay writing. The high scoring group (H) had a mean of 0.37 (*SD* = .47) for narrativity, while the low scoring group (L) had a mean of .00 (*SD* = .48). The group difference was significant at the alpha of 0.05 (*t* = -3.06, *p* = .003). In terms of word concreteness scores, the high scoring group had a mean of -0.47 (*SD*= .70), while the low scoring group had a mean of -0.94 (*SD* = .46). The difference was significant at the alpha of 0.05 (*t* = -3.03, *p* = .003). For referential cohesion z-scores, the high scoring group had a mean of 0.11 (*SD* = .74), while the low scoring group had a mean of -0.58 (*t* = -3.52, *p* = .000). Additionally, the referential cohesion scores for the high scoring group had a mean of 52.89 (*SD* = 24.23), whereas the low scoring group had a mean of 32.20 (*t* = -3.39, *p* = .001). Overall, the results indicate that high scoring essays were easier to read, with more natural sounding language, greater use of concrete words, and more referential cohesive devices employed.

TABLE 6
T-test Results of Text Easability Indices of Coh-Metrix

Variables	Group	*M*	*SD*	*t*	*p*
Text easability, PC narrativity, z score	H	.37	.47	-3.06	.003
	L	.00	.48		
Text easability, PC word concreteness, z score	H	-.47	.70	-3.03	.003
	L	-.94	.46		
Text easability, PC referential cohesion, z score	H	.11	.74	-3.52	.000
	L	-.58	.77		
Text easability, PC referential cohesion	H	52.89	24.23	-3.39	.001
	L	32.20	22.89		

Referential Cohesion

The results presented in Table 7 indicate a significant variation between the two groups of Korean L2 essay writers in terms of referential cohesion. The high scoring group achieved higher scores than the low scoring group for all five indices of referential cohesion presented in the table. For argument overlap in adjacent sentences (binary), the high scoring group had a mean of .56 (*SD* = .14), while the low scoring group had a mean of .46 (*t* = -2.82, *p* = .006). In terms of noun overlap in all sentences (binary), the high scoring group had a mean of .30 (*SD* = .12), while the low scoring group had a mean of .24 (*SD* = .11). The difference was significant at the alpha of .05 (*t* = -2.12, *p* = .037).

For argument overlap in all sentences (binary), the high scoring group had a mean of .47 (*SD* = .14), while the low scoring group had a mean of 0.38 (*t* = -2.67, *p* = .009). Additionally, for content word overlap in adjacent sentences (proportional), the high scoring group had a mean of .12 (*SD* = .04), while the low scoring group had a mean of .10 (*t* = -2.48, *p* = .016). Finally, for content word overlap in all sentences (proportional), the high scoring group had a mean of .10 (*SD* = .04), while the low scoring group had a mean of .07 (*t* = -2.97, *p* = .004). Overall, the results suggest a significant difference between the two groups, with the high scoring group demonstrating greater referential cohesion, generating more links between ideas that are easier to read.

TABLE 7
T-test Results of Referential Cohesion of Coh-Metrix

Variables	Group	*M*	*SD*	*t*	*p*
Argument overlaps, adjacent sentences, binary, mean	H	.56	.14	-2.82	.006
	L	.46	.15		
Noun overlap, all sentences, binary, mean	H	.30	.12	-2.12	.037
	L	.24	.11		
Argument overlaps, all sentences, binary, mean	H	.47	.14	-2.67	.009
	L	.38	.12		
Content word overlap, adjacent sentences, proportional, mean	H	.12	.04	-2.48	.016
	L	.10	.04		
Content word overlap, all sentences, proportional, mean	H	.10	.04	-2.97	.004
	L	.07	.02		

Latent Semantic Analysis (LSA)

The results of the t-test reported in Table 8 indicate that the high scoring group of essay writers demonstrated a greater use of LSA adjacent sentences, new sentences, and adjacent paragraphs, in comparison to the low scoring group. The results were statistically significant, revealing a clear difference between the two groups. Specifically, for LSA overlap adjacent sentences, the high scoring group had a mean of .18 (*SD* = .03), while the low scoring group had a mean .16 ($t = 2.33$, $p = .023$).

