Assessing Writing 66 (2025) 100975

Contents lists available at ScienceDirect

Assessing Writing

journal homepage: www.elsevier.com/locate/asw

# Lexical richness in young English learners' writing: A focus on opinion and listen-write task types

Hakyung Sung[a,c,*,1], Mikyung Kim Wolf[b,2], Michael Suhan[b], Kristopher Kyle[a,3]

[a] *University of Oregon, USA*
[b] *ETS, USA*
[c] *Rochester Institute of Technology, USA*

ARTICLE INFO

*Keywords:*
Lexical richness
Young English learners
Opinion and listen-write tasks

ABSTRACT

Ample research has examined the linguistic characteristics of second language (L2) writing across proficiency scores, with a focus on lexical diversity, sophistication, and density as key dimensions of lexical richness. However, the applicability of these indices to young L2 learners' writings, often characterized by limited vocabulary and constrained output in standardized writing tasks, remained underexplored. To address this gap, this study analyzed the lexical richness of young L2 learners' written productions from two TOEFL Junior Writing tasks (Opinion and Listen-Write tasks), using 37 tailored indices of lexical diversity, sophistication, and density. The results indicated that the lexical characteristics of young L2 learners vary by task score and task type, particularly when assessed through indices such as lexical diversity (e.g., moving-average type-token ratio) and sophistication (e.g., n-gram strength of association). Nonetheless, incorporating additional measures, such as syntactic complexity or discourse features, may be essential for distinguishing young L2 learners at higher proficiency levels.

## 1. Introduction

In second language (L2) acquisition research, a substantial body of research has explored the relation between linguistic features and L2 writing proficiency, demonstrating meaningful relationships that inform both assessment and instructional practices (e.g., Kim, 2014; Lu, 2011; Norris & Ortega, 2003; Ortega, 2003). Among the various linguistic features, exploring lexical richness in written tasks serves as a foundational step, providing a basis for examining other dimensions such as lexicogrammatical and syntactic complexity (Halliday, 1989; Laufer & Nation, 1995). These lexical features contribute to the construct of lexical proficiency and can also play a critical role in explaining variation in writing quality (e.g., Crossley et al., 2015; Engber, 1995; Nation, 2005). This connection holds because, as lexical proficiency increases, learners are more likely to draw on a broader and more sophisticated vocabulary, thereby often signaling higher L2 writing ability (Bulté et al., 2024; Bulté & Housen, 2012).

Motivated by the relationship, extensive work has examined mature learners' lexical richness (e.g., Crossley et al., 2009; Crossley et al., 2012; Crossley et al., 2014; Ferris, 1994; Kim et al., 2018; Kyle & Crossley, 2015; Kyle & Crossley, 2016; Kyle et al., 2021; Yoon,

[*] Correspondence to: 1 Lomb Memorial Drive, Rochester, NY 14623-5603, USA.
*E-mail addresses:* hksgla@rit.edu (H. Sung), mkwolf@ets.org (M.K. Wolf), msuhan@ets.org (M. Suhan), kkyle2@uoregon.edu (K. Kyle).
[1] ORCID: 0000-0002-5860-7353
[2] ORCID: 0000-0002-0622-7754
[3] ORCID: 0000-0001-5415-9672

https://doi.org/10.1016/j.asw.2025.100975
Received 28 January 2025; Received in revised form 25 July 2025; Accepted 26 July 2025
Available online 6 August 2025

2018; Zenker & Kyle, 2021). By contrast, relatively little attention has been paid to young L2 writers in standardized tests. In L2 assessment, young language learners are broadly identified as children between the ages of 5 and 13, aligning with kindergarten through early secondary or middle school education levels (Wolf, 2024). This study particularly focuses on examining the lexical characteristics of young L2 learners' writing in the TOEFL Junior Writing test, which assesses English proficiency for students aged 11 and older (ETS, 2019). To explore these characteristics, we investigate how specific lexical features—such as diversity, sophistication, and density—distinguish between task scores (0−4) and task types (opinion versus integrated writing). The findings illuminate key indices of lexical proficiency in young L2 writers that can carry implications for writing assessment design.

## 2. Literature review

### 2.1. Lexical richness, complexity, and difficulty

Lexical richness refers to the characteristics of word use (Kyle, 2019; Laufer & Nation, 1995; Lu, 2012). Originally introduced by Yule (1944) as a measure of vocabulary breadth, it is closely linked to the current concept of lexical diversity, which refers to the variety of unique words used in a text relative to its total word count. Over time, the concept has expanded to include lexical density, defined as the proportion of content words relative to function words (Halliday, 1989; Linnarud, 1986), and lexical sophistication, defined as the proportion of sophisticated or advanced words (Laufer & Nation, 1995; Meara, 2005).

In L2 acquisition research, these dimensions of lexical richness are particularly important for understanding productive language use. Conceptually, lexical richness often reflects learners' capacity to handle structural complexity and cognitive demands during vocabulary production (Bulté et al., 2024). The first dimension, lexical complexity, often includes both structural intricacy of lexical items (e.g., word length, morphological complexity) and the diversity of words produced (Bulté & Housen, 2012; Pallotti, 2015). The second, lexical difficulty, relates to cognitive effort required for word production, as more challenging words require greater processing resources (DeKeyser, 2005; DeKeyser, 2016). Researchers generally assume that words requiring greater cognitive effort tend to appear less frequent in learners' output (Laufer & Nation, 1995). Accordingly, lexical difficulty has been often operationalized using measures of lexical sophistication, which were initially computed based on word frequency, but were later expanded to incorporate additional indicators (Kyle et al., 2018).

### 2.2. Empirical studies on the relationship between lexical richness and L2 writing proficiency

Previous studies have explored how lexical richness, measured in terms of diversity, sophistication, and density, is related to L2 learners' writing performance.

#### 2.2.1. Lexical diversity

Lexical diversity refers to the range of different words a writer uses in a text. It is commonly used as an indicator of language proficiency, based on the assumption that more advanced learners have greater access to vocabulary and are therefore less likely to repeat words (Engber, 1995; Jarvis, 2002). To operationalize lexical diversity, researchers have developed a range of indices that quantify the extent of vocabulary variation in a text (Jarvis, 2013). These indices have been widely used to investigate the relationship between lexical diversity and L2 writing proficiency, particularly in independent writing tasks (e.g., argumentative writing). For example, Kyle et al. (2021) reported significant correlations between lexical diversity indices (e.g., the number of word types: $r = 0.698$; moving-average type-token ratio [MATTR]: $r = 0.566$) and L2 writing proficiency in argumentative independent essays from the TOEFL Public Use Data. Similarly, Woods et al. (2023) found that multiple lexical diversity indices (e.g., MATTR, the Measure of Textual Lexical Diversity [MTLD]) predicted opinion writing assessment scores among L2 students in a U.S. university.

In contrast, relatively few studies have explored the relationship between lexical diversity and proficiency scores in source-based writing tasks. Cumming et al. (2005), for example, demonstrated a positive relationship between lexical diversity (measured by type-token ratios [TTR]) and holistic scores in read-write and listen-write tasks. Cai and Yan (2024) substantiated these findings, showing that higher lexical diversity (as measured by TTR for function words and root TTR for content words) is positively associated with integrated writing scores.

Although these studies underscore the importance of lexical diversity measures across writing tasks and test settings, they also reveal two limitations. First, traditional indices (e.g., TTR, root TTR) are inherently sensitive to variations in text length (Koizumi & In'nami, 2012; Malvern et al., 2004; McCarthy & Jarvis, 2010): TTR tends to decrease and root TTR tends to increase with longer texts, potentially misrepresenting the true range of vocabulary in test-taker responses. Although some researchers investigating independent writing tasks have employed text-length-stable measures (e.g., MATTR, MTLD; Zenker & Kyle, 2021), source-based writing studies often rely on TTR-based indices, leaving them vulnerable to text-length effects (e.g., Cumming et al., 2005; Cai & Yan, 2024). In addition, lexical diversity indices fail to capture the sophistication of vocabulary use (Crossley et al., 2015; Kyle & Eguchi, 2023), which can limit their ability to reflect the cognitive difficulties associated with L2 vocabulary production (DeKeyser, 2005; DeKeyser, 2016).

