

RESEARCH ARTICLE

A Study on the Possibility of Replacing Holistic Evaluation with Linguistic Feature Assessment in Essay Evaluation

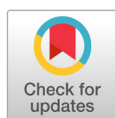
Min-Seok Choi, Dong Ju Lee^{*}
Korea National University of Education

^{*}Corresponding Author: Dong Ju Lee, maydjlee@knue.ac.kr

ABSTRACT

This study investigated the relationship between the linguistic features of English essays and the holistic evaluations conducted by non-native Korean teachers of English, while also examining the applicability of automated essay scoring. A total of 23 essays written by second- and third-year high school students were analyzed using Coh-Metrix 3.0, with particular attention to indices of lexical diversity, syntactic complexity, cohesion, and readability. Spearman's rank-order correlation analysis was performed using SPSS to examine the relationships between the linguistic indices and the teachers' evaluation rankings. The results revealed that lexical diversity (LDTTRa) demonstrated a relatively strong positive correlation with teachers' evaluations, whereas indices related to syntactic complexity, cohesion, and readability exhibited only limited correlations. These findings suggest that lexical features may play a significant role in teachers' holistic judgments of essay quality. Furthermore, the study indicates that Coh-Metrix-based automated scoring may serve as a supplementary tool to enhance the objectivity and reliability of essay assessment, rather than functioning as a replacement for human raters.

Keywords: English essay evaluation, Coh-Metrix, automated scoring, lexical diversity



OPEN ACCESS

Brain, Digital, & Learning
2026, Vol. 16, No. 2, 189–204

<https://doi.org/10.31216/BDL.2026.16.2.5>

Received: June 09, 2026

Revised: June 30, 2026

Accepted: June 30, 2026

© 2026 Institute of Brain based Education,
Korea National University of Education



This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Introduction

English writing is not merely an activity of sentence production; rather, it constitutes a complex domain of language performance in which learners' linguistic knowledge, cognitive processes, and writing competence are simultaneously manifested (Hyland, 2003). In particular, L2 essay writing demands higher-order language abilities that extend beyond the mere generation of grammatically accurate sentences. It requires writers to formulate a clear position on a given topic, develop and sustain arguments through relevant supporting reasons and evidence, and construct a coherent and cohesive text through the logical organization of ideas across and within paragraphs. In this regard, Graham and Perin (2007) argued that effective writing involves a range of interrelated competencies, including idea organization, logical argumentation, and critical thinking.

Given the multidimensional nature of L2 essay writing, writing assessment has increasingly

been conceptualized as a holistic process that simultaneously takes into account both linguistic and content-related dimensions of written performance. Accordingly, teachers frequently employ holistic scoring methods when evaluating students' essays in educational settings (Lee & Choi, 2012). The various dimensions assessed in a student's essay, such as content, organization, and language use, do not operate as discrete constructs; rather, they interact dynamically to shape the overall quality of the written text. Consequently, raters tend to evaluate writing performance by forming an overall judgment of textual effectiveness while considering the interplay among these dimensions (Lee, 2019).

From this perspective, Brown and Abeywickrama (2019) defined writing assessment as a process of evaluating learners' communicative competence and discourse-construction abilities rather than merely measuring their linguistic knowledge. Similarly, Lee et al. (2017) conceptualized writing assessment as an integrated evaluative activity through which learners' cognitive processes and communicative abilities are revealed. They further emphasized that effective writing assessment requires raters to holistically consider content quality, organizational effectiveness, and quality of written expression when evaluating students' written compositions.

Despite the pedagogical value of holistic essay assessment, the evaluation process is inherently dependent on raters' subjective judgments, raising persistent concerns regarding inter-rater reliability. This issue is particularly salient for Korean EFL teachers, who are required to evaluate English essays produced outside their native linguistic context and may therefore experience a considerable assessment burden (Kwon, 2010). Such concerns stem from the fundamental nature of essay assessment itself. Unlike objective-response assessments in subjects such as mathematics or science, where clearly defined correct answers exist, essay evaluation necessarily involves qualitative judgments regarding content quality, logical development, and communicative effectiveness.

Against this backdrop, increasing attention has been directed toward the potential use of Automated Essay Scoring (AES) systems as a means of enhancing objectivity and consistency in writing assessment. As a result, a growing body of research has examined the degree of agreement between human raters and automated scoring systems in order to determine the feasibility of incorporating automated assessment into educational contexts (Si et al., 2013). For example, Shin (2022) developed an automated scoring model that incorporated rating criteria commonly employed by human evaluators and reported a significant relationship between automated scores and human ratings. Similarly, Park et al. (2021) demonstrated that an AES model based on Long Short-Term Memory (LSTM) architecture achieved a relatively high level of predictive accuracy in estimating English essay scores, thereby highlighting the potential applicability of artificial intelligence-based assessment technologies in educational settings.

Nevertheless, the extent to which automated scoring systems can genuinely substitute for human evaluators remains an open question. Most AES systems rely primarily on the quantification of observable linguistic features and thus evaluate writing through measurable textual characteristics (Lee, 2013). Although such approaches can provide objective assessments of formal linguistic aspects, including lexical sophistication, syntactic complexity, and textual cohesion, they may be less capable of directly evaluating higher-order dimensions of writing, such as the validity of argumentation, the appropriateness of supporting evidence, and the demonstration of critical thinking. These dimensions constitute essential components of effective essay writing and are often central to human evaluators' judgments.

