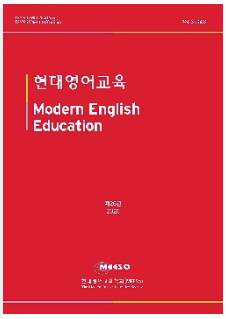
Modern English Education, Vol. 26, (2025) pp. 68-83
doi: https://doi.org/10.18095/meeso.2025.26.1.68

MEESO

Modern English Education

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



The Relationship Between Text Readability, Item Types, and Their Impact on Student Performance on CSAT English Reading Items

Jungran Kim (Gimhae Foreign Language High School / Pusan National University)
YeonJoo Jung (Pusan National University)

Received: 4 January 2025
Revised: 26 January 2025
Accepted: 30 January 2025

Kim, Jungran, & Jung, YeonJoo. (2025). The relationship between text readability, item types, and their impact on student performance on CSAT English reading items. *Modern English Education*, *26*, 68-83.

Keywords
CSAT, ARTE, readability, incorrect response rates
수학능력시험, 아르테, 이독성, 오답률

Abstract

Despite extensive research on readability formulas in language assessment, their relationship with student performance in high-stakes English as a Foreign Language (EFL) testing remains understudied. This study investigated the relationship between text readability and student performance across four item categories in English reading comprehension section of the Korean College Scholastic Ability Test (CSAT). Analysis was conducted on 335 reading passages from CSAT examinations (2017–2024) using four readability formulas (two traditional and two advanced). Results revealed varying patterns of incorrect response rates across item categories. Main Idea Comprehension items showed the most consistent performance patterns (CV = 0.137) while Discourse Structure Inference items displayed the highest variability (CV = 0.165). Correlation analyses indicated weak relationships between readability measures and incorrect response rates, with contextual inference items showing weak but significant correlations with certain readability indices. Subsequent multiple regression analyses revealed that readability measures explained very small or small amounts of variance in incorrect response rates across item categories (R^2 = 0.015–0.091). Findings of this study are discussed in terms of developing differentiated approaches to difficulty calibration tailored to each item category's distinct linguistic features and cognitive demands.

Jungran Kim
(First author)
PhD Candidate
Department of English Education
Pusan National University
kjr1021@kakao.com

YeonJoo Jung
(Corresponding author)
Professor
Department of English Education
Pusan National University
yjjung@pusan.ac.kr

* This work was supported by a Journal Research Promotion (2024) for Graduate Students of Pusan National University.

INTRODUCTION

The difficulty of reading comprehension tests has long been understood to be influenced by item difficulty as well as text difficulty (Alderson, 2000). Research on text readability has evolved significantly over recent decades, progressing from surface-level analysis to more sophisticated approaches. Early studies focused on quantifiable features like sentence length and word complexity, establishing fundamental metrics for assessing text difficulty (Chall & Dale, 1995; Flesch, 1948; Kincaid et al., 1975; Smith & Senter, 1967; McLaughlin, 1969). These traditional approaches have since been enhanced by advanced computational methods that incorporate multiple dimensions of text complexity, including lexical diversity, syntactic complexity, and text cohesion (Benjamin, 2012; Choi & Crossley, 2022a; Crossley et al., 2014, 2017, 2019, 2023; Crossley & McNamara, 2008; McNamara et al., 2014). The development of tools like the Automatic Readability Tool for English (ARTE; Choi & Crossley, 2022a, 2022b) has made these sophisticated measures more accessible to researchers and practitioners.

While these readability measures have been extensively studied in various educational settings, from classroom material selection to standardized testing and second language (L2) learning contexts (e.g., Crossley et al., 2023; Hwang & Lee, 2020a, 2020b, 2020c; Joo, 2017; Nahatame, 2021; Yum & Kim, 2023), their application in EFL assessment contexts, particularly in Korean high school English assessment such as the College Scholastic Ability Test (CSAT), a high-stakes examination that plays a crucial role in determining Korean students' university admission, requires further investigation. The CSAT English section encompasses various types of reading comprehension items, from basic comprehension to higher-order thinking skills (Kang et al., 2021). Research has demonstrated that these different item types significantly influence how students process and comprehend texts (Chang, 2004; Kwon & Lee, 2015), even having a greater impact on assessment outcomes than text characteristics themselves (Ji & Kim, 2014). Given that the CSAT employs an absolute evaluation method in English (Ministry of Education, 2014), maintaining consistent difficulty levels across different test administrations becomes paramount as scores directly reflect achievement levels rather than relative standing. Among various factors affecting test difficulty, text readability and item types warrant particular attention. These two factors are the primary controllable elements in test development, as opposed to other variables such as student characteristics or testing conditions; reading test difficulty is fundamentally determined by the interplay between text difficulty and item difficulty (Alderson, 2000).

This study aims to enhance our understanding of the relationship between text readability, item categories, and student performance in the CSAT English reading assessment through two main investigations. First, it examines the consistency of incorrect response rates across different categories of reading comprehension items. Second, it explores how various readability formulas predict student incorrect response rates in these different item categories. Through this analysis, this study seeks to contribute to more informed approaches to test development and difficulty control in the specific context of CSAT English assessment.

LITERATURE REVIEW

Text Readability in Language Assessment

Text readability refers to the extent to which a written text can be comprehended and processed efficiently by its readers based on its linguistic features (Dale & Chall, 1949; Richards & Schmidt, 2013). In language assessment, it serves as one of the text variables that predict item difficulty, incorporating multiple dimensions including lexical sophistication, topic unfamiliarity, requisite background knowledge, and structural complexity (Hwang & Lee, 2020a). Text readability has been widely used as a key indicator by teachers and test developers in selecting appropriate reading materials that match students' reading abilities (Crossley et al., 2017, 2023). While various readability formulas have been developed over the decades, their practical applications in educational settings have demonstrated both possibilities and limitations. Traditional readability measures such as Flesch Reading Ease formula (FRE; Flesch, 1948), Flesch-Kincaid Grade Level formula (FKGL; Kincaid et al., 1975), and Dale-Chall formula (Chall & Dale, 1995; Dale & Chall, 1948) have primarily relied on quantifiable surface-level text features (Benjamin, 2012; Flesch, 1948). Recent empirical research has highlighted the limitations of these traditional measures. For example, Song (2021) analyzed Korean high school English textbooks and found that readability scores alone failed to fully capture the complexity differences between textbook levels. The study revealed that texts with similar vocabulary levels showed significant differences in syntactic structure, demonstrating the necessity of considering both lexical and syntactic features in text difficulty evaluation.