#### 2.2.2. Lexical sophistication

Exploring the cognitive demands of word usage in L2 writing often involves analyzing lexical sophistication, typically defined as the use of advanced or infrequent words (Laufer & Nation, 1995; Meara & Bell, 2001). This construct is commonly assessed through word frequency relative to a reference corpus, based on the assumption that the sizable reference corpus represents the target language use for L2 learners in a given written domain (e.g., Crossley et al., 2014; Kyle & Crossley, 2015). While exceptions exist, the dominant view is that using less frequent words reflects more advanced lexical knowledge. Beyond frequency-based indices, subsequent studies have introduced a broader range of measures to capture lexical sophistication, including word range (to assess contextual dispersion), psycholinguistic norms (e.g., familiarity, concreteness, imageability), contextual distinctiveness, age of acquisition, and n-gram measures (e.g., n-gram frequency, range, and strength of association) (e.g., Crossley et al., 2009; Kim et al., 2018; Kyle & Crossley, 2015; Kyle & Crossley, 2016; Kyle et al., 2018; Yoon, 2018).

Among the proposed indices for measuring lexical sophistication, several have been shown to explain meaningful variance in writing scores within specific learner corpora. For example, Kim et al. (2018) analyzed L2 argumentative essays by Korean students and naturalistic journal writing by L2 learners at a U.S. university. They identified several lexical components, such as bigram and trigram association strength, content word properties, and word specificity, which accounted for 16.1 % and 31 % of the variance in L2 writing and lexical proficiency, respectively. Similarly, Yoon (2018) conducted a longitudinal study of L2 students in the U.S., analyzing narrative and argumentative tasks. The study found that n-gram frequency positively correlated with vocabulary scores (trigrams for narratives, bigrams for argumentative essays), while range measures negatively correlated, suggesting that context-specific words contribute to higher writing scores. These findings align with Kyle and Crossley (2015), who emphasized the predictive value of frequent n-grams and context-specific words for lexical proficiency.

On the other hand, Kyle and Crossley (2016) compared how lexical sophistication relates to writing proficiency across task types in 480 TOEFL iBT test-takers. Their models explained 36.8 % of the variance in independent essay scores but only 8.3 % in integrated task scores, indicating a stronger link between lexical sophistication and proficiency in opinion-based tasks compared to source-based tasks. They suggested that source use in source-based tasks may limit test-takers' ability to demonstrate lexical sophistication, an idea further explored in Kyle (2020), which examined the impact of source use on writing performance. Collectively, these studies demonstrate the importance of multivariate models for measuring lexical sophistication and that the effectiveness of lexical sophistication indices varies across writing task types.

### *2.2.3. Lexical density*

Lexical density refers to the degree of informational content in a text, reflecting how much meaning is carried by lexical items (e.g., nouns, verbs, adjectives) as opposed to grammatical items (e.g., articles, prepositions) (Halliday, 1989). In the context of L2 writing, lexical density provides insight into how much a text focuses on conveying information rather than managing interaction, with higher lexical density often associated with more formal, academic, or informational genres (e.g., Gregori-Signes & Clavel-Arroitia, 2015; Ure, 1971).

To operationalize this construct, previous studies have calculated the ratio of content words to the total number of words in a text (Ure, 1971). However, the application of the lexical density index in L2 research has been limited. One issue is the lack of formal theoretical grounding for treating content words as inherently more complex than function words, which complicates the interpretation of lexical density as a developmental measure of proficiency (Bulté et al., 2024). Moreover, lexical density has often been found to vary more strongly with register than with proficiency level (cf. Kyle, 2019). Less interactive genres, such as academic essays, naturally show higher lexical density than conversational or dialogic texts (e.g., O'Loughlin, 1995). Studies examining its relationship with L2 proficiency have shown mixed results, reporting either no significant findings (cf. Lu, 2012; spoken domain) or limited evidence of a relationship (e.g., Engber, 1995; Linnarud, 1986). Nonetheless, lexical density indices have been frequently investigated alongside lexical cohesion in L2 writing tasks, which reflects a writer's vocabulary choices aimed at enhancing writing quality through lexical overlap (Crossley et al., 2016).

## *2.3. Investigating lexical richness in young L2 learners' writing: Potentials and challenges*

While most L2 writing studies have focused on advanced or mature L2 learners (e.g., TOEFL iBT test-takers, university students), a growing body of work has begun to examine vocabulary use and its relationship to writing proficiency among younger L2 English learners (e.g., Durrant & Brenchley, 2019; Durrant & Brenchley, 2023; De Wilde, 2023; Maamuujav et al., 2021; Verspoor et al., 2012; Wolf et al., 2018). For instance, De Wilde (2023) found that lexical features such as diversity, sophistication, and spelling accuracy predicted proficiency scores among Dutch-speaking students in their first year of secondary school, with lexical models explaining 50 % of the variance. Similarly, Wolf et al. (2018) reported that higher holistic scores on the TOEFL Junior Writing test were linked to increased use of academic, low-frequency, and abstract vocabulary, which are key indicators of academic literacy (Schleppegrell, 2004). These findings highlight the central role of lexical richness in assessing writing proficiency among younger learners, particularly in standardized assessment contexts where vocabulary use is closely tied to scoring rubrics (Chapelle et al., 2011).

However, assessing young L2 learners' written production poses inherent challenges due to their limited vocabulary size and the resulting brevity of their texts in assessment contexts. For example, Wolf et al. (2018) found that test-takers, aged 10–15, produced an average of 110–120 words and 10 sentences in a 10-minute argumentative task, while De Wilde (2023) reported even shorter texts, averaging 31 words in a picture-narration task. This brevity raises key questions about the effectiveness of previously studied lexical richness indices, their relationship with task scores, and their variability across writing task types—questions that guide the focus of our current study.

## *2.4. Current study*

While extensive research on lexical richness (often measured through indices of lexical diversity, sophistication, and density) exists in L2 writing, most studies have focused on older or more advanced learners. Consequently, it remains unclear whether these measures effectively capture aspects of the writing performance of young L2 learners, whose texts tend to be shorter and whose vocabulary size may be more limited. To address this gap, the present study examines the relationship between lexical richness and English writing proficiency among young learners taking the TOEFL Junior Writing test, comparing these features across task scores and two task types (Opinion and Listen-Write). Specifically, the study is guided by the following research questions (RQs):

RQ1. What is the relationship between lexical richness indices and task scores in the Opinion and Listen-Write tasks?

RQ2. How do lexical richness indices distinguish young L2 learners across task scores in the Opinion and Listen-Write tasks?

RQ3. How do lexical richness indices differ between task types in young L2 learners' writing on the Opinion and Listen-Write tasks?

## 3. Methods

### 3.1. Dataset

#### 3.1.1. Test takers

As this is a secondary data analysis study, we obtained the dataset used in this study from ETS. This operational data includes scores, responses, and students' demographic information (e.g., L1, gender, and age). The scores were assigned by professionally trained ETS raters. The study sample comprised 637 students from Korea, Mexico, and Turkey who took the same writing prompts from the TOEFL Junior Writing test. The sample included Korean-, Spanish-, and Turkish-speaking students ranging in age from 11 to 16 years old. Table 1 presents the demographic characteristics of the test takers in terms of first language (L1), gender, and age.

#### 3.1.2. Writing tasks

The TOEFL Junior Writing test consists of four tasks: Edit, Email, Opinion, and Listen-Write. This study focuses on the Opinion and Listen-Write tasks because they require learners to produce extended written responses, providing a broader sample of vocabulary use. In contrast, the Edit task involves sentence-level revisions, and the Email task follows a relatively fixed communicative format, both of which may offer limited opportunities for demonstrating lexical richness. Second, prior research on lexical richness has often focused on argumentative writing and source-based writing tasks, which are writing tasks that require sustained language production and that align closely with the Opinion and Listen-Write tasks. Thus, selecting these two tasks allowed us to better situate our findings within the broader context of L2 writing research and to build comparability with previous studies. Sample tasks illustrating the TOEFL Junior Writing test format are available for public viewing at https://www.ets.org/toefl/junior/prepare.html. However, the tasks administered to test takers in this study were distinct from the publicly available samples and were not accessible to students prior to testing.

In the Opinion task, test takers are prompted with a question or statement on a specific topic. They are asked to write a paragraph that expresses their personal viewpoint on the subject with supporting details (e.g., *Some people think that students should not be allowed to have mobile phones in school. Others think that it is acceptable. What do you think?*). The expected length for this written response is between 100 and 150 words.

The Listen-Write task engages test takers in a listening exercise in which they hear a brief talk on an academic subject (e.g., *Different amounts of rainfall on a mountain in Hawaii*), accompanied by one or more visual aids. The listening duration is approximately 100 s and set in a classroom environment. Following the one-time audio presentation, students are asked to write a summary of the information presented based on the specific prompt (e.g., *Explain the reason for the different amounts of rainfall on a mountain in Hawaii*). They are allowed to take notes during the listening phase and use these notes to aid in drafting their written response.