At the same time, writing assessment in authentic educational contexts is rarely based exclusively on content-related

considerations. Previous research suggests that formal linguistic features, including grammatical accuracy, lexical sophistication, and fluency, substantially influence raters' overall impressions of writing quality (Li & Huang, 2022). Consequently, it is plausible that quantitative linguistic indices generated through automated analysis may exhibit meaningful relationships with human evaluators' holistic judgments. If such linguistic indicators consistently align with teachers' overall evaluations, automated scoring systems may possess considerable value as supplementary tools that support, rather than replace, human assessment practices.

Motivated by these considerations, the present study investigates the extent to which quantitative linguistic features identified in Korean high school students' English essays are associated with the holistic evaluations assigned by non-native Korean English teachers. More specifically, this study employs Coh-Metrix to quantify various linguistic characteristics of student essays and examines the relationships between these linguistic indices and teachers' holistic scoring outcomes. Through this analysis, the study seeks to explore the extent to which linguistic feature-based assessment may complement or potentially serve as an alternative component within holistic writing evaluation.

Accordingly, the following research questions guide the present study:

1. What linguistic indices can be identified in students' English essays through Coh-Metrix analysis?
2. What relationships exist between the identified linguistic indices and the holistic evaluation scores assigned by non-native Korean English teachers?
3. To what extent can linguistic feature-based assessment supplement or potentially substitute for holistic essay evaluation?

Theoretical Background

English Essay Assessment and Holistic Scoring

In educational settings, English essay assessment represents a prominent form of performance-based assessment that requires not only learners' linguistic competence but also higher-order cognitive abilities, including critical thinking, logical reasoning, and effective organization of ideas (Brown & Abeywickrama, 2019; Lee et al., 2017). Unlike assessments that primarily measure discrete linguistic knowledge, essay assessment aims to evaluate multiple dimensions of writing performance, encompassing content development, organizational effectiveness, and language use in addition to grammatical and lexical accuracy (Lee, 2019).

Given the multidimensional nature of writing, holistic scoring has been widely adopted in English essay assessment. Holistic scoring emphasizes the overall quality and communicative effectiveness of a written text rather than evaluating individual components separately, as is typically done in analytic scoring approaches (Hyland, 2003). Hughes and Hughes (2020) noted that raters frequently assign scores based on their overall impression of a composition rather than independently assessing specific dimensions such as content relevance, logical organization, and linguistic accuracy. Similarly, Cooper (1977) defined holistic scoring as an evaluative approach in which a single score is assigned on the basis of the evaluator's overall judgment of a text. He further argued that holistic scoring offers practical advantages, particularly in large-scale writing assessments, because it enables efficient evaluation of a substantial number of written

responses within a limited timeframe.

Despite its practical benefits, holistic scoring has long been criticized for potential issues related to reliability. Because scoring decisions are often influenced by raters' intuitive and subjective judgments, concerns regarding inter-rater consistency have been consistently raised (Park, 2006). Weigle (2002) similarly argued that although holistic scoring is advantageous in terms of efficiency in second-language writing assessment, it may be susceptible to rater bias and subjective interpretation, resulting in lower levels of inter-rater reliability compared with analytic scoring. In secondary school English assessment contexts, challenges associated with maintaining scoring consistency and managing evaluative workload have likewise been reported (So, 2023).

Furthermore, unlike selected-response assessments, such as multiple-choice tests that involve predetermined correct answers, English essay assessment is inherently open-ended in nature. As a result, evaluation outcomes may vary depending on the criteria or textual features that individual raters prioritize during the assessment process (Lee, 2022). Even when evaluating the same piece of writing, raters may assign different scores according to the relative importance they place on factors such as content quality, organizational coherence, linguistic accuracy, or overall communicative effectiveness. Consequently, ensuring fairness, consistency, and objectivity remains a persistent challenge in English essay assessment.

Automated Essay Scoring (AES)

Automated Essay Scoring (AES) refers to the process by which computer-based systems, rather than human raters, evaluate learners' written texts and assign scores. Interest in AES has grown substantially in educational contexts due to its potential to enhance scoring efficiency while ensuring objectivity and consistency in assessment practices (Dikli, 2006). AES systems assess writing quality by analyzing a range of linguistic features, including lexical sophistication, syntactic complexity, cohesion, and readability. By minimizing the influence of human subjectivity, these systems contribute to improved reliability and consistency in writing assessment. Furthermore, AES has attracted considerable attention for its pedagogical value, as it can provide learners with timely and objective feedback while reducing teachers' assessment workloads (Park et al., 2015).

Several studies have highlighted the effectiveness of AES in educational settings. For instance, Lim et al. (2014), in their investigation of an automated scoring program for English short-answer writing assessment, argued that AES systems can enhance both scoring efficiency and consistency in large-scale assessment contexts while facilitating more immediate feedback for learners. Similarly, Dikli and Bleyle (2014) compared automated essay scoring with teacher-provided feedback on essays written by ESL learners and found that automated systems were effective in delivering rapid and repeated feedback on grammatical, spelling, and mechanical errors, thereby alleviating teachers' grading burdens. In addition, Wilson and Roscoe (2020) reported that automated feedback generated through automated writing evaluation (AWE) systems positively influenced students' writing self-efficacy and English writing performance. Their findings further suggested that such systems can reduce the time and effort required for assessment while maintaining overall writing quality.