These traditional formulas demonstrated limited construct validity in capturing the multifaceted nature of text complexity including cohesion, semantics, and text structure (Crossley et al., 2017, 2019, 2023; McCarthy & McNamara, 2021), particularly in L2 learning contexts where reading comprehension extends beyond surface-level features (Green et al., 2010; Joo, 2017). Song's (2021) findings further supported this limitation, showing how syntactic complexity measures revealed hierarchical differences between textbook levels that were not apparent from traditional readability indices alone. Furthermore, Nahatame's (2021) eye-tracking study compared the predictive power of traditional formulas such as FRE, New-Dale Chall formula, and newer formulas such as Crowdsourced Algorithm of Reading Comprehension (CAREC), Crowdsourced Algorithm of Reading Speed (CARES), and Coh-Metrix L2 Reading Index (CML2RI) against L2 readers' processing effort. The study demonstrated that while newer readability formulas showed better predictive power than traditional ones, no single formula consistently predicted L2 readers' processing effort. This finding is particularly significant given that L2 readers exhibit distinct reading patterns, including longer reading times, more fixations, and lower word-skipping probabilities compared to L1 readers. The effectiveness of CAREC and CARES was further validated by Crossley et al. (2019), which demonstrated that these algorithms, developed through crowdsourced human judgments, significantly outperformed traditional readability measures in predicting both text comprehension and reading speed.

Currently, a variety of readability formulas, ranging from traditional to current metrics, can be automatically calculated through the Automatic Readability Tool for English (ARTE; Choi & Crossley, 2022a, 2022b). This tool incorporates traditional measures such as FRE (Flesch, 1948), FKGL (Kincaid et al., 1975), Automated Readability Index (ARI; Smith & Senter, 1967), SMOG Grading (McLaughlin, 1969), and New Dale-Chall Readability Formula (Chall & Dale, 1995), as well as advanced approaches including CML2RI (approximated) (Crossley & McNamara, 2008; Crossley et al., 2019), CAREC (Crossley et al., 2019), Crowdsourced Algorithm of Reading Comprehension Modified (Crossley et al., 2019), and CARES (Crossley et al., 2019). Among traditional formulas, FRE (Flesch, 1948) and New Dale-Chall Readability Formula (Chall & Dale, 1995) stand out for their widespread use. The FRE measures readability based on sentence length and word length, while the New Dale-Chall Readability Formula uniquely incorporates a list of familiar words in its calculations. More recent approaches have evolved to address the limitations of traditional measures. The CML2RI (Crossley & McNamara, 2008) specifically targets L2 learners by analyzing word overlap, sentence syntactic similarity, and word frequency. CAREC and CARES (Crossley et al., 2019) represent innovative approaches that integrate multiple linguistic features and crowd-sourced evaluations. These newer formulas make use of advanced natural language processing techniques to capture complex linguistic features that were previously unmeasurable, such as lexical diversity indices, semantic similarity between sentences, discourse connective patterns, and psycholinguistic properties of words (e.g., word concreteness, imageability, and age of acquisition).

While significant advances have been made in readability research, there remains room for improvement. Traditional formulas continue to be widely used in practice despite their limitations, largely because newer formulas are more difficult to interpret, access, and have limited exposure. Tools like ARTE aim to address these challenges by making state-of-the-art readability formulas more accessible to educational content creators and researchers, and further empirical research using those tools across diverse contexts will be valuable in establishing their broader applicability and effectiveness.

Item Difficulty and Student Performance

While these advances in readability research provide valuable insights for text selection and modification, understanding how different types of assessment items interact with text complexity is equally crucial for effective reading assessment design. The interaction between item types and student performance represents a critical area in reading assessment research (Alderson, 2000). Cohen and Upton's (2007) analysis of test-taker responses to the new TOEFL reading section revealed distinct strategy use patterns across different item types. While examinees used similar reading strategies across basic comprehension, inferencing, and reading-to-learn items, they engaged different test-management strategies specific to each item type. Their study found that reading-to-learn items, though intended to measure higher-level comprehension skills, proved easier (90% success rate) compared to basic comprehension (83%) and inferencing items (77%). This unexpected result challenges conventional assumptions about item difficulty levels, suggesting the need for careful examination of item type characteristics in high-stakes testing contexts.

Research examining the relationship between linguistic features and item characteristics has revealed systematic patterns across question types in EFL contexts. According to Kang et al. (2021), CSAT consists of items designed to measure higher-order thinking skills necessary for university study, and, specifically, the CSAT English reading section comprises seven-item categories: main idea identification, detail comprehension, context understanding, interactive context understanding, indirect writing, language form and vocabulary, and extended passage comprehension; each category contains multiple item types, with each category requiring its own distinctive cognitive processes and strategic approaches. The difficulty level across these item categories is influenced by external textual factors such as the degree of inference required for problem-solving and the unfamiliarity of item types as well as the attractiveness of answer choices (Chang, 2004). Ji and Kim's (2014) study, which administered three repeated assessments using the same English passages transformed into different item types, revealed that item type characteristics had a greater impact on reading assessment than text familiarity. This item-dependent nature of reading assessment was further evidenced by Kwon and Lee's (2015) research, which found that students employ different reading strategies depending on the item type they encounter. For instance, Kwon and Lee (2015) found that for gap-filling items, students utilized keywords (19.8%) and contextual vocabulary (6.3%), while for sentence ordering items, they focused on discourse markers such as connectives, pronouns, and repeated phrases rather than content words (2.5%). For main idea identification items, students employed both key words (16.7%) and contextual vocabulary (9.0%) but approached the text more selectively. Text organization is another variable which affects student performance differentially across different testing formats and learner proficiency levels. Kobayashi (2002) found that text organization had minimal impact on cloze test performance among lower proficiency EFL learners. However, when testing formats measured more integrative comprehension abilities, particularly among higher proficiency EFL learners, organizational differences in texts produced significant performance variations. This interaction effect between text organization and proficiency level is particularly relevant in the CSAT context, where items assess various levels of comprehension ability.

These findings collectively suggest that different item types may draw on distinct linguistic competencies and cognitive processes. In the context of high-stakes tests like CSAT, understanding these interaction patterns becomes crucial for maintaining consistent difficulty levels across different test administrations. Moreover, the absolute evaluation system employed in CSAT English section necessitates particular attention to how different item types interact with text features to affect student performance.

The Present Study

Research on readability in language assessment and item difficulty in reading comprehension has progressed over the years, yet several research gaps remain, which warrant further investigation. Despite the theoretical advances in readability assessment discussed by Crossley et al. (2017, 2019), the relationship between readability formulas and actual student performance in EFL assessment contexts, especially how different readability formulas predict student performance in high-stakes testing environments, remains understudied. Additionally, there is a lack of research investigating whether items within the same category demonstrate consistent incorrect response rates across different test administrations. Prior research has offered valuable insights into these aspects separately. For instance, Kwon and Lee's (2015) findings highlighted different linguistic competencies required for various item types, but their study did not examine whether these patterns remain consistent across different test administrations. Similarly, although Kobayashi (2002) demonstrated significant interaction between text organization and testing formats, a systematic investigation linking readability features to student performance in high-stakes EFL tests is still needed. Recent empirical analyses using actual student performance data have emerged (Hwang & Lee, 2020a, 2020b, 2020c), but few studies have comprehensively examined both the consistency of incorrect response rates and the predictive power of various readability formulas across different item categories.

Given these research gaps, particularly regarding the consistency of student performance patterns and the predictive power of readability formulas, this study aims to investigate the relationship between readability formulas and incorrect response rates in the English reading comprehension items of CSAT. Using ARTE, we analyzed both the consistency of incorrect response rates and their relationship with various readability metrics. Specifically, this study addresses the following research questions:

(1) Are the student incorrect response rates constant across different reading item categories in the CSAT English section?