For LSA overlap adjacent paragraphs, the high scoring group had a mean of .13 (*SD* = .05), while the low scoring group had a mean of .08 ($t = 3.12$, $p = .002$). Finally, for LSA new sentences, the high scoring group had a mean of .35 (*SD* = .04), whereas the low scoring group had a mean .31 ($t = 3.36$, $p = .000$). These results suggest that the use of LSA in adjacent sentences, paragraphs, and new sentences may be a useful indicator of writing quality in second language learners.

TABLE 8
T-test Results of LSA Indices of Coh-Metrix

Variables	Group	*M*	*SD*	*t*	*p*
LSA overlap, adjacent sentences, standard deviation	H	.18	.03	2.33	.023
	L	.16	.04		
LSA overlap, adjacent paragraphs, standard deviation	H	.13	.05	3.12	.002
	L	.08	.06		
LSA given/new, sentences, mean	H	.35	.04	3.36	.000
	L	.31	.04		

Lexical Diversity

The results presented in Table 9 revealed that the high scoring group had lower scores for lexical diversity, as measured by the type-token ratio (TTR) of word lemmas, with a mean of .56 (*SD* = .09). In contrast, the low scoring group had a higher TTR score, with a mean of .66 (*SD* = .07). The group difference was statistically significant ($t = 4.84$, $p = .000$). Similarly, for the type-token ratio of all words, the high scoring group had a mean of .38 (*SD* = .05), whereas the low scoring group had a mean of .47 (*SD* = .05). The group difference was statistically significant at the alpha of .05 ($t = 6.39$, $p = .001$).

Regarding the lexical diversity MTLD indices, the high scoring group had a mean of 73.09 (*SD* =15.34), while the low scoring group had a mean of 88.04 (*SD* = 17.84). The results were significant at the alpha of .05 ($t = 3.57$, $p = .001$). However, it should be noted that the use of TTR as an index of lexical diversity has been criticized for its sensitivity to text length, as pointed out by McCarthy and Jarvis (2010).

Although advanced learners are expected to have a wider range of vocabulary in their writing, the increased TTR in the low scoring group could be due to their lack of writing fluency rather than actual lexical diversity. This possibility can be supported by Lim (2023), in which more advanced learners produced more words in a speaking task, thereby resulting in a decreased TTR. Hence, the TTR scores presented in this study should be interpreted with caution.

TABLE 9
T-test Results of Lexical Diversity Indices of Coh-Metrix

Variables	Group	*M*	*SD*	*t*	*p*
Lexical diversity, type-token ratio, content word lemmas	H	.56	.09	4.84	.000
	L	.66	.07		
Lexical diversity, type-token ratio, all words	H	.38	.05	6.39	.000
	L	.47	.05		
Lexical diversity, MTLD, all words	H	73.09	15.34	3.57	.000
	L	88.40	17.84		

Connectives

The results of the t-test in Table 10 indicate that the group with high scores used a significantly greater number of all connectives, logical connectives, and additive connectives compared to the low-scoring group. The incidence of all connectives for the high-scoring group was found with mean of 103.95 (*SD* = 18.15), whereas the low-scoring group had a mean 94.45 (t = -2.07, p = .042). Similarly, the high-scoring group had a mean of 54.30 (*SD* = 13.25) for logical connectives, while the low-scoring group had a mean of 46.85 (t = -2.50, p = .015). Additionally, the high-scoring group used more additive connectives, with a mean of 20.84 (*SD* = 10.92), compared to the low-scoring group with a mean of 14.65 (t = -2.50, p = .014).