According to the TOEFL Junior Test Taker Handbook (ETS, 2023), the testing session is conducted under standardized conditions following ETS's official test administration protocols. These protocols include proctoring guidelines and test delivery via a secure online testing platform. Each task is timed, and responses are collected and stored electronically. No external materials (e.g., dictionaries, internet, or textbooks) are permitted during the writing sessions to ensure independent performance. For the present study, we selected data including only those test takers who responded to the same prompt within each task type, thereby minimizing potential prompt effects on the writing samples.

**Table 1**
Test takers' characteristics.

| | | *n* | % |
|---|---|---|---|
| L1 | Korean | 293 | 46.0 |
| | Spanish | 234 | 36.7 |
| | Turkish | 110 | 17.3 |
| Gender | Female | 310 | 48.7 |
| | Male | 324 | 50.9 |
| | Not specified | 3 | 0.5 |
| Age | 11 | 104 | 16.3 |
| | 12 | 92 | 14.4 |
| | 13 | 66 | 10.4 |
| | 14 | 113 | 17.7 |
| | 15 | 237 | 37.2 |
| | 16 | 25 | 3.9 |

#### 3.1.3. Scoring procedures and score/writing sample distributions

The data included the students' written responses along with scores from ETS's trained raters using the task rubrics with score points from 0 to 4 (cf. Appendix A). Two trained raters scored the responses, with identical scores assigned to 44 % of the essays, requiring no further adjudication. For the remaining 56 % of responses, where scores differed, a third rater determined the final score. The score distributions and the average writing lengths (in terms of the number of word tokens and word types) for each score point are presented in Table 2 and Fig. 1, respectively.

### 3.2. Lexical richness indices

#### 3.2.1. Measuring unit

The operational definition of a lexical item in this study is any string of letters separated by spaces or punctuation. For measuring lexical richness, individual words served as the primary unit of analysis, and lemmatized forms were used to assess both lexical diversity and density. In contrast, lexical sophistication required a more complex approach to lemmatization. The software employed for these analyses is sensitive to variations in tokens and automatically performs tokenization or lemmatization as needed; however, exact lemmatization procedures differ depending on the database source. A detailed description of each index, including database information and lemmatization protocols, is provided in the Index Guide (Kyle et al., 2018, p. 1037). Additionally, each word was automatically tagged for part of speech to distinguish homonyms (e.g., *run_NOUN* ["had a good *run*"] vs. *run_VERB* ["we *run* to the school"]), and to isolate content words (i.e., nouns, verbs, adjectives, adverbs) for indices focused on lexical richness exclusively among content words.

#### 3.2.2. Lexical diversity indices

Lexical diversity was measured using the moving-average type-token ratio (MATTR), selected for its stability across varying text lengths (e.g., Covington & McFall, 2010; Zenker & Kyle, 2021). Following Kyle et al. (2024), we optimized the window size for each task. Preliminary analyses indicated that a window size of 11 was optimal for the Opinion task and 20 for the Listen-Write task (cf. Appendix B); these were subsequently used for all further calculations. MATTR scores were computed using a modified version of TAALED Python package (Kyle et al., 2024).

#### 3.2.3. Lexical sophistication indices

Building on previous multivariate approaches in measuring lexical sophistication (e.g., Kyle & Crossley, 2015; Kyle & Crossley, 2016; Kyle et al., 2018; Yoon, 2018), we examined six categories of lexical sophistication: (1) word frequency, (2) word range, (3) psycholinguistic norms (familiarity, concreteness, imageability, and meaningfulness), (4) contextual distinctiveness, (5) age of acquisition, and (6) n-gram usage (frequency, range, and strength of association). All indices were computed using reference corpora and databases. Word frequency and range were logarithmically transformed to account for Zipfian distribution (Davies, 2009). Psycholinguistic norms were based on the MRC database (i.e., familiarity, concreteness, imageability; Coltheart, 1981) and concreteness ratings provided by Brysbaert et al. (2014). Contextual distinctiveness was measured using McD_CD (McDonald & Shillcock, 2001) and USF norms (Nelson et al., 2004). Age of acquisition values were drawn from two sources: Kuperman et al. (2012) and the VXGL project (Flor et al., 2024). Kuperman et al. (2012) provides AoA ratings for over 30,000 English content words, based on adult participants' estimates of when they learned each word. The VXGL project (Flor et al., 2024) offers grade-level mappings of English vocabulary based on U.S. grade-level expectations. We also calculated the strength of association for n-grams using multiple sub-indices (mutual information [MI], mutual information squared [MI2], T-score [T], DeltaP [DP], and approximate collexeme strength [AC]; Kyle et al., 2018). We used TAALES (version 2.8; Kyle et al., 2018) to compute the lexical sophistication indices, except for VXGL, which was added via an in-house Python package. In total, 34 lexical sophistication indices were calculated, as summarized in Table 3. More details on calculation methods are available in the TAALES package documentation (cf. https://www.linguisticanalysistools.org/taales.html). Note that spelling errors may result in unmatched words or n-grams in the reference corpora or databases, leading to a score of 0 for those cases and excluding them from the final score.

**Table 2**
Distribution of scores across writing tasks.

| Task type | | Score | | | | *M* | *SD* |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | | |
| Opinion | *n* | 142 | 235 | 208 | 52 | 2.27 | 0.90 |
| | (%) | (22.29) | (36.89) | (32.65) | (8.16) | | |
| Listen-Write | *n* | 155 | 198 | 181 | 88 | 2.32 | 1.00 |
| | (%) | (24.92) | (31.83) | (29.10) | (14.15) | | |

*Notes.* For the Listen-Write task, we excluded 15 responses that received a score of 0. As defined in the rubric (Appendix A), a score of 0 is assigned to responses that are blank, off-topic, in another language, that reject the prompt, consist of random keystrokes, or merely copy the prompt. Scores 1 through 4 reflect increasing levels of writing proficiency: score 1 indicates extremely limited and often inaccurate language use; score 2 reflects partial control of basic forms; score 3 represents generally accurate but less complex and varied language; and score 4 demonstrates advanced proficiency with a task-appropriate range of lexical and grammatical features.

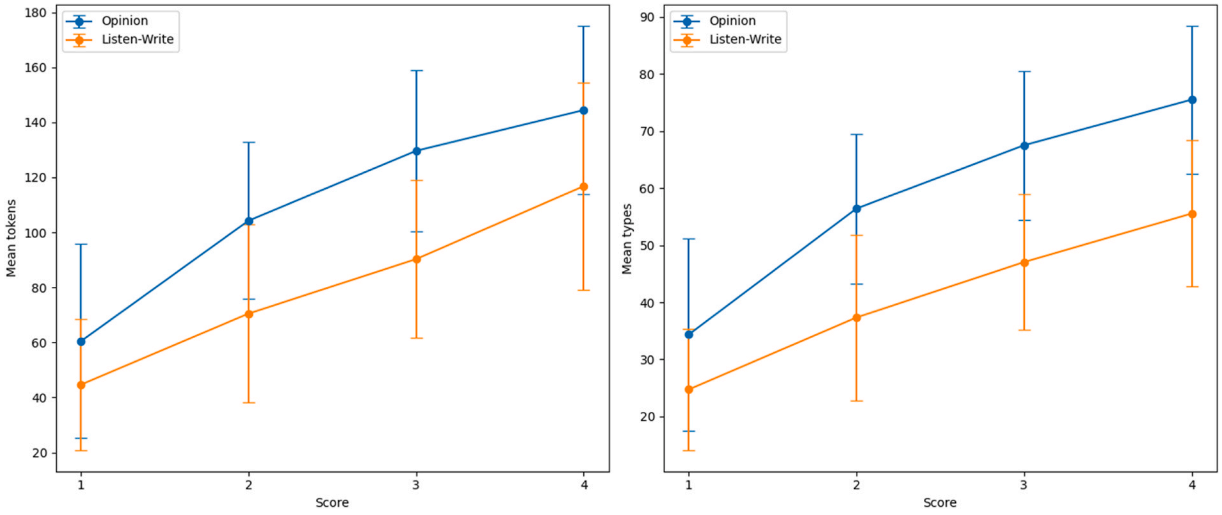

**Fig. 1.** Word token and type counts across scores and writing tasks.

**Table 3**
Summary of the calculated lexical sophistication indices.

| Category | Index | References |
|---|---|---|
| Word frequency | COCA_academic_Frequency_Log_AW | Davies (2009) |
| | COCA_academic_Frequency_Log_CW | |
| Word range | COCA_academic_Range_Log_AW | |
| | COCA_academic_Range_Log_CW | |
| Psycholinguistic norms | MRC_Familiarity_AW | Coltheart (1981) |
| | MRC_Concreteness_AW | |
| | MRC_Imageability_AW | |
| | MRC_Meaningfulness_AW | |
| | MRC_Familiarity_CW | |
| | MRC_Concreteness_CW | |
| | MRC_Imageability_CW | |
| | Brysbaert_Concreteness_AW | Brysbaert et al. (2014) |
| | Brysbaert_Concreteness_CW | |
| Contextual distinctiveness | USF_AW | Nelson et al. (2004) |
| | USF_CW | |
| | McD_CD_AW | McDonald and Shillcock (2001) |
| | McD_CD_CW | |
| Age of acquisition | AoA_AW | Kuperman et al. (2012) |
| | AoA_CW | |
| | VXGL | Flor et al. (2024) |
| n-gram frequency | COCA_academic_bi_Frequency_Log | Davies (2009) |
| | COCA_academic_tri_Frequency_Log | |
| n-gram range | COCA_academic_bi_Range_Log | |
| | COCA_academic_tri_Range_Log | |
| n-gram strength of association | COCA_academic_bi_MI | |
| | COCA_academic_bi_MI2 | |
| | COCA_academic_bi_T | |
| | COCA_academic_bi_DP | |
| | COCA_academic_bi_AC | |
| | COCA_academic_tri_MI | |
| | COCA_academic_tri_MI2 | |
| | COCA_academic_tri_T | |
| | COCA_academic_tri_DP | |
| | COCA_academic_tri_AC | |