Among the AES-related tools frequently employed in writing assessment research, Coh-Metrix has received considerable attention. Graesser et al. (2004) described Coh-Metrix as a computational tool capable of analyzing texts

from multiple dimensions, including cohesion, linguistic complexity, and readability. Because it provides quantitative measures of lexical diversity, syntactic complexity, and discourse cohesion, Coh-Metrix has been widely utilized in both English language education and automated writing assessment research.

English Essay Assessment and Linguistic Features

In the assessment of English essays, linguistic features constitute essential criteria for evaluating overall writing quality and proficiency. Among these features, lexical diversity refers to the extent to which learners employ a wide range of vocabulary in their writing. Engber (1995), in an analysis of essays written by ESL learners, demonstrated that lexical diversity is closely associated with writing quality. Similarly, Kyle and Crossley (2016) identified lexical diversity and lexical sophistication as significant predictors of English writing scores. Their analysis of essays produced by learners across varying proficiency levels revealed that higher-rated essays tended to contain a broader range of vocabulary and more sophisticated lexical items. The authors further reported that these lexical characteristics were significantly correlated with holistic writing scores. Consistent with these findings, Jarvis et al. (2003) found that highly rated essays written by English language learners exhibited greater lexical variety and more complex linguistic expressions. Essays that employed contextually appropriate and diverse vocabulary, rather than relying repeatedly on basic lexical items, were generally evaluated more favorably in terms of overall writing quality.

Syntactic complexity is widely regarded as an indicator of learners' ability to produce not only simple sentence structures but also more sophisticated and complex constructions. Ortega (2003) argued that syntactic complexity serves as an important measure of second language proficiency and demonstrated that learners' developmental levels are closely linked to their ability to use complex syntactic structures. Likewise, Lu (2011), through an analysis of ESL learners' written production, reported that learners with higher levels of English proficiency tended to produce more varied and complex sentence structures. These characteristics were reflected in greater clause length, more frequent use of subordinate clauses, and higher overall levels of syntactic complexity.

Cohesion refers to the linguistic property through which textual elements are semantically connected to create a unified and coherent discourse. Halliday and Hasan (1976) characterized cohesion as a fundamental feature that facilitates text comprehension, arguing that highly cohesive texts enable readers to follow the progression of ideas more effectively. In a study of ESL learners' writing, Lee (2002) similarly emphasized that the appropriate use of cohesive devices plays a critical role in establishing logical flow and coherence. Furthermore, Crossley and McNamara (2010) argued that cohesion and coherence significantly influence essay quality assessments, reporting that evaluators tend to place greater emphasis on the natural and logical development of ideas than on the mere presence of cohesive devices.

Readability, which reflects the ease with which a text can be read and understood, is another important indicator of language proficiency. Readability is typically influenced by factors such as sentence length, lexical difficulty, and textual complexity. Using the Coh-Metrix text analysis tool, Graesser et al. (2004) demonstrated that highly readable texts facilitate readers' comprehension and positively influence evaluations of writing quality. Their findings suggest that readability constitutes a significant factor in the assessment of English writing.

Methods

Participants

The participants in this study consisted of 23 high school students (11 second-year students and 12 third-year students) who participated in an English essay contest at J High School in Gangwon Province, South Korea. A total of 23 essay samples produced by these students were collected and analyzed. The students were asked to write an argumentative essay on the following topic: “Traditional classroom teaching is more effective than online classes. Choose your position and support it with two specific reasons and examples.”

The essays were evaluated independently by four Korean teachers of English, two for each grade level. All evaluators had more than five years of teaching experience and were not responsible for teaching the participating students in their respective grade levels, thereby minimizing potential bias in the assessment process. Following the completion of the evaluations, the handwritten essays were transcribed into electronic text format and converted into word-processing documents to facilitate subsequent natural language processing and computational analyses.

Tools and Procedures

Teacher Scoring Rubrics

Prior to the essay contest, the participating English teachers collaboratively developed an analytic scoring rubric to ensure consistency and fairness in the assessment process. The rubric was designed on a 100-point scale, allocating 60% of the total score to the content domain and 40% to the language use domain. The scoring rubric is presented in Table 1.

Table 1. Scoring rubric for the English essay contest

Category	Description	Points	Total
Content (60 points)	① Does the essay demonstrate a logical organizational structure, including an introduction, body, and conclusion?	10	100
	② Does the writer present and develop a persuasive and coherent argument throughout the essay?	30	
	③ Does the writer provide logical and relevant evidence to support the argument?	20	
Language Use (40 points)	④ Does the writer express ideas accurately and effectively in English?	20	20
	⑤ To what extent is the essay free from errors in grammar, word order, spelling, and other aspects of language form?	20	

The scoring rubric comprised two primary domains: content and language use. The content domain, accounting for 60 points, evaluated the extent to which students organized their essays according to a conventional essay structure, developed persuasive and coherent arguments, and supported their claims with logical and relevant evidence. The language use domain, worth 40 points, assessed students’ ability to communicate their ideas accurately and effectively in English, with particular attention to grammatical accuracy, word order, spelling, and other aspects of linguistic form.

By incorporating both rhetorical and linguistic dimensions of writing, the rubric reflected the principles of holistic writing assessment while providing structured criteria for evaluating multiple facets of students’ essay performance. Such an approach enabled raters to consider the overall quality of the essays while simultaneously attending to key components of effective academic writing.