TABLE 10
T-test Results of Connectives Indices of Coh-Metrix

Variables	Group	*M*	*SD*	*t*	*p*
All connectives incidence	H	103.95	18.15	-2.07	.042
	L	94.45	17.36		
Logical connectives incidence	H	54.30	13.25	-2.50	.015
	L	46.85	9.48		
Additive connectives incidence	H	20.84	10.92	-2.50	.014
	L	14.65	7.99		

Word Information

The results presented in Table 11 demonstrate the significant group differences in terms of lexical use. Regarding CELEX word frequency for content words, high scoring group had a mean of 2.46 (*SD* = .14), while low scoring group have a mean of 2.37 (t = -3.00, p = .012). Furthermore, the CELEX log minimum frequency for content words was also higher for high scoring group, with a mean of 1.41 (*SD* = .23). Low scoring group, in contrast, had a mean of 1.29 (*SD* = .21). The group differences were significant at the alpha of .05 (t = 3.03, p = .033). High scoring group also exhibited greater familiarity with content words, with a mean of 582.33 (*SD* = 4.88). Low scoring group, by comparison, had a mean of 576.24 (*SD* = 5.62). The group difference was significant at the alpha of .05 (t = -3.51, p =.000). Similarly, the imagability of content words was higher for the high scoring group, with a mean of 401.61(*SD* = 13.90). The low scoring group, in contrast, showed a mean of 388.41(t = -3.52, p = .000).

Moreover, the index of meaningfulness, as per the Colorado norms, was also higher for high scoring group, with an average score of 432.62 (*SD* = 9.56). In comparison, low scoring group showed an average score of 432.62 (*SD* = 11.21). The group difference was statistically significant (t = -3.39, p = .000). Finally, hypernymy for nouns is higher for low scoring group, with a mean of 6.56 (*SD* = .43). High scoring group, on the other hand, have a mean of 6.29 (*SD* = .39). The group difference was significant at the alpha of 0.05 (t = 2.82, p = .012). Taken all, the high-scoring Korean ESL essay writers appeared to use more frequent, familiar, and imaginable words that are also more meaningful and have more subordinate and superordinate words than the low scoring essay writers.

TABLE 11
T-test Results of Word Information Indices of Coh-Metrix

Variables	Group	*M*	*SD*	*t*	*p*
CELEX word frequency for content words, mean	H	2.46	.14	-3.00	.012
	L	2.37	.11		
CELEX log minimum frequency for content words	H	1.41	.23	-3.03	.033
	L	1.29	.21		
Familiarity with content words, mean	H	582.33	4.88	-3.51	.000
	L	576.24	5.62		
Imagability for content words, mean	H	401.61	13.90	-3.52	.000
	L	388.41	10.79		
Meaningfulness, Colorado norms, cont. words. mean	H	432.62	11.21	-3.39	.000
	L	419.02	9.56		
Hypernymy for nouns, mean	H	6.29	.39	2.82	.012
	L	6.56	.43		

Readability

The results in Table 12 presents the indices of readability for the essays written by Korean L2 learners, revealing significant differences between the high and low scoring groups. The scores for Flesch Reading Ease show that the high scoring group had a mean of 62.19 (*SD* = 11.32), while the low scoring group had a mean of 56.01 (t = -2.12, p = .023). This suggests that the text written by the high-scoring group was easier to read, with a Flesch Reading Ease score falling within the range of 60-70, which is considered to be standard in terms of readability.

Likewise, Coh-Metrix, a tool used to measure second language readability, indicates that the high scoring group had a mean of 22.08 (*SD* = 5.03), while the low scoring group showed a mean of 18.47 (t = -2.67, p = .001). The essays written by the high-scoring group appeared to be easier to read and comprehend. Overall, these results suggest that there were significant differences in the readability of essays between the high and low scoring groups of Korean L2 learners, with the former exhibiting greater ease of reading and comprehension.