*Notes.* AoA: age of acquisition; AW: All words; CW: content words; bi: bigram; tri: trigram; log: log-transformed; VXGL: vocabulary by (US-school) grade level.

### *3.2.4. Lexical density indices*

Compared to the other two types of lexical richness indices, calculating lexical density is relatively straightforward. It is operationally defined as the proportion of content words relative to the total number of words. In this study, we used two variations of this calculation, inspired by Crossley et al. (2019): (1) lexical_density_tokens, which is the percentage of text tokens (i.e., individual word occurrences) that are content words; and (2) lexical_density_types, which is the percentage of word types (i.e., unique words) that are content words. We employed TAACO (version 2.1.8; Crossley et al., 2019) to compute these indices.

### 3.3. Analytical approach

To answer the RQs, we first identified statistically meaningful features among 37 indices by examining their correlations, normality, and multicollinearity. We filtered indices that meaningfully correlated with the Opinion and Listen-Write task scores, discarding those that violated normality assumptions or showed negligible correlations ($|r| < 0.1$; Cohen, 2013). Among the multiple sub-indices measuring n-gram association strength (T, MI, MI2, DP, AC), we selected the strongest correlate for both bigrams and trigrams. We then assessed inter-correlations among the remaining indices, retaining only the feature with the stronger task-score correlation when two indices were highly correlated ($r > 0.7$).

For RQ1, we conducted Pearson correlations using the filtered indices and built preliminary regression models with the *lm* function in R (R Core Team, 2023). After confirming normality of residuals via Q-Q plots, we refined models using the Akaike Information Criterion (AIC), exploring predictor combinations with the *dredge* function (MuMIn package; Barton, 2009). We retained models with $\Delta\text{AIC} < 4$ (Akaike, 1974; Tan & Biswas, 2012) and selected the model with the lowest AIC for final evaluation. While some predictors did not reach conventional levels of statistical significance (e.g., $p > .05$), we retained them if they contributed meaningfully to overall model fit based on AIC comparisons. To answer RQ2, we used ANOVA to test for significant differences in the selected features across four score levels, followed by Tukey post-hoc tests (*multicomp* package; Hothorn et al., 2008) to identify specific score-level contrasts. For RQ3, we conducted a MANOVA with task type (Opinion vs. Listen-Write) as a fixed factor and the regression-selected predictors as dependent variables to examine how lexical richness features varied by task.

## 4. Results

### 4.1. RQ1. Relationship between lexical richness indices and task scores

After statistical filtering and addressing multicollinearity, we identified 10 indices for the Opinion task to include in the correlation and regression analyses. Table 4 presents the results of the correlation analysis, while Table 5 summarizes the best-fitting regression model for the Opinion task.

The best regression model accounted for 20 % of the variance in the task scores assigned by the raters. The results indicated that lexical diversity, age of acquisition (all words), n-gram strength of association, bigram range and word frequency (all words) are significant predictors of the Opinion task scores.

For the Listen-Write task, we identified 7 indices to include in the correlation and regression analyses. Table 6 shows the result of the correlation analysis, while Table 7 demonstrates the best-fitting regression model for the Listen-Write task scores.

The best regression model accounted for 24 % of the variance in the task scores assigned by the raters. The results indicate that lexical diversity, n-gram strength of association, contextual distinctiveness (content words), and trigram range were important predictors of the Listen-Write task scores.

### 4.2. RQ2. Lexical richness across different task scores

We further examined the indices included in the regression models using ANOVAs. Fig. 2 shows the distribution of six selected lexical richness indices across four scores for the Opinion task. Table 8 presents the mean differences in scores, alongside the adjusted *p*-values from the Tukey tests. The results showed that MATTR, with a large effect size ($\eta^2 = 0.16$), revealed distinct patterns for lower (1, 2) versus higher scores (3, 4), though no statistical difference was observed between scores 3 and 4. Indices with medium effect sizes, including AoA_AW ($\eta^2 = 0.06$), COCA_academic_tri_MI and COCA_academic_bi_MI ($\eta^2 = 0.05$), indicated some variation across proficiency scores, yet they did not consistently distinguish adjacent scores. Meanwhile, COCA_academic_bi_range_log and COCA_academic_frequency_log_AW ($\eta^2 = 0.02$), showed limited variation with a small effect size.

Fig. 3 illustrates the distribution of five selected lexical richness indices across four scores for the Listen-Write task. Table 9 presents

**Table 4**
Correlations of the lexical indices (Opinion).

| Index | *r* | *p* |
|---|---|---|
| MATTR | 0.385 | < .001 |
| AoA_AW | 0.240 | < .001 |
| COCA_academic_tri_MI | 0.210 | < .001 |
| COCA_academic_bi_MI | 0.207 | < .001 |
| Brysbaert_Concreteness _AW | 0.138 | < .001 |
| lexical_density_tokens | 0.136 | < .001 |
| COCA_academic_Frequency_Log_AW | −0.134 | < .001 |
| COCA_academic_bi_Range_Log | −0.139 | < .001 |
| USF_CW | −0.166 | < .001 |
| MRC_familiarity_AW | −0.191 | < .001 |

**Table 5**
Best regression model with selected predictors (Opinion).

| | Estimates | SE | *t* | *p* | Relative importance |
|---|---|---|---|---|---|
| (Intercept) | −10.447 | 0.369 | −3.723 | < .001 | |
| MATTR | 7.757 | 0.828 | 9.371 | < .001 | 0.602 |
| AoA_AW | 0.564 | 0.126 | 4.492 | < .001 | 0.167 |
| COCA_academic_tri_MI | 0.082 | 0.042 | 1.939 | .053 | 0.086 |
| COCA_academic_bi_MI | 0.184 | 0.808 | 2.292 | .022 | 0.084 |
| COCA_academic_bi_Range_Log | −0.318 | 0.212 | −1.496 | .135 | 0.032 |
| COCA_academic_Frequency_Log_AW | 0.564 | 0.231 | 2.443 | .015 | 0.029 |

*Note*. multiple $R^2 = 0.21$, adjusted $R^2 = 0.20$

**Table 6**
Correlations of the lexical indices (Listen-Write).

| Index | *r* | *p* |
|---|---|---|
| MATTR | 0.441 | < .001 |
| COCA_academic_tri_MI | 0.303 | < .001 |
| COCA_academic_bi_MI | 0.251 | < .001 |
| COCA_academic_tri_Range_Log | −0.144 | < .001 |
| COCA_academic_bi_Frequency_Log | −0.134 | < .001 |
| USF_CW | −0.154 | < .001 |
| MRC_Concreteness_CW | −0.160 | < .001 |

**Table 7**
Best regression model with selected predictors (Listen-Write).

| | Estimates | SE | *t* | *p* | Relative importance |
|---|---|---|---|---|---|
| (Intercept) | −10.447 | 0.369 | −3.723 | < .001 | |
| MATTR | 4.424 | 0.432 | 10.241 | < .001 | 0.622 |
| COCA_academic_tri_MI | 0.073 | 0.038 | 1.908 | .057 | 0.159 |
| COCA_academic_bi_MI | 0.163 | 0.093 | 1.750 | .081 | 0.103 |
| USF_CW | −0.014 | 0.004 | −3.868 | < .001 | 0.078 |
| COCA_academic_tri_Range_Log | −0.110 | 0.069 | −1.595 | .111 | 0.037 |