During the essay assessment process, two English teachers independently assessed the students’ essays using the

scoring rubric described above. Upon completion of the independent evaluations, the raters participated in a consultation meeting to review and discuss their scores, thereby ensuring consistency in the application of the assessment criteria. Final rankings were determined based on the combined scores assigned by the two raters. The detailed scoring results for each grade level are presented in Tables 2 and 3.

Table 2. Detailed scoring results for second-year students

Student (2nd Grade)	Content (60 points)			Language Use (40 points)		Total	Rank
	①	②	③	④	⑤		
Student 1	9	27	19	13	16	84	5
Student 2	10	25	19	15	17	86	4
Student 3	9	23	18	15	18	83	6
Student 4	10	29	20	19	19	97	1
Student 5	10	25	18	16	18	87	3
Student 6	9	27	17	18	17	88	2
Student 7	8	25	19	16	15	83	6
Student 8	9	26	17	15	15	82	7
Student 9	10	25	15	15	16	81	8
Student 10	10	25	18	16	17	86	4
Student 11	10	26	18	18	16	88	2

Table 2 summarizes the detailed scoring results of the second-year students' English essays. Because the content domain constituted the majority of the assessment criteria (60 points), students' overall rankings were largely determined by their performance in content-related aspects of writing. Student 4 ranked first with a total score of 97 points, demonstrating exceptional performance across both content and language domains. This outcome reflects the successful integration of logical argumentation, coherent organization, and accurate language use.

Conversely, some students exhibited relatively strong linguistic performance but received lower scores in the content domain, which negatively affected their overall rankings. These results suggest that while language proficiency contributed to writing quality, the ability to develop persuasive arguments and provide logical support was a more influential factor in the final evaluation. Such findings underscore the importance of content development in English essay assessment when content-related criteria are assigned substantial weight.

Table 3. Detailed scoring results for third-year students

Student (3rd Grade)	Content (60 points)			Language Use (40 points)		Total	Rank
	①	②	③	④	⑤		
Student 1	6	20	12	13	13	64	10
Student 2	10	27	18	16	16	87	3
Student 3	10	26	18	17	17	88	2
Student 4	6	20	11	15	15	67	9
Student 5	7	25	15	16	16	79	6
Student 6	9	25	15	17	17	83	5
Student 7	6	15	11	14	14	60	11
Student 8	6	20	13	15	15	69	8
Student 9	9	25	16	17	17	84	4
Student 10	10	26	18	18	17	89	1
Student 11	8	20	14	17	18	77	7
Student 12	8	22	17	18	18	83	5

Table 3 presents the detailed scoring results for the third-year students' English essays. As the same scoring rubric was applied across both grade levels, the third-year results similarly indicate that students' overall rankings were substantially influenced by their performance in the content domain. Student 10 achieved the highest overall ranking, with a total score of 89 points, demonstrating strong performance in both the content and language domains.

Moreover, although some students across the two grade levels received identical total scores, their domain-specific scores differed. This pattern suggests that the essay assessment captured multiple dimensions of writing performance rather than relying solely on an overall score. In this regard, the rubric allowed for a more comprehensive evaluation of students' English writing by considering both content-related qualities, such as organization and argument development, and linguistic features, including accuracy and clarity of expression.

Coh-Metrix 3.0

To quantitatively examine the linguistic features of students' English essays, this study employed Coh-Metrix 3.0, a computational text analysis tool based on natural language processing (NLP). Coh-Metrix was developed to provide multidimensional analyses of textual and discourse characteristics, including cohesion, lexical sophistication, syntactic complexity, and readability. Owing to its ability to generate a wide range of linguistic indices, it has been extensively utilized in discourse analysis and second language (L2) writing research (Graesser et al., 2004).

Among the numerous indices generated by Coh-Metrix, the present study focused on measures of lexical diversity, syntactic complexity, cohesion, and readability, as these dimensions are closely associated with writing quality and correspond to the language-related criteria included in the English Essay Contest Scoring Rubric.

To assess lexical diversity, two indices were selected: LDTTRa (Lexical Diversity Type–Token Ratio, adjusted) and LDTTRc (Lexical Diversity Type–Token Ratio for Content Words). LDTTRa measures the range of vocabulary used throughout a text while controlling for the effects of text length and repetitive word use. LDTTRc specifically captures the diversity of content words, including nouns, verbs, adjectives, and adverbs, thereby providing an indication of learners' ability to express meaning through varied lexical choices.

Syntactic complexity was examined using SYNLE (Syntactic Complexity: Mean Number of Words Before Main Verb). This index reflects the average number of words preceding the main verb in a sentence and serves as an indirect measure of learners' ability to produce syntactically elaborated sentence structures.

To investigate textual cohesion, DRNP (Density of Referential Noun Phrase Overlap) was employed. This index measures the extent to which identical or semantically related noun phrases recur across a text, thereby indicating the degree of referential cohesion and the overall coherence of written discourse.

Readability was assessed using two widely recognized indices: RDFRE (Readability: Flesch Reading Ease) and RDFKGL (Readability: Flesch–Kincaid Grade Level). RDFRE is based on the Flesch Reading Ease formula and incorporates sentence length and syllable counts to estimate the ease with which a text can be read; higher scores indicate greater readability. In contrast, RDFKGL is derived from the Flesch–Kincaid Grade Level formula and estimates the educational grade level required to comprehend a text. Higher RDFKGL values indicate greater linguistic complexity and a higher level of reading proficiency required for comprehension.