TABLE 12
T-test Results of Text L2 Readability Indices of Coh-Metrix

Variables	Group	*M*	*SD*	*t*	*p*
Flesch reading ease	H	62.19	11.32	-2.12	.023
	L	56.01	9.16		
Coh-Metrix L2 readability	H	22.08	5.03	-2.67	.001
	L	18.47	3.32		

DISCUSSION AND CONCLUSION

The Coh-Metrix analyses revealed that the high scorers produced a greater number of paragraphs, sentences, and words; those who wrote more within a given time received higher scores. This is consistent with previous findings. MacArthur et al. (2018) found that length alone predicted 30 percent of the post-test variance in quality. In Goh et al. (2020), the total number of words and the number of words per sentence predicted high scores on the writing assessment. Kobrin and Patterson (2011) also reported that the essay length determines writing scores as words are needed to express and support ideas, logic, and arguments. The Coh-Metrix descriptive statistics corroborate that the high-scoring group wrote lengthy texts, including a greater number of paragraphs and sentences in a paragraph compared to the low scoring group. It is noteworthy that producing more language in a given time provides evidence for a higher level of fluency. Not surprisingly, the decreased TTR for high scoring group also seems attributable to their higher level of fluency; as they produced more tokens, which is commonly found in the case of discussing a single topic, TTR is supposed to drop.

Fluency, along with accuracy and complexity, consists of the construct of second language proficiency (Housen & Kuiken, 2009). In all four language skills, fluency is often achieved later in the language-learning process because it requires hours of practice in a variety of contexts. As well indicated in the published rubric for productive tests (e.g., TOEFL writing rubric), fluency can discern advanced learners from high-intermediate learners. Note that the notion of fluency embraces the accurate use of vocabulary and grammar as well as fast processing (Lim & Godfroid, 2014). Even though the rubric currently used did not have the sub-category for fluency, raters still awarded more points to longer texts. Conceivably, learners would need more words to develop ideas in depth, concretize concepts, and effectively support claims. Therefore, word count may well contribute to the quality of content. The use of an online dictionary did not seem to be of great help for low scorers because of the limited time; those who had either a relatively small vocabulary size or a low resting activation level of lexical presentations (Plag et al., 2015) would encounter challenge retrieving words from their mental lexicon while writing. In the context of L2 writing assessment, therefore, we do not need to hesitate to allow students to use online tools, as their accessibility does not significantly affect test scores.

According to the Coh-Metrix analyses, high scorers also used more concrete words, more frequent words, more familiar words, more imaginable words, and more meaningful words. At first glance, this seems to contradict previous findings in L2 vocabulary research, since concrete and high-frequency words are relatively easy to learn (Peters, 2019). However, as the readability indices show, human raters gave high scores to the easy-to-read texts. In view of the socio-cultural theory, the ultimate purpose of writing is to communicate effectively with readers. For L2 learners to make them understood better, they should be able to use vocabulary appropriately as well as accurately and fluently. In this regard, the absolute proportion of low-frequency words does not seem to account for the quality of L2 essays; using more low-frequency words does not guarantee better writing quality. We cannot rule out the possibility that low scorers could have simply listed low-frequency words out of context. In a similar vein, Lim (2023) uncovered that advanced Korean EFL learners beyond the B2 level did not necessarily use more low-frequency words than those below the B2 level during speaking assessment.

The Coh-Metrix analyses also showed that high scorers could organize ideas more coherently both at the sentence and paragraph level, deploying a greater number of logical and additive connectives. In addition to easy-to-read vocabulary, cohesion seems to contribute to readability, which in turn leads to higher scorers. As the given rubric directs raters' attention to *organization*, it makes sense that the Coh-Metrix output relevant to text cohesion, both referential cohesion and laten semantic analysis indices, made a meaningful distinction between the high and the low scoring group. More proficient learners were able to link sentences and paragraphs better by repeating key words and appropriate conjunctions. This is consistent with Crossley et al. (2016), in which high-score texts contained more given information, better maintained cohesion and thus showed a better degree of comprehensibility.