*Notes*. multiple $R^2 = 0.25$, adjusted $R^2 = 0.24$

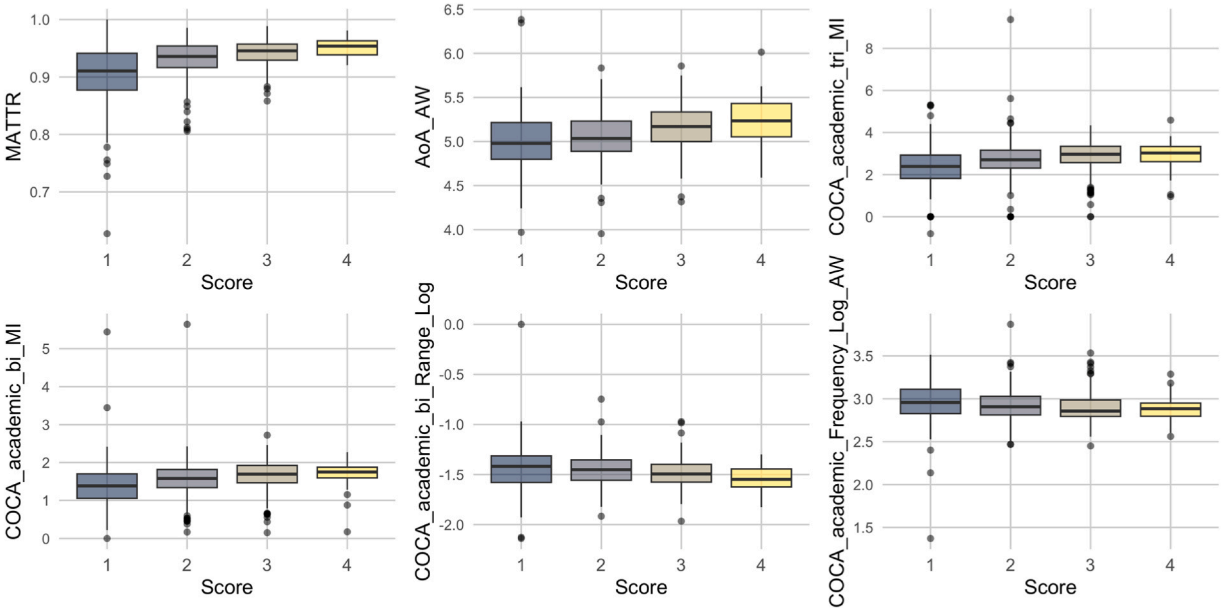

**Fig. 2.** Distribution of lexical richness indices across scores (Opinion).

**Table 8**
Results of ANOVA and Tukey post-hoc test across task scores (Opinion).

| Index | F | *p* | η2 | Post-hoc | Mean difference | *p* adj |
|---|---|---|---|---|---|---|
| MATTR | 41.04 | < .001*** | 0.16 | 2–1 | 0.028 | *** |
| | | | | 3–1 | 0.039 | *** |
| | | | | 4–1 | 0.049 | *** |
| | | | | 3–2 | 0.011 | ** |
| | | | | 4–2 | 0.021 | ** |
| | | | | 4–3 | 0.010 | |
| AoA_AW | 13.31 | < .001*** | 0.06 | 2–1 | 0.049 | |
| | | | | 3–1 | 0.148 | *** |
| | | | | 4–1 | 0.224 | *** |
| | | | | 3–2 | 0.099 | ** |
| | | | | 4–2 | 0.175 | *** |
| | | | | 4–3 | 0.076 | |
| COCA_academic_tri_MI | 11.02 | < .001*** | 0.05 | 2–1 | 0.347 | *** |
| | | | | 3–1 | 0.510 | *** |
| | | | | 4–1 | 0.566 | *** |
| | | | | 3–2 | 0.163 | |
| | | | | 4–2 | 0.219 | |
| | | | | 4–3 | 0.056 | |
| COCA_academic_bi_MI | 10.06 | < .001*** | 0.05 | 2–1 | 0.169 | . |
| | | | | 3–1 | 0.267 | *** |
| | | | | 4–1 | 0.326 | *** |
| | | | | 3–2 | 0.098 | |
| | | | | 4–2 | 0.157 | |
| | | | | 4–3 | 0.059 | |
| COCA_academic_bi_Range_Log | 4.837 | .002** | 0.02 | 2–1 | −0.039 | |
| | | | | 3–1 | −0.096 | ** |
| | | | | 4–1 | −0.031 | |
| | | | | 3–2 | −0.088 | ** |
| | | | | 4–2 | −0.057 | |
| | | | | 4–3 | −0.039 | |
| COCA_academic_Frequency_Log_AW | 3.967 | .008** | 0.02 | 2–1 | −0.034 | |
| | | | | 3–1 | −0.064 | * |
| | | | | 4–1 | −0.078 | . |
| | | | | 3–2 | −0.030 | |
| | | | | 4–2 | −0.044 | |
| | | | | 4–3 | −0.013 | |

*Notes*. According to Cohen's (2013) benchmarks, an $\eta^2$ of 0.01 indicates a small effect size, 0.06 represents a medium effect size, and 0.14 or higher signifies a large effect size, reflecting the proportion of variance explained by the independent variable; Sig. codes for *p*-value (adjusted): '***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1, ' ' 1

the results of the ANOVA and post-hoc Tukey tests for the task. The findings indicate that MATTR, with a large effect size ($\eta^2 = 0.21$), revealed distinct patterns across task scores, though it showed limited distinction between score groups 3 and 4. Indices with medium effect sizes, such as COCA_academic_tri_MI ($\eta^2 = 0.11$) and COCA_academic_bi_MI ($\eta^2 = 0.07$), indicated some variation across task scores but did not consistently differentiate adjacent scores within the upper groups. Meanwhile, indices such as USF_CW and COCA_academic_tri_range_log, which had a small effect size ($\eta^2 = 0.03$), showed limited variation across scores.

### *4.3. RQ3. Differences between Opinion and Listen-Write tasks*

We conducted a MANOVA to explore how the lexical richness indices differ between Opinion and Listen-Write tasks. We selected the eight indices (MATTR, COCA_academic_bi_MI, COCA_academic_tri_MI, AoA_AW, COCA_academic_bi_Range_Log, COCA_academic_Frequency_Log_AW, USF_CW, and COCA_academic_tri_Range_Log) based on their statistical significance in the regression models for predicting task scores in at least one task type. Additionally, we included Tokens (i.e., the total number of words) and Types (i.e., the total number of unique words) as reference measures. As shown in Table 10, the MANOVA results indicate large effects (Cohen, 2013) for Types, MATTR, and COCA_academic_bi_MI, medium effects for Tokens and AoA_AW, and small effect for COCA_academic_tri_MI, USF_CW, and COCA_academic_tri_Range_Log.

## 5. Discussion

This study examined lexical richness in young L2 learners' written productions from the Opinion and Listen-Write tasks, sourced from the standardized and timed context of the TOEFL Junior Writing test. Lexical richness served as the conceptual framework and was operationalized through multiple indices, including diversity, sophistication, and density, which collectively measure the breadth and depth of word usage in written production.