(2) How do readability formulas predict student incorrect response rates for CSAT English reading items?

METHOD

Structure and Development of CSAT English Section

CSAT, administered by the Korea Institute for Curriculum and Evaluation (KICE), has been Korea's standardized college entrance examination since its first implementation in 1994. This national-level assessment is conducted once annually in

November and serves as a crucial determinant for college admissions. CSAT consists of six subject areas: Korean Language Arts, Mathematics, English, Korean History, Electives (including Social Studies/Science/Vocational Education), and Second Foreign Languages/Chinese Characters. The English section consists of two parts: listening comprehension (37 points, 17 items) and reading comprehension (63 points, 28 items). The reading section consists of multiple-choice items with five options and plays a particularly significant role in determining students' overall English scores. The reading comprehension section comprises 25 passages in total: twenty-three short passages with one item each, one longer passage with two associated items, and one extended passage with three related items. In addition to the annual CSAT, KICE administers two national-level mock tests exclusively for third-year high school students each year (in June and September). These mock examinations mirror the format of the actual CSAT, allowing students to gauge their preparedness while providing valuable data for test developers to set appropriate difficulty levels for the upcoming CSAT in November. KICE implemented an absolute grading system for the November 2017 CSAT (for 2018 college admission) for the English section, replacing the previous relative grading system. According to the Ministry of Education (2014), this change aimed to reduce excessive competition for marginal score differences while promoting communication-focused English education in schools. Under this system, students receive grades based on predetermined score thresholds rather than their relative performance rankings. The grading system change required maintaining consistent item difficulty across different administrations, as score distributions directly indicate students' absolute performance levels. The present study focused on English reading items from both the CSAT and its mock tests from 2017 to 2024, during which the absolute grading system has been in effect.

Corpus Data

The corpus data for this study comprised two components: the reading passages from the English section of CSAT and their corresponding incorrect response rates. The reading passages were collected from both the CSAT mock examinations and actual CSAT examinations administered by KICE from 2017 (when absolute grading was introduced) through June 2024. These passages were obtained from the 'Download Past Test Questions for High School Seniors' section of EBSi (Korea Educational Broadcasting System, 2024a), while the incorrect response rates were sourced from EBSi's 'Top 15 Historical Grade Cutoffs/Incorrect Response Rates' service (Korea Educational Broadcasting System, 2024b). This service presents incorrect response rates for the 15 most challenging items from each examination, based on student-submitted answers. The number of students who submitted responses ranged from 98,962 (for the June 2023 Mock Exam) to 177,371 (for the June 2024 Mock Exam). As this study aimed to analyze the readability and linguistic features of reading passages, the listening comprehension section (Questions 1–17) was excluded. Furthermore, only items with available incorrect response rates were included in the analysis to enable comparison between text readability and student incorrect response rates. Long reading passages containing two different items (title inference and vocabulary inference items) were analyzed using both item types. Long passages with three different items (Questions 43-45) typically test sequence organization, reference inference, and specific details in narrative texts. These passages were excluded from the analysis as they have not been included in the top 15 items with the highest incorrect response rates provided by EBSi. Consequently, the final corpus comprised 335 passages from 23 examinations. Table 1 shows how reading passages were classified into different item categories.

TABLE 1
Classification of Reading Passages by Item Category

Category	Item Type	Number of Items	Subtotal	Total
Main Idea Comprehension	Identifying Main Argument	3	56	335
	Identifying Central Theme	19		
	Identifying Title	34		
Contextual Inference	Interpreting Implied Meanings	19	110	
	Inferring Text Completion	91		
Language Component Analysis	Analyzing Grammatical Structures	21	58	
	Assessing Lexical Appropriateness	37		

Category	Item Type	Number of Items	Subtotal	Total
Discourse Structure Inference	Finding an Irrelevant Sentence	5	111	
	Ordering sentences	44		
	Inserting a sentence	42		
	Completing Summary	20		

Note. Each Item was coded as YY_MM_IT where YY indicates year, MM indicates month, and IT indicates item type (e.g., 17_06_22 represents an item identifying a main argument from June 2017).

The passages were categorized into four main categories (Main Idea Comprehension, Contextual Inference, Language Component Analysis, and Discourse Structure Inference), following the classification schemes proposed by Kang et al. (2021) and Yum and Kim (2023). Each item type within these categories was identified by specific item numbers in the test: Main Idea Comprehension included identifying main argument (22), central theme (23), and title (24, 41); Contextual Inference comprised interpreting implied meanings (21) and text completion (31–34); Language Component Analysis consisted of analyzing grammatical structures (29) and assessing lexical appropriateness (30, 42); and Discourse Structure Inference included finding an irrelevant sentence (35), ordering sentences (36, 37), inserting a sentence (38, 39), and completing summary (40). For analyzing the readability and linguistic features of the reading passages, all texts were compiled into separate text files, following the data processing methodology established by Yum and Kim (2023). The preprocessing procedures were implemented as follows:

1. All multiple-choice identifiers and markers were removed from the main text.
2. In items designed to test grammar or vocabulary errors, the incorrect forms were corrected according to their answer keys.
3. For text completion questions, the correct answers were inserted to create complete passages.
4. Irrelevant sentences were eliminated from the passages.
5. For items testing paragraph ordering, the text was reorganized to reflect the correct sequential order.
6. In sentence insertion items, the target sentence was positioned in its correct location to create a coherent passage.

The analysis was conducted after implementing these data cleaning procedures to ensure consistency and accuracy in the assessment of readability and linguistic features.

Readability Formula Selection in ARTE

This study employed ARTE (Version 1.1; Choi & Crossley, 2022a, 2022b) to analyze text complexity through multiple formulas, ranging from traditional to newer approaches. ARTE's web-based platform and an Application Programming Interface (API) facilitate the application of both traditional and contemporary readability formulas, making sophisticated text analysis more accessible to educators and researchers (Choi & Crossley, 2022a). From the multiple formulas available in ARTE, four formulas were selected for calculating the readability scores for each text: two traditional formulas – Flesch Reading Ease (FRE) and New Dale-Chall (NDC) – and two newer formulas – the Coh-Metrix L2 Reading Index (CML2RI) and the Crowdsourced Algorithm of Reading Comprehension (CAREC). Detailed information about each formula is presented in Table 2 (Choi & Crossley, 2022b).

TABLE 2
Overview of Readability Formulas Used in the Analysis

Model	Formula	Features
FRE	206.835 – 0.836(totalSyllables/totalWords) – 1.015(totalWords/totalSentences)	Uses surface-level linguistic features: average syllable length (word difficulty) and sentence length (syntactic complexity).
NDC	3.6365 + 0.0496(totalWords/totalSentences) + 0.1579(PDW)	Uses a wordlist to determine PDW (Percentage of Difficult Words). The constant (3.6365) is only added when PDW exceeds 5%.