Collectively, these indices provided objective measures of linguistic performance and enabled a systematic

examination of the relationship between students' essay scores and the linguistic characteristics of their written production.

Data Analysis

The present study analyzed students' English essays using Coh-Metrix 3.0 to investigate the relationship between the linguistic characteristics of the essays and the evaluation outcomes assigned by non-native Korean English teachers. More specifically, the study sought to examine the extent to which objective linguistic features of students' written texts were associated with teachers' holistic judgments of essay quality. The data analysis procedures were conducted in three stages.

First, each student essay was processed using Coh-Metrix 3.0 to obtain a set of linguistic indices representing lexical diversity, syntactic complexity, cohesion, and readability. The selected indices included overall lexical diversity (LDTTRa) and content-word lexical diversity (LDTTRc) as measures of lexical diversity; the mean number of words before the main verb (SYNLE) as an indicator of syntactic complexity; the density of referential noun phrase overlap (DRNP) as a measure of textual cohesion; and the Flesch Reading Ease score (RDFRE) as an indicator of readability.

Second, the relationships between the Coh-Metrix indices and the essay rankings assigned by the English teachers were examined using IBM SPSS Statistics. Because the teacher evaluation outcomes were expressed as ordinal rankings, Spearman's rank-order correlation coefficient was employed to identify the strength and direction of rank-based associations between the linguistic variables and the teachers' evaluation results.

Third, the correlation results were interpreted to explore how specific linguistic features were associated with teachers' holistic evaluations of students' essays. Particular attention was given to identifying which textual characteristics were most strongly related to higher evaluation outcomes, thereby providing insights into the linguistic factors that may influence teachers' judgments of English writing quality.

Results and Discussions

The results of the Coh-Metrix analysis of students' essays by grade level are presented in Tables 4 and 5. The analysis focused on lexical diversity, syntactic complexity, cohesion, and readability, which were selected as key linguistic features relevant to English essay quality.

Table 4. Coh-Metrix analysis results for second-year students

Student (2nd Grade)	LDTTRa	LDTTRc	SYNLE	DRNP	RDFRE	RDFKGL
Student 1	0.436	0.577	2.304	333.333	82.534	4.632
Student 2	0.403	0.544	4.353	343.75	72.603	6.874
Student 3	0.466	0.555	6.294	268.199	66.467	7.803
Student 4	0.421	0.586	4.364	305.994	66.748	7.529
Student 5	0.372	0.432	3.72	332.326	58.713	8.359
Student 6	0.413	0.578	3.083	338.279	75.581	6.206
Student 7	0.439	0.552	3.524	352.941	62.147	7.206
Student 8	0.564	0.688	4.357	366.337	55.222	9.142
Student 9	0.576	0.681	3.421	314.136	74.807	5.323
Student 10	0.368	0.532	3.103	385.417	71.503	6.344
Student 11	0.433	0.594	3.182	362.989	63.84	7.528

Table 4 presents the Coh-Metrix analysis results for the English essays written by the second-year students. With regard to lexical diversity, Student 8 (LDTTRa = 0.564, LDTTRc = 0.688) and Student 9 (LDTTRa = 0.576, LDTTRc = 0.681) showed relatively high values, suggesting that these students employed a wider range of vocabulary in their essays. In contrast, Student 5 (LDTTRa = 0.372, LDTTRc = 0.432) and Student 10 (LDTTRa = 0.368, LDTTRc = 0.532) demonstrated comparatively lower levels of lexical diversity, indicating a more limited range of lexical choices.

In terms of syntactic complexity, Student 3 recorded the highest SYNLE value (6.294), suggesting that this student may have produced more syntactically elaborated sentence structures. By contrast, students with lower SYNLE values appeared to rely on relatively simpler sentence structures, although further qualitative analysis would be necessary to determine the precise nature of their syntactic patterns.

Regarding cohesion, Student 10 showed the highest DRNP value (385.417), indicating a greater degree of referential noun phrase overlap across the essay. This may suggest stronger referential cohesion through the repeated use of identical or semantically related noun phrases. Conversely, Student 3 displayed a relatively low DRNP value, which may indicate less frequent noun phrase overlap and comparatively weaker referential cohesion.

The readability indices also revealed notable differences among the second-year students. In particular, Student 8 and Student 5 showed relatively high RDFKGL values, suggesting that their essays may have required a higher level of reading proficiency due to longer sentences and/or more complex lexical choices. Overall, these findings indicate that the second-year students' essays varied considerably in terms of lexical diversity, syntactic complexity, cohesion, and readability, highlighting individual differences in their English writing performance.

Table 5 presents the results of the Coh-Metrix analysis for the English essays written by the third-year students. Overall, the third-year students tended to demonstrate higher levels of lexical diversity than the second-year students, suggesting a greater ability to employ a wider range of vocabulary in their written production.

In terms of lexical diversity, Student 7 (LDTTRa = 0.604, LDTTRc = 0.723) and Student 4 (LDTTRa = 0.575, LDTTRc = 0.745) exhibited relatively high values, indicating the use of a broad range of lexical items and content words throughout their essays. In contrast, Student 5 (LDTTRa = 0.427, LDTTRc = 0.551) showed comparatively lower lexical diversity scores, suggesting a tendency to rely on a more restricted range of vocabulary. These findings imply considerable individual variation in lexical resource use among the third-year students.