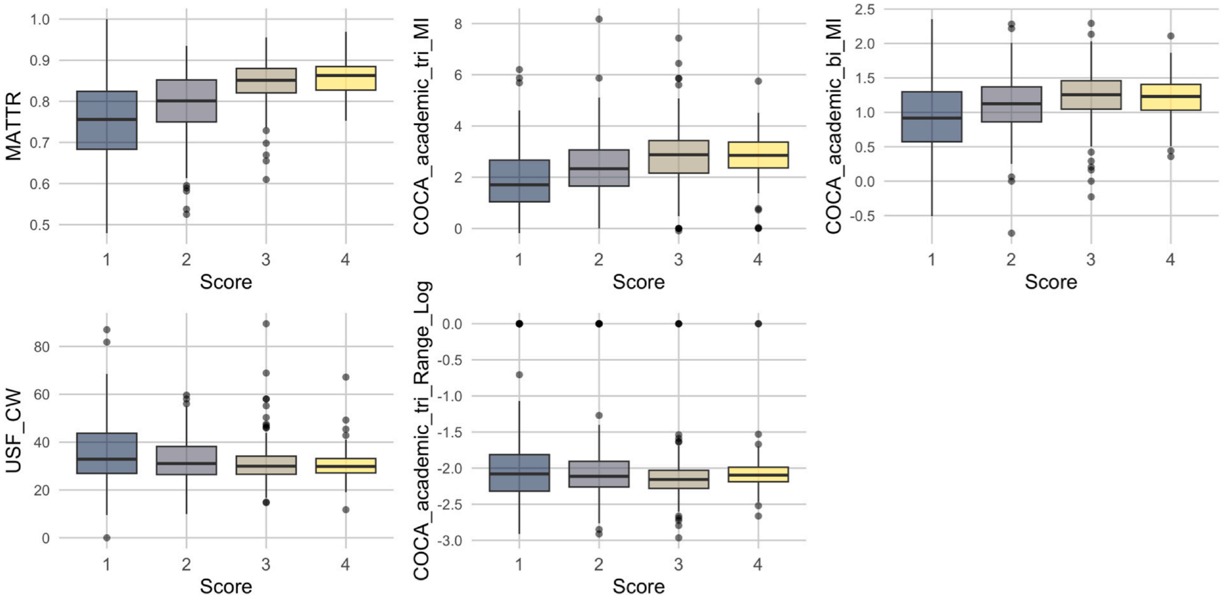

**Fig. 3.** Distribution of lexical richness indices across scores (Listen-Write).

**Table 9**
Results of ANOVA and Tukey post-hoc test across task scores (Listen-Write).

| Index | F | *p* | η2 | Post-hoc | Mean difference | *p* adj |
|---|---|---|---|---|---|---|
| MATTR | 53.40 | < .001*** | 0.21 | 2–1 | 0.040 | *** |
| | | | | 3–1 | 0.093 | *** |
| | | | | 4–1 | 0.103 | *** |
| | | | | 3–2 | 0.053 | *** |
| | | | | 4–2 | 0.063 | *** |
| | | | | 4–3 | 0.010 | |
| COCA_academic_tri_MI | 23.37 | < .001*** | 0.11 | 2–1 | 0.473 | *** |
| | | | | 3–1 | 0.015 | *** |
| | | | | 4–1 | 0.944 | *** |
| | | | | 3–2 | 0.542 | *** |
| | | | | 4–2 | 0.470 | ** |
| | | | | 4–3 | −0.071 | |
| COCA_academic_bi_MI | 16.55 | < .001*** | 0.07 | 2–1 | 0.187 | *** |
| | | | | 3–1 | 0.305 | *** |
| | | | | 4–1 | 0.295 | *** |
| | | | | 3–2 | 0.118 | * |
| | | | | 4–2 | 0.108 | |
| | | | | 4–3 | −0.010 | |
| USF_CW | 5.74 | < .001*** | 0.03 | 2–1 | −2.915 | * |
| | | | | 3–1 | −3.697 | ** |
| | | | | 4–1 | −4.681 | ** |
| | | | | 3–2 | −0.782 | |
| | | | | 4–2 | −1.766 | |
| | | | | 4–3 | −0.984 | |
| COCA_academic_tri_Range_Log | 6.07 | < .001*** | 0.03 | 2–1 | −0.115 | |
| | | | | 3–1 | −2.263 | *** |
| | | | | 4–1 | −0.189 | . |
| | | | | 3–2 | −0.148 | . |
| | | | | 4–2 | −0.074 | |
| | | | | 4–3 | 0.074 | |

### *5.1. RQ1*

The results showed that higher-scoring texts displayed a similar lexical richness profile that has been documented for more mature L2 writers in previous studies: they contained a broader range of vocabulary, relied on less frequent but more strongly associated n-grams, and showed no advantage in lexical density.

First, lexical diversity, as captured by the length-stable MATTR index, was positively related to proficiency scores in both tasks. This

**Table 10**
MANOVA results comparing linguistic features across task types.

| Index | Opinion<br>*Mean (SD)* | Listen-Write<br>*Mean (SD)* | *F* | $\eta^2$ | *p* |
|---|---|---|---|---|---|
| Tokens | 106.053 (41.218) | 75.206<br>(38.838) | 188.993 | 0.129 | < .001 |
| Types | 56.688<br>(19.311) | 39.055<br>(16.778) | 302.645 | 0.192 | < .001 |
| MATTR | 0.931<br>(0.039) | 0.808<br>(0.090) | 985.129 | 0.436 | < .001 |
| COCA_academic_bi_MI | 1.563<br>(0.503) | 1.123<br>(0.485) | 252.020 | 0.165 | < .001 |
| COCA_academic_tri_MI | 2.723<br>(0.891) | 2.398<br>(1.260) | 28.192 | 0.022 | < .001 |
| AoA_AW | 5.090<br>(0.290) | 5.280<br>(0.561) | 57.714 | 0.043 | < .001 |
| COCA_academic_bi_Range_Log | −1.470<br>(0.177) | −1.454<br>(0.270) | 1.573 | 0.001 | < .001 |
| COCA_academic_Frequency_Log_AW | 2.918<br>(0.193) | 2.867<br>(0.302) | 12.892 | 0.001 | < .001 |
| USF_CW | 35.246<br>(0.730) | 32.616<br>(11.032) | 20.368 | 0.016 | < .001 |
| COCA_academic_tri_Range_Log | −2.096<br>(0.346) | −1.982<br>(0.623) | 16.235 | 0.013 | < .001 |

finding aligns with previous studies, which observed that more advanced L2 users tend to use a more diverse vocabulary in independent (e.g., Kyle et al., 2021; Woods et al., 2023) and integrated writing tasks (e.g., Cumming et al., 2005; Cai & Yan, 2024). The present findings extend that pattern downward to younger writers, demonstrating that diversity remains a reliable marker of quality even in very short, timed compositions.

Lexical sophistication added a complementary layer of differentiation. N-gram indices, particularly bigram and trigram strength of association (as measured by MI scores), were significant predictors in both tasks, indicating that advanced L2 learners favored phraseological combinations that are more strongly associated in advanced writers' norms. This finding aligns with previous research on lexical sophistication among university-level L2 learners (e.g., Kim et al., 2018; Yoon, 2018), which highlights the importance of n-gram use in L2 argumentative writing. Interestingly, in contrast to an earlier finding that lexical sophistication played a limited role in the integrated writing task (Kyle & Crossley, 2016), our results showed that n-gram strength of association significantly predicted scores on the Listen-Write task. Task-specific patterns also emerged. The Opinion task showed that word frequency and bigram range indices were significant predictors, suggesting that advanced learners used less frequent words and/or more context-specific bigrams, which aligns with the previous findings (e.g., Yoon, 2018; Kyle & Crossley, 2015). In the Listen-Write responses, by contrast, word frequency was not informative. Instead, scores were more closely tied to the use of context-distinct words and a limited trigram range, implying that high scorers mined the listening source for precise lexical items and formulaic trigrams.

Lexical density did not correlate meaningfully with proficiency in either task, reinforcing the consensus that density is a weak indicator of productive ability (Bulté et al., 2024; Engber, 1995; Kyle, 2019). Given the brevity of the responses (i.e., typically 100–150 words produced in ten minutes), writers had little opportunity to deploy cohesion-building lexical material, limiting the diagnostic value of density measures.

Note that although word token and type counts were not direct indicators of lexical richness in this study, Fig. 1 suggests that they may still function as useful diagnostic features. This observation aligns with previous studies involving young L2 learners, which have shown that simple quantity-based measures can be meaningful predictors of early writing development (e.g., Kyle et al., 2021; Treffers-Daller, 2013).

## 5.2. RQ2

The results showed that lexical richness indices identified as meaningful in RQ1could differentiate proficiency scores to some extent. Taken at face value, the findings mirror the test rubric (Appendix A): writers who earn a 4 do, on average, display "lexical variation appropriate for the task," whereas those who receive a 1 rely on "extremely limited vocabulary used incorrectly." However, even the most significant lexical richness index in this study (i.e., MATTR) struggles to distinguish between scores of 3 and 4.

In the Opinion task, students who received a score of 4 produced a mean MATTR of.953 (*SD* =.017), while those who scored a 3 clustered just below, at.943 (*SD* =.022). Because the two standard deviation bands overlap (.936–.970 for score 4 vs.921–.965 for score 3), many score 3 responses are statistically indistinguishable from score 4 responses in terms of lexical diversity. A similar pattern appeared in the Listen-Write task: dispersion steadily narrowed from the lowest band (*SD* =.160 at score 0) to the highest (*SD* =.041 at score 4). Yet, the score-4 interval (.815–.896, around a mean of.856) still overlapped with the score-3 interval (.791–.900, mean.845). In short, even among top-scoring responses, levels of lexical diversity varied considerably. Some writers achieved a score of 4 despite having lexical richness values comparable to those of score 3 writers, while others displayed much greater variety. This pattern suggests that lexical richness indices alone, even the most predictive ones, do not fully account for how raters distinguish between high-scoring responses, particularly at the upper end of the scale.

One possibility is that, due to the nature of the holistic rubric, raters may have considered a range of features such as grammatical accuracy, lexical use, coherence, and task relevance in addition to lexical richness when assigning a score. If two essays receive a score of 4 for slightly different reasons (e.g., one was strong for lexical richness, the other for strong rhetorical structure), the lexical richness examined in this study may not account for that distinction. Future study could model how lexical, grammatical, and discursive dimensions interact at upper score levels. Using multi-trait rubrics or analytic scoring would better reveal which configurations of linguistic strengths raters reward when assigning top scores.