Taken together, the current study suggests the feasibility of Coh-Metrix as an evaluation and an instructional tool. As the correlation analyses show, the Coh-Metrix output seems well-aligned with human rating, in terms of assessing learners' vocabulary use and text cohesion. The Coh-Metrix output complements human rating by quantifying raters' impressionistic evaluation. Jung et al. (2019) also supported the use of computational indices of Coh-Metrix to examine human evaluations of L2 writing proficiency. The study revealed that different linguistic features contribute differently to overall second language writing proficiency scores. Specifically, features related to text length and lexical complexity were found to be more important predictors of writing quality than those related to cohesion and syntax complexity. However, it is worth noting that individual raters bring in their own idiosyncrasy, however experienced or well-trained they are, and thus human judgments may well come out differently. As shown in the earlier correlation analyses (Table 5), albeit with the same rubric, raters seem to put a differing amount of value on various aspects of writing. Recall that rater 3 performed similarly to Coh-Metrix when rating vocabulary and cohesion. Rater 2 seemed to accord with Coh-Metrix in view of latent semantic overlaps, lexical diversity, and some lexical characteristics. Rater 1 did not show any meaningful correlations with Coh-Metrix. To minimize the effect of rater bias, or individual differences among raters, it appears critical to have a sufficient norming session before rating. Coh-Metrix can serve as a complement, increasing the objectivity of human rating and reducing rater training costs.

In addition to the testing purpose, Coh-Metrix can play as a personal learning tool. Learners could benefit from the Coh-Metrix output to understand their weaknesses as well as strength in L2 writing, especially for their use of vocabulary and text cohesion. As in other CALL tools, however, it is important for instructors to provide proper guidance — how to use Coh-Metrix, when to use it, how to interpret the output, and how to incorporate the output into revision. For instance, students should be told to interpret TTR with caution as it may explain fluency rather than lexical diversity. Note that Coh-Metrix was not created for L2 writing instruction, but for language research in general. Depending on one's learning style, preference, and familiarity with CALL tools, the effectiveness of Coh-Metrix as a learning tool could vary. From a teacher's perspective, however, Coh-Metrix can shine as a teaching assistant (Shiroza, 2020). It certainly reduces teachers' workload of providing feedback, while improving the quality of the feedback. As suggested in the past (e.g., Lim & Kahng, 2012),

teachers may divide their work with the online text analysis tool, as they can better take care of content, logic, and creativity, while Coh-Metrix is limited in analyzing a text at the pragmatical level (Crossley et al., 2014).

This study is far from perfect. Caution needs to be taken when it comes to interpreting the findings. The sample size may be insufficient to draw general conclusions. Only argumentative essays were considered in the study. Notably, all of the participants in this study were relatively motivated and advanced EFL learners, as they volunteered to participate in an on-campus English essay writing content as an extra-curricular activity. If the data were collected in a regular English class, or if the learners were asked to write narrative genres, the results might have yielded different outcomes. In respect of data analysis, the groups were evenly divided into the two groups; the high-scoring group (n = 30) and the low-scoring group (n =30). If we had removed the gray area in the middle, then the statistical results might have turned out to be different; otherwise, the picture might have been clearer.

The recent advancement of online technologies, including machine translation and chat GPT, has transformed the nature and the process of L2 writing. Future research can further analyze such technology-assisted L2 writing with Coh-Metrix and see which area of a text can improve with the help of technology or what makes good writers from poor writers. For the practical use of Coh-Metrix in L2 writing class, action research by practitioners would be also necessary.

References

Ahn, S. (2020). Discriminating L2 writing proficiency through the use of computational tool Coh-Metrix. *The Journal of Linguistics Science*, *92*, 209-232.

Aryadoust, V., & Liu, S. (2015). Predicting EFL writing ability from levels of mental representation measured by Coh-Metrix: A structural equation modeling study, *Assessing Writing*, *24*, 35-58.

Blake, R. J. (2009). The use of technology for second language distance learning. *The Modern Language Journal*, *93*, 822-835.

Blass, L., & Pike-Baky, M. (2007). *Mosaic one: Writing paragraph review and essay development*. McGraw-Hill.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, *32*, 1-16.

Crossley, S. A., & McNamara, D. S. (2009). Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing, 18*(2), 119-135.

Crossley, S. A., & McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life-Long Learning, 21(2/3),* 170-191.

Crossley, S. A., Skalicky, S., & Dascalu, M. (2019). Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading, 42*(3-4), 541-561.

Crossley, S. A., Yang, H. S., & McNamara, D. S. (2014). What's so simple about simplified texts? A computational and psycholinguistic investigation of text comprehension and text processing. *Reading in a Foreign Language*, *26*(1), 92-113.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, *86*(1), 67-96.

Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational Research Bulletin*, *27*(1), 37-54.

Evangelopoulos, N., Zhang, X., & Prybutok, V. R. (2012). Latent semantic analysis: Five methodological recommendations. *European Journal of Information Systems*, *21*(1), 70-86.

Ferrari, M., Bouffard, T., & Rainville, L. (1998). What makes a good writer? Differences in good and poor writers' self-regulation of writing. *Instructional Science, 26,* 473-488.

Ferris, D., & Roberts, B. (2001). Error feedback in L2 writing classes. *Journal of Second Language Writing*, *10*(3), 161-184.

Ferster, B., Hammond, T. C., Alexander, R. C., & Lyman, H. (2012). Automated formative assessment as a tool to scaffold student documentary writing. *Journal of Interactive Learning Research*, *23*(1), 81-99.

Goh, T. T., Sun, H., & Yang, B. (2020). Microfeatures influencing writing quality: The case of Chinese students' SAT essay. *Computer Assisted Language Learning*, *33*(4), 455-481.

Goo, Jihae. (2012). Transfer of L1 writing proficiency among EFL students: Testing the threshold hypothesis with Korean 9th grade EFL learners. *Studies in Linguistics, 17*(2), 135-156.

Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-Metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal*, *115*(2), 210-229.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193-202.

Hadjerrouit, S. (2020, September). *Exploring the affordances of Numbas for mathematical learning: A case study* [Paper presentation]. DRUM 2020, Tunis, Tunisia. https://hal.science/hal-03113967v1/document

Halliday, M. A. K., & Hasan, R. (2014). *Cohesion in English*. Routledge.

Hongwei, W., & Liqin, Y. (2013). A computational analysis of textual features and L2 writing proficiency. *International Journal of Academic Research in Progressive Education and Development*, *2*(4), 170-185.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*(4), 461-473.

Intaraprawat, P., & Steffensen, M. S. (1995). The use of metadiscourse in good and poor ESL essays. *Journal of Second Language Writing*, *4*(3), 253-272.

Jacobs, H. L., Zingraf, S. A., Wormuth, D. R., Hartfield, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Newbury House.

Jung, Y. J., Crossley, S., & McNamara, D. (2019). Predicting second language writing proficiency in learner texts using Computational Tools. *The Journal of Asia TEFL*, *16*(1), 37-52.

Kobrin, J. L., & Patterson, B. F. (2011). Contextual factors associated with the validity of SAT scores and high school GPA for predicting first-year college grades. *Educational Assessment, 16*(4), 207-226.

Kong, Eun Jung., & Yoon, In Hee. (2013). L2 proficiency effect on the acoustic cue-weighting pattern by Korean L2 learners of English: Production and perception of English stops. *Journal of the Korean Society of Speech Sciences*, *5*(4), 81-90.

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics*, *27*(4), 590-619.

Lee, Eun-Hee. (2014). An analysis of Korean EFL university learners' use of cohesion in writing with Coh-Metrix. *The Journal of Foreign Studies*, *27*, 195-216.

Lee, J-W., & Schallert, D. L. (1997). The relative contribution of L2 language proficiency and L1 reading ability to L2 reading