Table 5. Coh-Metrix analysis results for third-year students

Student (3rd Grade)	LDTTRa	LDTTRc	SYNLE	DRNP	RDFRE	RDFKGL
Student 1	0.538	0.646	5.333	318.75	51.089	9.446
Student 2	0.481	0.574	3.818	334.884	64.242	9.154
Student 3	0.471	0.574	4.2	308.37	77.552	5.262
Student 4	0.575	0.745	5.778	375.691	73.82	7.959
Student 5	0.427	0.551	3.176	330.275	63.45	7.595
Student 6	0.592	0.7	2.143	346.369	59.174	8.182
Student 7	0.604	0.723	2.294	366.864	70.606	5.881
Student 8	0.508	0.632	3.929	372	72.639	7.564
Student 9	0.492	0.573	4.471	297.297	52.647	8.62
Student 10	0.519	0.686	4.136	347.826	61.949	8.255
Student 11	0.598	0.69	2.231	329.167	62.444	6.843
Student 12	0.511	0.687	3.808	375.691	62.758	7.965

Regarding syntactic complexity, Student 4 (SYNLE = 5.778) and Student 1 (SYNLE = 5.333) recorded relatively high SYNLE values, suggesting a greater tendency to produce syntactically elaborated sentence structures. By comparison, Student 6 (SYNLE = 2.143) and Student 11 (SYNLE = 2.231) demonstrated lower values, indicating a preference for relatively less complex sentence constructions.

For the cohesion measure, DRNP, Student 4 (DRNP = 375.691) and Student 12 (DRNP = 375.691) obtained the highest values, indicating a relatively greater degree of referential noun phrase overlap throughout their essays. In contrast, Student 9 (DRNP = 297.297) showed the lowest value, suggesting less frequent referential connections among ideas within the text. These differences may reflect variation in the extent to which students maintained cohesion across their written discourse.

The readability indices also revealed notable differences among the third-year students. Student 2 (RDFRE = 64.242, RDFKGL = 9.154) produced an essay characterized by a relatively high grade-level requirement while maintaining a moderate level of readability. In contrast, Student 3 (RDFRE = 77.552, RDFKGL = 5.262) produced a text that was comparatively easier to read and required a lower level of reading proficiency for comprehension. These findings suggest that students varied considerably in their use of sentence structures and lexical choices, resulting in different levels of textual complexity and readability.

Taken together, the results presented in Tables 4 and 5 indicate that students exhibited distinct strengths across different linguistic dimensions, including lexical diversity, syntactic complexity, cohesion, and readability. Such variation suggests that multiple linguistic features contribute to the quality of English essay writing and may serve as important indicators of students' writing proficiency.

To address the primary purpose of this study—namely, to examine the extent to which quantitative linguistic indices are associated with teachers' holistic evaluations of English essays—the relationships between the Coh-Metrix indices and the rankings assigned by the English teachers were subsequently analyzed using IBM SPSS Statistics. Specifically, Spearman's rank-order correlation analyses were conducted to investigate the associations between the linguistic features identified through Coh-Metrix and the evaluation outcomes. The results of these analyses are presented in Tables 6 and 7.

Table 6. Correlation analysis of second-year students' English essays ($p < .05$)

Variable		Correlation Coefficient (r)	Significance Level (p)
Second-Year Rank	LDTTRa	.72*	.01
	LDTTRc	.20	.55
	SYNLE	.12	.70
	DRNP	.02	.93
	RDFRE	-.09	.77

Table 7. Correlation analysis of third-year students' English essays ($p < .05$)

Variable		Correlation Coefficient (r)	Significance Level (p)
Third-Year Rank	LDTTRa	.55	.06
	LDTTRc	.46	.12
	SYNLE	-.04	.89
	DRNP	.29	.35
	RDFRE	.07	.82

Table 6 presents the results of the correlation analysis between the Coh-Metrix indices and the essay rankings assigned to the second-year students. As shown in the table, a statistically significant positive correlation was found between essay rankings and overall lexical diversity (LDTTRa) ($r = .72, p < .05$). This finding suggests that students who employed a wider range of vocabulary tended to receive higher rankings in the essay evaluation. In other words, lexical diversity appears to have been an important factor associated with teachers' judgments of writing quality among the second-year students.

In contrast, the remaining linguistic indices did not demonstrate statistically significant relationships with essay rankings. Specifically, content-word lexical diversity (LDTTRc; $r = .20, p = .55$), syntactic complexity (SYNLE; $r = .12, p = .70$), referential noun phrase overlap (DRNP; $r = .02, p = .93$), and readability (RDFRE; $r = -.09, p = .77$) showed only weak associations with evaluation outcomes. These findings suggest that, for the second-year students, teachers' evaluations were more closely related to the breadth of vocabulary use than to syntactic complexity, textual cohesion, or readability.

Table 7 presents the corresponding results for the third-year students. Similar to the findings for the second-year group, overall lexical diversity (LDTTRa) exhibited the strongest positive association with essay rankings ($r = .55, p = .06$). Although this relationship did not reach statistical significance, the magnitude of the correlation suggests a moderate positive tendency between lexical diversity and teachers' evaluations. Content-word lexical diversity (LDTTRc) also showed a moderate positive correlation ($r = .46, p = .12$), whereas syntactic complexity (SYNLE; $r = -.04, p = .89$), referential noun phrase overlap (DRNP; $r = .29, p = .35$), and readability (RDFRE; $r = .07, p = .82$) demonstrated weak or negligible relationships with essay rankings.