### *5.3. RQ3*

The results showed that young L2 learners demonstrated significantly greater lexical productivity and richness in the Opinion task than in the Listen-Write task. The differences were particularly pronounced for the number of unique words (Types), the diversity of word use (MATTR), and the strength of association in bigrams (COCA_academic_bi_MI), all showing large effect sizes. Moderate differences were also observed for the total number of words (Tokens) and the age of acquisition of words (AoA_AW). Smaller but significant differences appeared in trigram association strength (COCA_academic_tri_MI), semantic network connectivity (USF_CW), and trigram range (COCA_academic_tri_Range_Log).

To accomplish the Listen-Write task, the learners summarized a short talk about an academic topic, which narrows the range of expected lexical items, as learners must repeat lexical items from the talk to effectively write a response. Furthermore, the Listen-Write task requires learners to take notes while listening and then use their notes to write a summary. Limitations in listening or note-taking skills, as well as the increased cognitive load of an integrated-skills task are thus factors that may increase the difficulty of this task. We therefore posit that the genre and difficulty of the task limited the lexical items learners were able to use, pushing them toward more frequent, easily accessible words or collocations. By contrast, in the Opinion task, it may have been easier for learners to tap into personal experience as well as to use broader and more varied lexical items.

Overall, these patterns parallel findings with older L2 users and reinforce the view that open-ended argumentative writing elicits broader and more sophisticated lexical choices than integrated writing tasks (Crossley et al., 2014; Kyle & Crossley, 2016).

## 6. Implications and limitations

Beyond the findings discussed, this study offers broader implications for writing assessment development and score interpretation. The use of lexical indices as predictors of young English learners' proficiency contributes to the growing body of validity evidence supporting lexical features as meaningful indicators of writing ability. These findings may inform rater training by providing empirical support for incorporating lexical diversity and sophistication into scoring rubrics or exemplar materials. More broadly, they underscore the value of using automated lexical indices to support scoring calibration in standardized writing assessments targeting younger L2 learners. However, the results also highlight the limits of relying on lexical measures alone for capturing the full range of proficiency, particularly at the upper end of the scale—a point further addressed in the limitations below.

Several limitations may affect the generalizability and interpretation of this study's findings. First, the analysis centered on two specific writing tasks administered under the standardized, timed conditions of the TOEFL Junior Writing test. While this design ensured consistency, it also limited the scope of writing contexts examined. Nevertheless, the observed differences in lexical indices across the two task types underscore the value of incorporating diverse task formats when assessing students' lexical richness in relation to their writing abilities. Future studies could extend this work by including a broader range of writing genres (e.g., personal narratives) to determine whether patterns of lexical richness hold across task types.

Second, the study employed specific measures of lexical richness, including lexical diversity (e.g., MATTR), lexical sophistication (e.g., n-gram indices), and lexical density. While these indices effectively captured certain aspects of lexical use, they are constrained by the properties of the reference database (e.g., COCA-academic, USF), which may not fully represent the lexical profiles of young L2 learners or domain-specific language use (cf. Brysbaert & New, 2009; Keuleers & Balota, 2015). To improve alignment between reference data and learner output, future research should consider developing or adopting proficiency-appropriate reference corpora (cf. Monteiro et al., 2020). For example, large peer writing corpora from similar age groups or curated collections of graded readers and instructional materials could provide more representative baselines.

Third, the reliability of automated analyses may have been affected by spelling errors. As noted earlier, unmatched words or n-grams due to misspellings were excluded from the analyses, which may have disproportionately affected low-proficiency learners who tend to produce more spelling errors (e.g., De Wilde, 2023; Verspoor et al., 2012). Although spelling accuracy was not a focal construct in this study, its influence on lexical indices warrants further investigation. Future research should incorporate lexical accuracy and explore how it interacts with other lexical features across proficiency levels.

Fourth, while the findings highlighted the importance of certain lexical richness indices (e.g., MATTR, n-gram strength of association), they also pointed to challenges in distinguishing higher proficiency scores, as discussed earlier. This suggests a need to supplement lexical measures with other linguistic indicators, such as syntactic complexity or discourse structure (cf. Wolf et al., 2018). In addition, in source-based tasks like Listen-Write, further insights may be gained by comparing learners' lexical choices with the source materials (e.g., listening passages) (cf. Cai & Yan, 2024; Kyle, 2020).

## CRediT authorship contribution statement

**Hakyung Sung:** Writing – review & editing, Writing – original draft, Visualization, Conceptualization, Investigation. **Mikyung Kim Wolf:** Conceptualization, Data curation, Supervision, Writing – original draft, Writing – review & editing. **Michael Suhan:** Data curation, Writing – review & editing. **Kristopher Kyle:** Methodology, Writing – review & editing.

## Declaration of Competing Interest

The authors declare that there are no conflicts of interest regarding the publication of this study.

## Acknowledgements

We extend our sincere gratitude to the ETS TOEFL research team for supplying the essential data for this study. We also appreciate the ETS R&D technical reviewers for their valuable feedback and constructive suggestions.

## Appendix

A. TOEFL Junior writing scoring rubrics

**Table 11**
Scoring rubric (Opinion)

| Score | Development and language use descriptors |
|---|---|
| 4 | A typical response at this level is characterized by the following:<br>- states a position on the topic<br>- provides support for the position, with specific details and/or examples<br>- is mostly well organized and coherent<br>- shows lexical variation appropriate for the task<br>- displays a varied sentence structure appropriate for the task<br>- may contain minor errors but they do not interfere with meaning or clarity |
| 3 | A typical response at this level is characterized by the following:<br>- states a position on the topic<br>- provides support for the stated position, but may have difficulty doing so fully<br>- is generally well organized, with an occasional lapse of clarity when connecting ideas<br>- shows some lexical variation appropriate for the task<br>- may display some variation in sentence structure appropriate for the task<br>- may contain some errors that occasionally interfere with meaning |
| 2 | A typical response at this level is characterized by the following:<br>- states a position on the topic, but provides inadequate/incomplete support<br>- only vaguely implies a position on the topic, and provides inadequate/incomplete support<br>- connections between ideas are attempted, but are sometimes unclear or missing<br>- shows little lexical variation (e.g., vocabulary is simple and repetitive), or frequently uses vocabulary incorrectly<br>- shows little variation in sentence structure (e.g., sentences are mostly simple and short), and shows little control of sentence structures<br>- may contain errors that frequently interfere with meaning |
| 1 | A typical response at this level is characterized by the following:<br>- states a position but provides incoherent or no support OR<br>- does not state a position, or makes only a minimal connection to the prompt and provides minimal or no support<br>- is generally unorganized and incoherent<br>- displays extremely limited vocabulary that is frequently used incorrectly<br>- uses mostly incorrect sentence structures<br>- displays many errors that seriously interfere with meaning |
| 0 | Only copies words from the prompt, rejects the prompt, is completely off topic, consists of keystroke characters, is written in a foreign language, or is blank. |

*Source.* https://www.ets.org/toefl/junior/scoring-reporting.html (copyright ETS)

**Table 12**
Scoring rubric (Listen-Write)

| Score | Development and language use descriptors |
|---|---|
| 4 | A typical response at this level is characterized by the following:<br>- accurately provides all key points<br>- provides support using relevant details from the talk<br>- is mostly well organized and coherent<br>- shows lexical variation appropriate for the task<br>- displays a varied sentence structure appropriate for the task<br>- may contain errors but they do not interfere with meaning or clarity |
| 3 | A typical response at this level is characterized by the following:<br>- accurately provides most key points<br>- provides some supporting details from the talk<br>- is generally organized, with an occasional lapse of clarity when connecting ideas<br>- shows some lexical variation appropriate for the task<br>- may display some varied sentence structure appropriate for the task<br>- may contain errors that occasionally interfere with meaning |
| 2 | A typical response at this level is characterized by the following:<br>- provides some accurate content from the key points<br>- provides minimal or no supporting details from the talk<br>- connections between ideas are attempted but are often unclear or missing<br>- shows little lexical variation (e.g., vocabulary is simple and repetitive), or frequently uses vocabulary incorrectly<br>- shows little variation in sentence structure (e.g., sentences are mostly simple and short), or shows little control of sentence structures<br>- may contain errors that frequently interfere with meaning |
| 1 | A typical response at this level is characterized by the following:<br>- provides minimal or no content from the key points<br>- does not provide details beyond those shown in the visuals<br>- provides incoherent or no support for any of the points<br>- is generally unorganized and incoherent<br>- displays extremely limited vocabulary that is frequently used incorrectly<br>- uses mostly incorrect sentence structures<br>- displays many errors that seriously interfere with meaning |
| 0 | Only copies words from the prompt, rejects the prompt, is completely off topic, consists of keystroke characters, is written in a foreign language, or is blank. |