Model	Formula	Features
CAREC	1.811 + 0.022(Age) + 0.746(BigramR) - 0.742(TrigramP) – 0.001(Image) + 0.000625(Freq) - 0.699(TTR) – 0.111(ParaO) – 2.067(TempC) + 0.035(NounP) + 0.002(CWTypes) – 0.08(PosAdj) + 0.047(WordL) – 0.395(CharE)	Combines 13 linguistic features including word acquisition age, n-gram measures, text cohesion, and lexical diversity.
CML2RI	-43.142 + 0.642(NumSent) + 12.671(SUBTLEXfreq) + 29.619(LemmaOverlap)	Measures text complexity using sentence count, SUBTLEXus word frequency, and cross-sentence noun/pronoun overlap.

According to Choi and Crossley (2022a), the FRE and the NDC formulas are the most widely adopted traditional formulas in educational contexts due to their straightforward interpretation and established use. FRE (Flesch, 1948) evaluates text difficulty using two primary components: word-level difficulty (the number of syllables per word) and structural difficulty (the number of words per sentence). FRE rates text on a 100-point scale, with higher scores indicating easier readability. While this approach has limitations in capturing text complexity comprehensively, its clear interpretability has contributed to its continued use in educational contexts (Choi & Crossley, 2022a). In contrast to FRE's focus on syllable and sentence length, the NDC formula (Dale & Chall, 1948; Chall & Dale, 1995) assesses text difficulty by identifying the proportion of challenging words using a predefined wordlist. The formula incorporates a threshold effect, where the constant is added only when the percentage of difficult words exceeds 5% of the total text. This reflects the non-linear nature of text comprehension, as the presence of challenging words beyond this threshold significantly increases the perceived difficulty level. The scores generated by this formula can be converted into grade level ranges, with higher scores indicating greater text complexity (Dale & Chall, 1948).

The selection of newer formulas was informed by empirical evidence regarding their theoretical validity in measuring L2 reading processing effort (Nahatame, 2021). The CAREC examines text complexity through 13 key linguistic dimensions, based on crowdsourced comprehension judgments using Bradley-Terry model (Crossley et al., 2019). These include content word diversity and lexical characteristics (such as age of acquisition, word occurrence patterns, mental imagery evocation potential, and information distribution measured by character entropy). The tool also evaluates sequential word patterns through n-gram analysis (incorporating the distribution of two-word combinations, relative frequency of three-word sequences, and trigram type-token ratios). Additionally, it assesses text coherence by examining word repetition at both paragraph and sentence levels, while also considering the emotional tone of the text through the presence of positive descriptors. These components were validated through a comprehensive regression analysis that demonstrated their predictive power in assessing text comprehension difficulty (Crossley et al., 2019). The CAREC formula assigns higher scores to more challenging texts and lower scores to more accessible ones, with the ratio of temporal connectives (TempC) and bigram range patterns (BigramR, measuring word co-occurrence) emerging as the strongest predictors of text comprehension difficulty (Choi & Crossley, 2022b). The CML2RI, developed specifically for L2 learners, analyzes word frequency, sentence syntactic similarity, and word overlap between sentences, with higher scores indicating easier-to-read texts (Crossley et al., 2008). The original CML2RI formula incorporates content word overlap (CWO), sentence syntax similarity (SSS), and CELEX word frequency (CF) (Choi & Crossley, 2022a). For web-based implementation, ARTE (Choi & Crossley, 2022b) provides a modified version using different parameters: the number of sentences in the text, the average frequency score based on SUBTLEXus for content words, and the proportion of noun and pronoun lemma types occurring in consecutive sentences. This modified version maintains the core focus on text cohesion and frequency measures, with particular emphasis on tracking lexical connections across consecutive sentences through noun and pronoun lemma types, as indicated by the largest coefficient (29.619) in the formula.

Data Analysis

To examine the consistency of student incorrect response rates (RQ1) and their relationship with readability formulas (RQ2), analyses were conducted using R statistical software (Version 4.2.2; R Core Team, 2024). Preliminary analyses examined the distributions of variables using Shapiro-Wilk tests and Q-Q plots to verify normality assumptions. The results confirmed normal distributions for all variables ($p > .05$), with skewness values across item categories remaining within a narrow range (-0.43 to 0.41), demonstrating nearly symmetrical distributions.

To investigate relationships between readability formulas and incorrect response rates, Pearson correlation analyses were performed between incorrect response rates and each readability measure (FRE, NDC, CAREC, and CML2RI), as well as among the readability measures themselves. The strength of correlations was evaluated using Cohen's (1988) criteria: weak association ($0.100 \leq |r| < 0.300$), moderate association ($0.300 \leq |r| < 0.500$), and strong association ($|r| \geq 0.500$). Multiple regression analyses were then conducted to examine the combined predictive power of readability measures for each item category. The car package (Fox & Weisberg, 2019) was used for regression diagnostics and the lmtest package (Zeileis & Hothorn, 2002) was employed to calculate *p* values from the models. Separate models were created with incorrect response rates as the dependent variable and four readability measures (FRE, NDC, CAREC, and CML2RI) as predictors. For each regression model, diagnostic tests were performed to check assumptions: multicollinearity was assessed using Variance Inflation Factors (VIF), residual normality was examined using Shapiro-Wilk test, and homoscedasticity was evaluated using Breusch-Pagan test. For all statistical analyses, significance level was set at $p < .05$.

FINDINGS AND DISCUSSION

Consistency of Student Incorrect Response Rates (IRR)

This section addresses the first research question by examining whether student IRR remained constant across different categories in the CSAT English section. The preliminary statistical analysis examined the normality of distributions for IRR across different categories using the Shapiro-Wilk test. The results indicated that the distributions of IRR across all categories were normal ($p > .05$). The skewness values for IRR across different categories were close to 0 (ranging from -0.03 to 0.33), indicating nearly symmetrical distributions, which support the assumption of normality for subsequent analyses.

Table 3 presents the descriptive statistics of IRR by categories. The mean (*M*) IRR varied across categories, ranging from 53.93% to 64.16%. Main Idea Comprehension showed the lowest mean IRR (53.93%), while Contextual Inference exhibited the highest (64.16%). Language Component Analysis and Discourse Structure Inference showed identical mean IRR (60.70%).

TABLE 3
Descriptive Statistics of IRR by Category

Category	*N*	*M*	*SD*	*CV*
Main Idea Comprehension	56	53.93	7.4	0.137
Contextual Inference	110	64.16	9.61	0.15
Language Component	58	60.7	8.44	0.139
Discourse Structure Inference	111	60.7	10	0.165

The consistency of IRR within each category was examined through standard deviation (*SD*) and coefficient of variation (*CV*). Main Idea Comprehension demonstrated relatively lower variability (*SD* = 7.40, *CV* = 0.137) compared to the other item categories. In contrast, Discourse Structure Inference showed the highest variability (*SD* = 10.00, *CV* = 0.165). Contextual Inference (*SD* = 9.61, *CV* = 0.150) and Language Component Analysis (*SD* = 8.44, *CV* = 0.139) showed intermediate levels of variability, with Language Component Analysis displaying slightly more consistency than Contextual Inference.