performance: A test of the threshold hypothesis in an EFL context. *TESOL Quarterly*, *31*(4), 713-739.
Lentz, L., & De Jong, M. (1997). The evaluation of text quality: Expert-focused and reader-focused methods compared. *IEEE Transactions on Professional Communication*, *40*(3), 224-234.
Lim, Hyojung. (2023). Exploring the development of second Language (L2) vocabulary in speaking. *Studies in Foreign Language Education, 37*(1), 145-163.
Lim, H., & Godfroid, A. (2014). Automatization in second language sentence processing: A partial, conceptual replication of Hulstijn, Van Gelderen, and Schoonen's 2009 study. *Applied Psycholinguistics*, *36*(5), 1247-1282.
Lim, H., & Kahng, J. (2012). Review of criterion for English language learning. *Language Learning & Technology*, *16*(2), 38-45.
Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474-496.
MacArthur, C. A., Jennings, A., & Philippakos, Z. A. (2018). Which linguistic features predict quality of argumentative writing for college basic writers, and how do those features change with instruction? *Reading and Writing*, *32*(6), 1553-1574.
Machicao, J. C. (2019). Higher education challenge characterization to implement automated essay scoring model for universities with a current traditional learning evaluation system. In Á. Rocha, C. Ferrás & M. Paredes (Eds.), *Information technology and systems: Advances in intelligent systems and computing* (pp. 835-844). Springer.
Malmcrona, H. (2020). Variation in assessment: A Coh-Metrix analysis of evaluation of written English in the Swedish upper secondary school. [Unpublished master's thesis]. Stockholm University.
McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381-392.
McNamara, D. S., & Graesser, A. C. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In P. M. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing and content analysis: Identification, investigation and resolution* (pp. 188-205). IGI Global.
McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of Cohesion. *Discourse Processes*, *47*(4), 292-330.
Mody, M. E., & Silliman, E. R. (2008). *Brain, behavior, and learning in language and reading disorders*. Guilford Press.
Park, Myo-Young (2012). Exploring the raters' bias on an EFL writing assessment using multi-faceted rasch measurement. *Studies in English Education, 17*(2), 175-202.
Petchprasert, A. (2021). Utilizing an automated tool analysis to evaluate EFL students' writing performances. *Asian-Pacific Journal of Second and Foreign Language Education*, *6*(1), 1-16.
Peters, E. (2019). Factors affecting the learning of single-word items. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 125-142). Routledge.
Pitler, E., & Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In M. Lapata, H. & Tou Ng (Eds.), *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 86-195). Association for Computational Linguistics.
Plag, I., Arndt-Lappe, S., Braun, M., & Schramm, M. (2015). *Introduction to English Linguistics*. Walter de Gruyter GmbH & Co KG.
Plakans, L., & Gebril, A. (2017). An assessment perspective on argumentation in writing. *Journal of Second Language Writing*, *36*, 85-86.
Rao, P. S. (2019). The significance of writing skills in ELL environment. *Academician: An International Multidisciplinary Research Journal*, *9*(3), 5-17.
Sawaki, Y., Quinlan, T., & Lee, Y. W. (2013). Understanding learner strengths and weaknesses: Assessing performance on an integrated writing task. *Language Assessment Quarterly, 10*(1), 73-95.
Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465-493.
Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research*, *12*(3), 329-363.
Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, *20*, 53-76.
Shiroza, S. (2020). Changing language, continuing discourse: A shift toward ELF and persistent native-speakerism in Japan's ELT policy. In R. A. Giri, A. Sharma, & J. D'Angelo (Eds.), *Functional Variations in English: Theoretical Considerations and Practical Challenges* (pp. 277-293). Springer.
Snow, E. L., Allen, L. K., Jacovina, M. E., Crossley, S. A., Perret, C. A., & McNamara, D. S. (2015). Keys to detecting writing flexibility over time: entropy and natural language processing. *Journal of Learning Analytics*, *2*(3), 40-54.
Sukandi, S. S., & Syafar, D. N. (2018). EFL students' responses to learning basic reading and writing skills. *Studies in English Language and Education*, *5*(1), 40-53.
Vygotsky, L. S., & Cole, M. (1978). *Mind in society: Development of higher psychological processes*. Harvard university press.
Wang, Q., Woo, H. L., & Zhao, J. (2009). Investigating critical thinking and knowledge construction in an interactive learning environment. *Interactive Learning Environments*, *17*(1), 95-104.
Weigle, S. C. (2007). Teaching writing teachers about assessment. *Journal of Second Language Writing*, *16*(3), 194-209.
Winke, P., & Lim, H. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing, 25*, 38-54.
Zaytseva, V., Miralpeix, I., & Pérez-Vidal, C. (2019). The effects of contexts on the acquisition of oral lexical ability in English as a foreign language. *The Language Learning Journal*, *49*(5), 597-613.