Taken together, the findings from both grade levels indicate that lexical diversity was the linguistic feature most consistently associated with teachers' holistic evaluations of students' essays. In particular, the positive correlations observed for LDTTRa suggest that teachers may place considerable value on students' ability to employ a varied vocabulary when judging overall writing quality. This finding is consistent with previous L2 writing research reporting that lexical richness is often perceived as an important indicator of language proficiency and writing competence.

By contrast, syntactic complexity, cohesion, and readability showed relatively weak relationships with evaluation outcomes across both grade levels. These results suggest that such linguistic features may not be directly reflected in teachers' holistic judgments or may interact with other aspects of writing quality, such as content development, organization, and argumentation. Given that the essay scoring rubric allocated 60% of the total score to the content domain, it is plausible that teachers' evaluations were influenced by multiple dimensions of writing performance rather than by any single linguistic feature. Therefore, while lexical diversity appears to be a salient factor in essay assessment, the overall quality of English essays is likely to be determined by a combination of linguistic and content-related characteristics.

Conclusions and Implications

This study examined the extent to which quantitative linguistic indices generated by Coh-Metrix were associated with holistic evaluations of English essays conducted by non-native Korean English teachers. By doing so, the study explored the potential of automated language analysis tools as a means of supporting, rather than replacing, teacher-based assessment in English essay evaluation.

The findings revealed that overall lexical diversity (LDTTRa) was the linguistic feature most consistently associated with teachers' evaluation rankings. In the second-year students' essays, LDTTRa showed a strong positive correlation with teacher-assigned rankings ($r = .72, p < .05$). In the third-year students' essays, the same index also demonstrated a relatively strong positive association ($r = .55$), although the correlation did not reach statistical significance. In contrast, the other indices, including syntactic complexity, cohesion, and readability, showed relatively weak relationships with teacher evaluations across both grade levels.

These results suggest that students' ability to use a wide range of vocabulary may function as a salient cue in teachers' holistic judgments of English writing quality. Although the scoring rubric formally included content-related criteria such as logical organization, argument development, and support for claims, lexical diversity appeared to be closely associated with actual evaluation outcomes. This interpretation is plausible in school-based assessment contexts, where teachers are often required to read and compare numerous essays within a limited period of time. Under such conditions, visible linguistic features, particularly vocabulary range, may contribute to teachers' overall impressions of writing quality.

Lexical diversity may serve as an immediately recognizable indicator of learners' English proficiency and expressive ability. Even when students address the same essay topic with similar arguments, essays that rely on repetitive or limited vocabulary may be perceived as less sophisticated or less effective. By contrast, essays that present ideas through varied lexical choices, appropriate paraphrasing, and more advanced vocabulary may be evaluated more positively in terms of overall quality and completeness. Thus, even when teachers intend to focus on content-related aspects of writing, lexical diversity may still influence their holistic evaluations.

At the same time, the relatively weak correlations between teacher evaluations and the indices of syntactic complexity, cohesion, and readability indicate that English essay assessment cannot be explained solely by formal linguistic features. Rather, essay evaluation is a multidimensional process that involves the interaction of various factors, including logical organization, content development, appropriateness of ideas, textual coherence, and language use. This tendency was particularly evident in the third-year essays, where students displayed more varied patterns across linguistic features. As learners advance in grade level and writing proficiency, it becomes increasingly difficult to account for writing quality using a single linguistic index. Teachers' evaluations therefore appear to reflect broader judgments of how effectively students construct, develop, and communicate their ideas in written English.

Based on these findings, it would be premature to conclude that automated scoring or language analysis systems can fully replace teachers' holistic evaluations in English essay assessment. Rather, such systems appear to have greater potential as supplementary tools that can enhance the reliability, consistency, and transparency of assessment practices. In school contexts, automated language analysis tools such as Coh-Metrix can provide objective numerical information about students' vocabulary use, syntactic features, cohesion, and readability. These data may help reduce teachers'

assessment burden, mitigate potential subjectivity in scoring, and support more evidence-based evaluation practices.

Pedagogically, automated language analysis tools may be used in several ways. First, they can serve as preliminary diagnostic tools for identifying the linguistic features of students' essays before teacher evaluation. Second, they can support feedback provision by offering objective information on students' lexical diversity, sentence complexity, cohesion, and readability. Third, they can be used for formative assessment by tracking changes in students' English writing development over time. In this sense, automated analysis tools can contribute not only to assessment efficiency but also to students' awareness of specific linguistic aspects of their writing.

Accordingly, future approaches to English essay assessment should consider hybrid evaluation models that integrate automated linguistic analysis with teacher judgment. Such models may allow teachers to make more informed and balanced evaluations by combining objective textual indices with professional interpretations of content, organization, and communicative effectiveness. Further research is also needed to examine the applicability of automated scoring and language analysis tools using larger samples of learner essays across different grade levels, proficiency groups, and writing task types. These efforts would contribute to a more comprehensive understanding of how automated analysis can be meaningfully incorporated into English writing assessment and instruction.

Conflicts of Interest

The author(s) declared no conflicts of interest.