The *CV* values across all categories ranged from 0.137 to 0.165. Differences in variability patterns across categories warrant further discussion. IRR consistency appears to be influenced by several factors beyond the number of items in each category. The stability of Main Idea Comprehension (*CV* = 0.137) might be attributed to their focused nature, consistently requiring the identification of central themes or main arguments. In contrast, the higher variability in Discourse Structure Inference (*CV* = 0.165) could reflect the broader range of skills required, from analyzing paragraph organization to understanding rhetorical relationships.

Text readability variability might be another factor contributing to these differing IRR patterns. The relationship between text readability and IRR patterns were examined in the following section, potentially offering important implications for test development and standardization processes in high-stakes testing contexts like CSAT.

Readability and Student IRR

This section presents the descriptive statistics of readability scores across different item categories, followed by an analysis of their correlation with student IRR to address the second research question. To ensure statistical validity, the Shapiro-Wilk test was conducted to examine the normality of readability score distributions within each item category. The test results confirmed normal distributions for all variables (p > .05). Further supporting this finding, the skewness values across item categories remained within a narrow range (-0.43 to 0.41), demonstrating nearly symmetrical distributions and validating the normality assumption for subsequent analyses.

Table 4 presents the descriptive statistics of four readability formulas across different item categories. The FRE scores ranged from 38.62 to 43.78, indicating consistently challenging reading levels. Contextual Inference showed the highest mean (*M* = 43.78, *SD* = 13.83), while Main Idea Comprehension displayed the lowest (*M* = 38.62, *SD* = 12.52), the scores falling within the “difficult” range (30–50) of the FRE scale. Texts in this range typically contained around 167 syllables per 100 words and average sentence lengths of 25 words, which are characteristics of academic publications (Flesch, 1948). The relatively large standard deviations (11.99–13.83) suggest considerable variation in sentence length and word length patterns within each category, though this surface-level linguistic variation may not necessarily reflect the full spectrum of text complexity. Notably, the two intermediate categories – Language Component Analysis (*M* = 42.07, *SD* = 11.99) and Discourse Structure Inference (*M* = 42.28, *SD* = 13.37) – showed similar FRE scores, suggesting comparable levels of linguistic complexity despite their different assessment focuses.

TABLE 4
Descriptive Statistics of Readability Formulas by Category

Variable	Category	*N*	*M*	*SD*
FRE	Main Idea Comprehension	56	38.62	12.52
	Contextual Inference	110	43.78	13.83
	Language Component Analysis	58	42.07	11.99
	Discourse Structure Inference	111	42.28	13.37
NDC	Main Idea Comprehension	56	9.88	1.18
	Contextual Inference	110	9.42	1.14
	Language Component Analysis	58	9.51	1.05
	Discourse Structure Inference	111	9.55	1.06
CAREC	Main Idea Comprehension	56	0.3	0.07
	Contextual Inference	110	0.27	0.06
	Language Component Analysis	58	0.28	0.07
	Discourse Structure Inference	111	0.27	0.07
CML2RI	Main Idea Comprehension	56	10.7	5.27
	Contextual Inference	110	11.09	4.95
	Language Component Analysis	58	11.99	4.65
	Discourse Structure Inference	111	10.73	4.54

The NDC readability scores, with scores ranging from 4.9 (4th grade or below) to 10+ (college graduate level), showed consistent means across categories (9.42–9.88). The small standard deviations (1.05–1.18) suggest that the lexical difficulty levels are tightly controlled across passages within each category. These scores correspond to college undergraduate reading levels, confirming the advanced vocabulary demands of CSAT passages. The consistency between NDC scores across different item categories, coupled with their small standard deviations, indicates a deliberate standardization of vocabulary complexity in CSAT passages. This controlled complexity is further supported by the CAREC scores. The CAREC scores (0–1 scale) showed consistent patterns across different item categories (0.27–0.30). The small standard deviations (0.06–0.07) across all categories indicate uniform text complexity levels within each item type. Main Idea Comprehension showed slightly higher complexity (*M* = 0.30, *SD* = 0.07), while Contextual Inference and Discourse Structure Inference displayed lower complexity levels (*M* = 0.27, *SD* = 0.06–0.07). The narrow range of scores across different item types suggests that CSAT maintains consistent levels particularly in lexical sophistication (as measured by word frequency and age of acquisition) and text cohesion features (including lexical overlap and temporal connective usage), regardless of the specific item category. The CML2RI scores, where higher values indicate easier text processing for L2 readers based on text cohesion and lexical accessibility measures, ranged from 10.70 to 11.99. Language Component Analysis demonstrated the highest processing ease (*M* = 11.99, *SD* = 4.65), while Main Idea Comprehension showed the lowest (*M* = 10.70, *SD* = 5.27). The relatively large standard deviations (4.54–5.27) suggest considerable variation in the measured linguistic features within each category: content word overlap between adjacent sentences (text cohesion), sentence syntactic similarity (syntactic complexity), and word frequency based on CELEX corpus (lexical accessibility). This within-category variation indicates that CSAT reading passages, even within the same item category, incorporate diverse linguistic patterns in terms of cohesive devices, syntactic structures, and vocabulary usage. However, the relatively narrow range of mean scores across categories (10.70–11.99) suggests that these passages maintain similar overall processing demands for L2 readers despite their internal linguistic diversity.

Overall, the above descriptive statistics across different item categories revealed several distinct patterns that suggest a complex relationship between text complexity and student performance. Most notably, the Contextual Inference category, despite showing the highest mean IRR (64.16%), presented the highest FRE score (43.78) and a relatively high CML2RI score (11.09), suggesting that the challenges in this category may stem more from other factors such as the cognitive demands of inferencing rather than from linguistic complexity. Conversely, Main Idea Comprehension demonstrated an inverse pattern, with the lowest IRR (53.93%) despite having the lowest FRE score (38.62) and CML2RI score (10.70), indicating that students can successfully identify main ideas even in linguistically complex texts. This pattern suggests that global comprehension skills may effectively compensate for increased textual difficulty in main idea tasks. The identical IRR means (60.70%) observed in both Language Component Analysis and Discourse Structure Inference categories, despite their differing CML2RI scores (11.99 and 10.73, respectively), suggests that multiple factors beyond linguistic complexity may contribute to item difficulty.

Table 5–8 reveal correlations among readability measures and IRR across different item categories. Main Idea Comprehension as shown in Table 5 showed relatively weak correlations between IRR and readability measures (*r* = -0.174 to 0.195, all *p* > .05). Strong correlations were observed among readability measures themselves, particularly between FRE and NDC (*r* = -0.871) and between CML2RI and NDC (*r* = -0.705).

TABLE 5

Correlations among Readability Formulas and IRR for Main Idea Comprehension

	1	2	3	4	5
1 IRR	1				
2 FRE	-0.064	1			
3 NDC	0.195	-0.871	1		
4 CAREC	0.077	-0.463	0.495	1	
5 CML2RI	-0.174	0.65	-0.705	-0.25	1

Table 6 reveals that Contextual Inference items showed significant, albeit weak, correlations between IRR and two readability measures: a negative correlation with FRE (*r* = -0.226, *p* < .05) and a positive correlation with CAREC (*r* = 0.210, *p* < .05). Similar to Main Idea Comprehension, strong correlations existed between readability measures, with FRE and NDC showing a strong negative correlation (*r* = -0.867) and CAREC and NDC showing a moderate positive correlation (*r* = 0.609).

TABLE 6

Correlations among Readability Formulas and IRR for Contextual Inference

	1	2	3	4	5
1 IRR	1				
2 FRE	-0.226*	1			
3 NDC	0.147	-0.867	1		
4 CAREC	0.210*	-0.655	0.609	1	
5 CML2RI	-0.045	0.628	-0.72	-0.395	1

Note. * indicates *p* <.05

Table 7 indicates that for Language Component Analysis, no significant correlations existed between IRR and readability measures (*r* = -0.168–0.130, all *p* > .05). The strongest correlations were found among readability measures: FRE and NDC showed a strong negative correlation (*r* = -0.887), while CAREC and NDC demonstrated a moderate positive correlation (*r* = 0.564).

TABLE 7

Correlations among Readability Formulas and IRR for Language Component Analysis

	1	2	3	4	5
1 IRR	1				
2 FRE	0.074	1			
3 NDC	-0.127	-0.887	1		
4 CAREC	-0.168	-0.651	0.564	1	
5 CML2RI	0.13	0.583	-0.528	-0.13	1

Table 8 shows that Discourse Structure Inference items exhibited no significant correlations between IRR and readability measures (*r* = -0.058–0.106, all *p* > .05). However, consistent with other categories, strong correlations were observed among readability measures: FRE and NDC showed a strong negative correlation (*r* = -0.812), and CAREC and NDC displayed the strongest positive correlation (*r* = 0.717) among all item categories.

TABLE 8

Correlations among Readability Formulas and IRR for Discourse Structure Inference

	1	2	3	4	5
1 IRR	1				
2 FRE	-0.058	1			
3 NDC	0.106	-0.812	1		
4 CAREC	0.046	-0.631	0.717	1	
5 CML2RI	-0.043	0.572	-0.563	-0.322	1

The consistent pattern across all four item categories was the strong negative correlation between FRE and NDC (ranging from *r* = -0.812–0.887), suggesting a systematic inverse relationship between these two readability measures regardless of item type. Additionally, CAREC and NDC maintained moderate to strong positive correlations across categories (*r* = 0.495–0.717), while CML2RI showed moderate to strong positive correlations with FRE (*r* = 0.572–0.650) and moderate to strong negative correlations with NDC (*r* = -0.528–0.705). However, the apparent disconnect between readability measures and IRR across all categories suggests that effective reading assessment may need to consider multiple factors in determining item difficulty.

Multiple regression analyses revealed varying predictive patterns across item categories. Table 9 shows multiple regression results for Main Idea Comprehension. The model including all four readability measures explained 9.1% (R^2 = .091) of the variance in IRR, with an adjusted R^2 of .020 (F(4, 51) = 1.275, p = .292). None of the individual readability measures emerged as significant predictors at α = .05 level, though both FRE and NDC showed marginally significant positive relationships (FRE: B = 0.271, *SE* = 0.162, t = 1.672, p = .101; NDC: B = 3.235, *SE* = 1.899, t = 1.703, p = .095). Other predictors showed no significant effects: CAREC (B = 0.872, *SE* = 17.238, t = 0.051, p = .960) and CML2RI (B = -0.150, *SE* = 0.270, t = -0.555, p = .581). The diagnostic tests indicated no violations of regression assumptions: residuals showed normal distribution (Shapiro-Wilk test: W = 0.993, p = .989) and homoscedasticity (Breusch-Pagan test: *BP* = 0.850, p = .927). However, VIF values for FRE (4.23) and NDC (5.14) suggested potential multicollinearity concerns.

TABLE 9
Multiple Regression Results for Main Idea Comprehension

Predictor	B	*SE*	*t*	*p*
(Intercept)	*12.849*	*24.182*	*0.531*	*0.598*
FRE	0.271	0.162	1.672	0.101
NDC	3.235	1.899	1.703	0.095
CAREC	0.872	17.238	0.051	0.960
CML2RI	-0.150	0.270	-0.555	0.581

Note. N = 56. B = unstandardized regression coefficient; *SE* = standard error

Table 10 shows multiple regression results for Contextual Inference. The model including all four readability measures explained 7.6% (R^2 = .076) of the variance in IRR, with an adjusted R^2 of .041 (F(4, 105) = 2.157, p = .079). None of the individual readability measures emerged as significant predictors at α = .05 level, though FRE showed a marginally significant negative relationship (B = -0.236, *SE* = 0.138, t = -1.706, p = .091). Other predictors showed no significant effects: NDC (B = -1.109, *SE* = 1.795, t = -0.618, p = .538), CAREC (B = 18.165, *SE* = 19.897, t = 0.913, p = .363), and CML2RI (B = 0.231, *SE* = 0.264, t = 0.875, p = .384). The diagnostic tests indicated no violations of regression assumptions: residuals showed normal distribution (Shapiro-Wilk test: W = 0.991, p = .678) and homoscedasticity (Breusch-Pagan test: *BP* = 6.603, p = .158). However, VIF values for FRE (4.49) and NDC (5.15) suggested potential multicollinearity concerns.

TABLE 10
Multiple Regression Results for Contextual Inference

Predictor	B	*SE*	*t*	*p*
(Intercept)	*77.456*	*23.154*	*3.345*	*0.001*
FRE	-0.236	0.138	-1.706	0.091
NDC	-1.109	1.795	-0.618	0.538
CAREC	18.165	19.897	0.913	0.363
CML2RI	0.231	0.264	0.875	0.384

Note. N = 110.

Table 11 shows multiple regression results for Language Component Analysis. The model including all four readability measures explained 8.4% (R^2 = .084) of the variance in IRR, with an adjusted R^2 of .014 (F(4, 53) = 1.209, p = .318). None of the individual readability measures emerged as significant predictors at α = .05 level, though CAREC showed a marginally significant negative relationship (B = -38.227, *SE* = 22.450, t = -1.703, p = .094). Other predictors showed no significant effects: FRE (B = -0.380, *SE* = 0.242, t = -1.570, p = .122), NDC (B = -2.358, *SE* = 2.295, t = -1.027, p = .309), and CML2RI (B = 0.451, *SE* = 0.321, t = 1.404, p = .166). The diagnostic tests indicated no violations of regression assumptions: residuals showed normal distribution (Shapiro-Wilk test: W = 0.988, p = .838) and homoscedasticity (Breusch-Pagan test: *BP* = 3.641, p = .457). However, VIF values for FRE (6.86) and NDC (4.71) suggested potential multicollinearity concerns.

TABLE 11

Multiple Regression Results for Language Component Analysis

Predictor	B	*SE*	*t*	*p*
(Intercept)	*104.423*	*31.413*	*3.324*	*0.002*
FRE	-0.380	0.242	-1.570	0.122
NDC	-2.358	2.295	-1.027	0.309
CAREC	-38.227	22.450	-1.703	0.094
CML2RI	0.451	0.321	1.404	0.166

Note. N = 58.