References

- Brown, H. D., & Abeywickrama, P. (2019). *Language assessment: Principles and classroom practices* (3rd ed.). Pearson Education.
- Cooper, C. R. (1977). Holistic evaluation of writing. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging*. National Council of Teachers of English, 3-31.
- Crossley, S., & McNamara, D. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 32(32), 984-989.
- Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment*, 5(1).
- Dikli, S., & Bley, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing*, 22, 1-17. <https://doi.org/10.1016/j.asw.2014.03.006>
- Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, 4(2), 139-155. [https://doi.org/10.1016/1060-3743\(95\)90004-7](https://doi.org/10.1016/1060-3743(95)90004-7)
- 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. <https://doi.org/10.3758/BF03195564>
- Graham, S., & Perin, D. (2007). *Writing next: Effective strategies to improve writing of adolescents in middle and high schools: A report to Carnegie Corporation of New York*. Alliance for Excellent Education.
- Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman.
- Hughes, A., & Hughes, J. (2020). *Testing for language teachers* (3rd ed.). Cambridge University Press. <https://doi.org/10.1017/9781009024723>

- Hyland, K. (2003). *Second language writing*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511667251>
- Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing*, 12(4), 377-403. <https://doi.org/10.1016/j.jslw.2003.09.001>
- Kwon, O. R. (2010). The teachers' perceptions and psychological burdens involved in the implementation of the speaking/writing tests of the National English Ability Test. *Foreign Language Education Research*, 13, 64-88.
- Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, 34, 12-24. <https://doi.org/10.1016/j.jslw.2016.10.003>
- Lee, S. G., Lee, Y. J., Hwang, E. G., Lee, J. W., Bae, J. G., Kim, J. O., Lee, S. E., & Lee, D. J (2017). Practical guidelines for English test development. Hankook Munhwasa.
- Lee, H. (2019). Analyzing L2 essay with FACETS and Coh-Metrix in L2 writing assessment. *Korean Journal of English Language and Linguistics*, 19(4), 796-816. <https://doi.org/10.15738/kjell.19.4.201912.796>
- Lee, I. (2002). Teaching coherence to ESL students: A classroom inquiry. *Journal of Second Language Writing*, 11(2), 135-159. [https://doi.org/10.1016/S1060-3743\(02\)00065-6](https://doi.org/10.1016/S1060-3743(02)00065-6)
- Lee, S. H. (2013). A corpus-based analysis of the difficulty of the reading materials in High School English 1 and 2 textbooks using Coh-Metrix. *Studies in Foreign Language Education*, 27(2), 131-148. <https://doi.org/10.16933/sfle.2013.27.2.131>
- Lee, S. Y. (2022). Comparison between native and non-native English professors on L2 writing assessment: A study of assessment areas and perceptions of writer identity. *Journal of Learner-Centered Curriculum and Instruction*, 22(21), 785-798. <https://doi.org/10.22251/jlcci.2022.22.21.785>
- Lee, S. Y., & Choi, H. S. (2012). Rating of Korean students' L2 writing: Similarities and differences between native and non-native raters. *Journal of Research in Curriculum Instruction*, 16(3), 629-655. <https://doi.org/10.24231/rici.2012.16.3.629>
- Li, J., & Huang, J. (2022). The impact of essay organization and overall quality on the holistic scoring of EFL writing: Perspectives from classroom English teachers and national writing raters. *Assessing Writing*, 51, 100604. <https://doi.org/10.1016/j.asw.2021.100604>
- Lim, H. G., Park, D. Y., & Si, G. J. (2014). Sophistication of an automated scoring system for large-scale essay writing tests. *Multimedia-Assisted Language Learning*, 17(1), 84-105. <https://doi.org/10.15702/mall.2014.17.1.84>
- 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>
- 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>
- Park, J. G. (2006). The study on the rater reliability of three scoring methods in assessing argumentative essays: Holistic, analytic, and multiple-trait scoring methods. *Foreign Language Education Research*, 9, 63-84.
- Park, K., Lee, Y., & Shin, D. (2021). Exploring the feasibility of an automated essay scoring model based on LSTM. *Journal of Curriculum Evaluation*, 24(4), 223-238. <https://doi.org/10.29221/jce.2021.24.4.223>
- Park, Y. H., Min, H. G., Kim, J. S., & Kim, Y. H. (2015). An investigation into the KICE writing assessment software for high schools: Teacher and student perceptions. *Studies in English Education*, 20(3), 195-237.

- Si, G. J., Lee, Y. S., Park, D. Y., Lim, H. G., Koo, S. G., Park, S. W., & Lim, E. Y. (2013). Exploring the applicability of automated scoring in large-scale English writing assessment in Korea. *Journal of Educational Evaluation*, 26(2), 319–345.
- Shin, D. G. (2022). Effects of scoring features on the accuracy of the automated scoring model of English. *Teacher Education*, 38(6), 73-91. <https://doi.org/10.14333/KJTE.2022.38.6.04>
- So, Y. S. (2023). Investigating performance and essay assessment practices in Korean high schools: Current state and improvement strategies. *English Language Assessment*, 18(2), 115–137. <https://doi.org/10.37244/ela.2023.18.2.115>
- Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511732997>
- Wilson, J., & Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research*, 58(1), 87-125. <https://doi.org/10.1177/0735633119830764>

Authors' Information

Choi, Min-Seok: Korea National University of Education, Graduate Student, First Author.

Lee, Dong Ju: Korea National University of Education, Professor, Corresponding Author. <https://orcid.org/0000-0002-9906-835X>