Table 12 shows multiple regression results for Discourse Structure Inference. The model including all four readability measures explained 1.5% (R^2 = .015) of the variance in IRR, with an adjusted R^2 of -.022 (F(4, 106) = 0.405, p = .805). None of the individual readability measures emerged as significant predictors at α = .05 level: FRE (B = 0.049, *SE* = 0.129, t = 0.380, p = .704), NDC (B = 2.002, *SE* = 1.809, t = 1.107, p = .271), CAREC (B = -8.601, *SE* = 20.822, t = -0.413, p = .680), and CML2RI (B = 0.045, *SE* = 0.269, t = 0.167, p = .868). The diagnostic tests indicated no violation of the normality assumption (Shapiro-Wilk test: W = 0.984, p = .206), though the heteroscedasticity test showed potential concern (Breusch-Pagan test: BP = 11.777, p = .019). VIF values for FRE (3.19) and NDC (3.95) were lower than other item categories, suggesting less severe multicollinearity concerns.

TABLE 12

Multiple Regression Results for Discourse Structure Inference

Predictor	B	*SE*	*t*	*p*
(Intercept)	*41.379*	*20.051*	*2.064*	*0.042*
FRE	0.049	0.129	0.380	0.704
NDC	2.002	1.809	1.107	0.271
CAREC	-8.606	20.822	-0.413	0.680
CML2RI	0.045	0.269	0.167	0.868

Note. N = 111.

The multiple regression analyses examined the combined predictive power of readability measures for IRR across different categories in CSAT English reading items. The results showed that the explanatory power of readability measures was consistently low across all item categories (R^2 ranging from .015 to .091), with readability measures explaining less than 10% of the variance in IRR. These findings suggest that text readability measures alone have limited predictive power for IRR in CSAT English reading comprehension items, particularly for Discourse Structure Inference items where the model explained only 1.5% of the variance in IRR. Although weak correlations were found between IRR and both FRE and CAREC in the contextual inference category, these relationships were not significant in the multiple regression analysis. This reduction in statistical significance might be partly explained by the multicollinearity observed among predictors, particularly the strong correlation between FRE and NDC, as indicated by their high Variance Inflation Factors (VIF = 4.23 and 4.15, respectively).

The consistently low predictive power of readability measures across all item categories suggests the presence of more influential factors specific to each item type. Our findings provide several empirical grounds for this interpretation. In Main Idea Comprehension items, despite showing the highest linguistic complexity (lowest FRE = 38.62), students maintained relatively stable performance (lowest CV = 0.137). This pattern, coupled with weak correlations between readability measures and IRR (r = -0.174 to 0.195), suggests that global reading strategies may play a more crucial role than linguistic complexity in determining performance. For Contextual Inference items, the highest IRR (64.16%) coincided with the highest FRE score (43.78), and uniquely showed significant correlations between IRR and both FRE (r = -0.226, p < .05) and CAREC (r = 0.210, p < .05). This distinctive pattern, not observed in other item categories, suggests that text complexity has a modest but significant influence on performance in contextual inference tasks. The presence of these correlations might be attributed to the nature of contextual inference processes, which require both basic text comprehension and higher-order inferencing skills. The negative correlation with FRE indicates that easier texts (higher FRE scores) are associated with lower error rates, suggesting that a certain level of text comprehensibility is necessary for successful inference-making. Similarly, the positive correlation with CAREC, which measures various linguistic features including word acquisition patterns, n-gram means, text cohesion, and lexical diversity, implies that these textual characteristics support the inference process. However, the limited predictive power in regression analysis ($R^2 = .076$) indicates that while text complexity plays a role, inferencing skills still remain a crucial factor in determining performance.

An additional consideration for both Main Idea Comprehension and Contextual Inference items is the linguistic complexity of answer choices themselves. Unlike Language Component Analysis and Discourse Structure Inference items where responses often involve analyzing given text structures, these two item categories require students to choose among English answer choices. The varying linguistic complexity and discriminability of these answer options could significantly affect IRR patterns. This limitation suggests that future studies should analyze the linguistic complexity of answer choices by using similar readability measures and examine their relationship with student performance. In Language Component Analysis items, despite having the highest processing ease (CML2RI = 11.99), performance remained moderate (IRR = 60.70%) with no significant correlations between readability measures and IRR. This disconnect between text processability and performance suggests that specialized linguistic analysis skills may be more critical than general text comprehension ability. For Discourse Structure Inference items, the highest variability in performance ($CV = 0.165$) coupled with the lowest explanatory power of readability measures ($R^2 = .015$) indicates that discourse-level comprehension skills, particularly the ability to understand text organization and coherence, may be more influential than surface-level or complex linguistic features. These patterns consistently demonstrate that while text complexity plays a role, task-specific reading skills likely exert greater influence on student performance in CSAT English reading items. This finding has important implications for both assessment development and reading instruction, suggesting the need to focus more on developing these specific comprehension skills (e.g., global reading strategies for main ideas, inferencing skills for contextual understanding, linguistic analysis skills for language components, and discourse structure awareness) rather than solely managing text complexity.

CONCLUSION

This study examined the relationship between text readability and student performance on CSAT English reading items, revealing several important findings about the nature of reading assessment in high-stakes EFL contexts. First, the analysis of IRR across different item categories demonstrated that student performance patterns vary systematically by item type, with Main Idea Comprehension items showing the most consistent performance patterns and Discourse Structure Inference items displaying the highest variability. This finding suggests that the cognitive demands associated with different comprehension tasks may play a more crucial role in determining item difficulty than linguistic complexity alone. Second, the investigation of readability formulas' predictive power revealed limited correlations with student performance, particularly after controlling for multiple comparisons. Notably, while Contextual Inference items showed unique significant correlations with both FRE and CAREC measures, the overall explanatory power remained low across all item categories ($R^2 < 10\%$). This finding challenges the conventional reliance on readability measures as primary indicators of text difficulty in assessment development. The weak correlations observed suggest that the relationship between text complexity and student performance is more nuanced than traditional readability measures might suggest, particularly in EFL contexts where additional cognitive and linguistic factors may influence comprehension. These findings have important implications for both assessment development and EFL pedagogy. For test developers, the results suggest the need for a more comprehensive approach to difficulty calibration that considers both the linguistic features of texts and the specific cognitive demands associated with different comprehension tasks. The consistent NDC scores across categories indicate successful standardization of vocabulary complexity, while the varying patterns of performance across item categories suggest that maintaining consistent difficulty levels requires careful attention to both text complexity and task-specific demands. For EFL educators, the findings highlight the importance of preparing students for different types of reading comprehension tasks, as performance patterns vary significantly across item categories. This is particularly evident in how Main Idea Comprehension items showed stable performance despite high linguistic complexity, suggesting the effectiveness of global reading strategies. The results also suggest that focusing solely on linguistic complexity may not be sufficient; students need practice with various comprehension tasks that require different cognitive processing skills.

This study has several limitations that should be noted. First, while the corpus included a substantial number of passages (N = 335), it was limited to the CSAT and its mock test from 2017 to 2024. A longer time span might reveal different patterns or trends. Second, the analysis focused on IRR as the primary measure of item difficulty, which provides only one aspect of student performance. Third, while this study examined four readability formulas, there might be other aspects of text complexity not captured by these measures. Fourth, the linguistic complexity of answer choices, particularly in Main Idea Comprehension and Contextual Inference items, was not analyzed, which could have influenced the IRR patterns. Future research could address these limitations in several ways. First, analyzing how the discriminating power of distractors across different item types affects student performance could provide deeper insights into item functioning. Particularly for Main Idea Comprehension and Contextual Inference items where students must choose among English answer choices, understanding how distractor characteristics influence performance patterns could contribute to more effective item development. Second, exploring additional factors that might influence student performance on different types of reading comprehension items, such as topic familiarity, text organization, or specific linguistic features beyond current readability measures, could yield valuable insights. Additionally, qualitative analysis of student responses (e.g., investigating the relationship between task-specific reading skills such as inferencing skills, discourse structure awareness, and student performance) might provide deeper insights into the cognitive processes involved in different types of reading comprehension tasks. Despite these limitations, this study makes a significant contribution to our understanding of how text readability relates to student performance in high-stakes EFL assessment contexts. The findings provide valuable insights for both test developers and educators working to improve the assessment and teaching of reading comprehension in EFL contexts.

References

Alderson, J. C. (2000). *Assessing reading.* Cambridge University Press.

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review, 24*(4), 68-88.

Chall, J. S., & Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.

Chang, Kyung-Suk. (2004). A model of predicting item difficulty of the reading test of College Scholastic Ability Test. *Foreign Languages Education, 11*(1), 111-130.

Choi, J., & Crossley, S. A. (2022a). Automated readability web app for English. *Proceedings of the 23 IEEE International Conference on Advanced Learning Technologies (ICALT 2022)*, (pp. 1-6). Bucharest, Romania.

Choi, J., & Crossley, S. A. (2022b). Automatic Readability Tool for English (ARTE) [Web application]. https://www.linguisticanalysistools.org/arte.html

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). L. Erlbaum Associates.

Cohen, A. D., & Upton, T. A. (2007). ‘I want to go back to the text’: Response strategies on the reading subtest of the new TOEFL. *Language Testing, 24*(2), 209-250.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly, 42*(3), 475-493.

Crossley, S. A., & McNamara, D. S. (2008). Assessing L2 reading texts at the intermediate level: An approximate replication of Crossley, Louwerse, McCarthy & McNamara (2007). *Language Teaching: Surveys and Studies, 41*(3), 409-429.

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), 541-561.

Crossley, S. A., Skalicky, S., Dascalu, M., McNamara, D. S., & Kyle, K. (2017). Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes, 54*(5-6), 340-359.

Crossley, S., Heintz, A., Choi, J. S., Batchelor, J., Karimi, M., & Malatinszky, A. (2023). A large-scaled corpus for assessing text readability. *Behavior Research Methods, 55*(2), 491-507.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

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

Dale, E., & Chall, J. S. (1949). The concept of readability. *Elementary English, 26*, 19-26.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221-233

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression* (3rd ed.). Sage.

Green, A., Ünaldi, A., & Weir, C. (2010). Empiricism versus connoisseurship: Establishing the appropriacy of texts in tests of academic reading. *Language Testing, 27*(2), 191-211.

Hwang, Lee-su, & Lee, Je-Young. (2020a). Correlation analysis between the text variables and item difficulty in CSAT: Focusing on syntactic complexity. *Studies in English Language & Literature, 46*(1), 265-283.

Hwang, Lee-su, & Lee, Je-Young. (2020b). Analysis of correlation between cohesion and item difficulty in English reading section of

CSAT. *The Journal of the Korea Contents Association, 20*(5), 344-350.

Hwang, Lee-su, & Lee, Je-Young. (2020c). Correlation between readability/word information and item difficulty in CSAT English reading passage. *The Journal of Humanities and Social Sciences, 11*(2), 389-400.

Ji, Sulki, & Kim, Haedong. (2014). A comparison of the effects of test-item type and text familiarity on results of an English reading test. *Foreign Languages Education, 21*(1), 215-239.

Joo, Hunwoo. (2017). Investigation of text readability of the college scholastic ability test and high school English textbooks based on lexical familiarity and syntactic complexity. *Studies in English Education*, *22*(2), 1-24.

Kang, Moon-gu, Kim, Kyeong-hwan, Park, Seon-ha, Cho, Keum-hui, & Hwang, Jin-ho. (2021). Yeong-eo-gwa seonda-hyeong siheom pyeong-ga mun-hang-eun eotteoh-ge mandeul-eoji-na? [How multiple-choice test items are created in English assessment]. EBS Books.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (Automated Reliability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel (Research Branch Report 8-75). Memphis, TN: Naval Air Station; 1975. *Naval Technical Training, US Naval Air Station: Millington, TN*.

Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing, 19*(2), 193-220.

Korea Educational Broadcasting System. (2024a). Go3 gichulmunje daunlodeu [Download Past Test Questions for High School Seniors]. https://www.ebsi.co.kr/ebs/xip/xipc/previousPaperList.ebs?targetCd=D300

Korea Educational Broadcasting System. (2024b). Yeokdae deunggeukcut/odabyul TOP15 [Top 15 Historical Grade Cutoffs/ Incorrect Response Rates]. https://www.ebsi.co.kr/ebs/xip/xipa/retrievePastGrdCutWrongAnswerRate.ebs?tab=1

Kwon, Jeong-hwa, & Lee, Jeong-won. (2015). A study on reading test-taking strategies in item types of an English reading test. *Journal of Humanities Research, 101*, 79-100.

McCarthy, K. S., & McNamara, D. S. (2021). The multidimensional knowledge in text comprehension framework. *Educational Psychologist, 56*(3), 196-214.

McLaughlin, G. H. (1969). SMOG grading: A new readability formula. *Journal of Reading, 12*, 639-646.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix* (1st ed.). Cambridge University Press.

Ministry of Education. (2014, December 26). Daehak suhak neungryeok siheom yeong-eo yeongyeok jeoldae pyeongga doip [Introduction of absolute evaluation for English section of College Scholastic Ability Test] [Press release]. https://www.moe.go.kr/boardCnts/viewRenew.do?boardID=294&lev=0&statusYN=C&s=moe&m=020402&opType=N&boardSeq=58100

Nahatame, S. (2021). Text readability and processing effort in second language reading: A computational and eye-tracking investigation. *Language Learning, 71*(4), 1004-1043.

R Core Team. (2024). R: A Language and environment for statistical computing. *R Foundation for Statistical Computing.* https://www.R-project.org/

Richards, J. C., & Schmidt, R. W. (2013). *Longman dictionary of language teaching and applied linguistics*. Routledge.

Smith, E. A., & Senter, R. J. (1967). *Automated readability index* (Vol. 66, No. 220). Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command.

Song, Juha. (2021). An analysis of Korean high school English textbooks through syntactic complexity and readability. *Modern English Education, 22*(1), 57-69.

Yum, Young-hee, & Kim, Haedong. (2023). Comparative analysis of linguistic features in English reading passages on national achievement English tests and CSATs. *English21, 36*(3), 175-198.

Zeileis, A., & Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News 2*(3), 7-10. https://CRAN.R-project.org/doc/Rnews/