*Source.* https://www.ets.org/toefl/junior/scoring-reporting.html (copyright ETS)

B. MATTR optimization

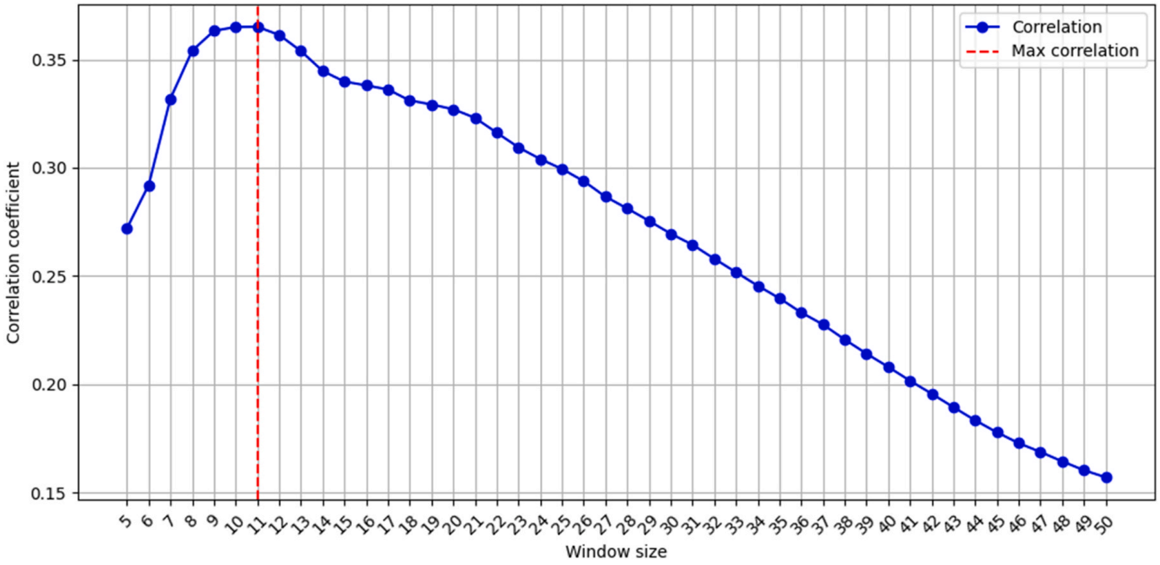

**Fig. 4.** MATTR optimization plot (Opinion)

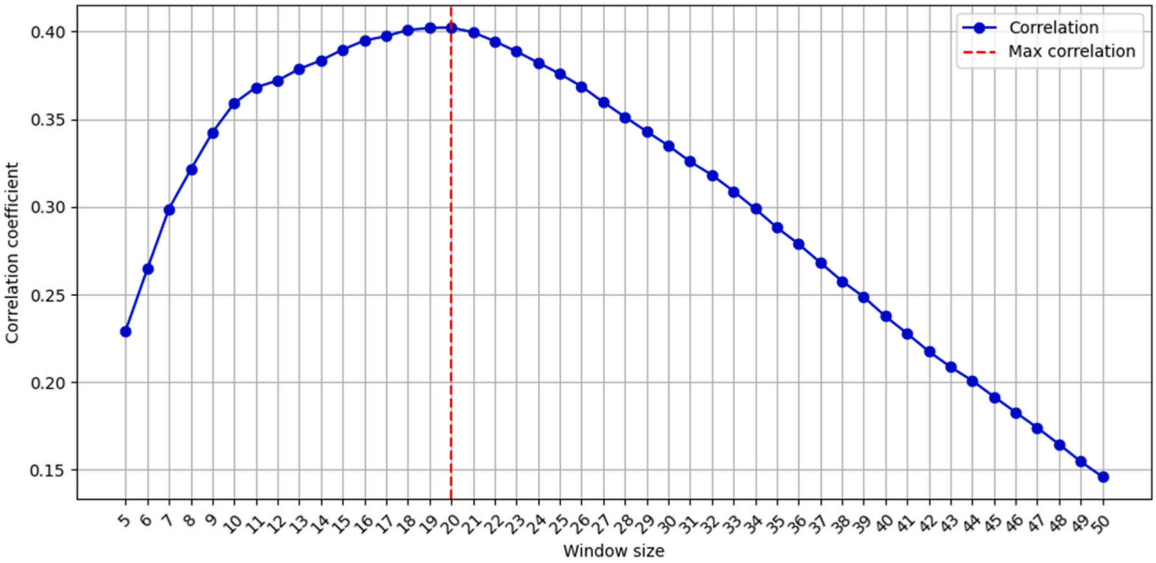

**Fig. 5.** MATTR optimization plot (Listen-Write)

*Notes.* These plots show the correlations between MATTR scores and proficiency scores across different window sizes (ranging from 5 to 50). The blue line represents the correlation coefficient at each window size, and the red dashed line marks the window size that yielded the highest correlation.

## Data Availability

The authors do not have permission to share data.

## References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Barton, K. (2009). *MuMIn: Multi-model inference*. R package. Retrieved from http://r-forge.r-project.org/projects/mumin/.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity accuracy and fluency in SLA* (pp. 21–46). Benjamins.

Bulté, B., Housen, A., & Pallotti, G. (2024). Complexity and difficulty in second language acquisition: A theoretical and methodological overview. *Language Learning*.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*, 904–911.

Cai, H., & Yan, X. (2024). Examining the direct and indirect impacts of verbatim source use on linguistic complexity in integrated argumentative writing assessment. *Assessing Writing, 61*, Article 100868.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2011). *Building a validity argument for the Test of English as a Foreign Language*. Routledge.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A, 33*(4), 497–505.

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

Crossley, S., Clevinger, A., & Kim, Y. (2014). The role of lexical properties and cohesive devices in text integration and their effect on human ratings of speaking proficiency. *Language Assessment Quarterly, 11*(3), 250–270.

Crossley, S., Kyle, K., & Dascalu, M. (2019). The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods, 51*, 14–27.

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

Crossley, S., Salsbury, T., & McNamara, D. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning, 59*(2), 307–334.

Crossley, S., Salsbury, T., & McNamara, D. S. (2012). Predicting the proficiency level of language learners using lexical indices. *Language Testing, 29*(2), 243–263.

Crossley, S., Salsbury, T., & Mcnamara, D. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics, 36*(5), 570–590.

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL®. *Assessing Writing, 10*(1), 5–43.

Davies, M. (2009). The 385+ million word corpus of contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics, 14*(2), 159–190.

DeKeyser, R. (2005). What makes learning second-language grammar difficult? A review of issues. *Language Learning, 55*(S1), 1–25.

DeKeyser, R. (2016). Of moving targets and chameleons: Why the concept of difficulty is so hard to pin down. *Studies in Second Language Acquisition, 38*(2), 353–363.

De Wilde, V. (2023). Lexical characteristics of young L2 English learners' narrative writing at the start of formal instruction. *Journal of Second Language Writing, 59*, Article 100960.

Durrant, P., & Brenchley, M. (2019). Development of vocabulary sophistication across genres in English children's writing. *Reading and Writing, 32*(8), 1927–1953.

Durrant, P., & Brenchley, M. (2023). Development of noun phrase complexity across genres in children's writing. *Applied Linguistics, 44*(2), 239–264.

ETS. (2019). *TOEFL Junior framework and test development*. ⟨https://www.ets.org/content/dam/ets-india/pdfs/toefl/toefl-ibt-insight-s1v7.pdf⟩.

ETS. (2023). *TOEFL Junior test taker handbook*. ⟨https://www.ets.org/pdfs/toefl/toefl-junior-handbook.pdf⟩.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing, 4*(2), 139–155.

Ferris, D. R. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly, 28*(2), 414–420.

Flor, M., Holtzman, S., Deane, P., & Bejar, I. (2024). Mapping of American English vocabulary by grade levels. *ITL-International Journal of Applied Linguistics*.

Gregori-Signes, C., & Clavel-Arroitia, B. (2015). Analysing lexical density and lexical diversity in university students' written discourse. *Procedia-Social and Behavioral Sciences, 198*, 546–556.

Halliday, M. A. K. (1989). *Spoken and written language*. Oxford, England: Oxford University Press.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal: Journal of Mathematical Methods in Biosciences, 50*(3), 346–363.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing, 19*(1), 57–84.

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning, 63*, 87–106.

Keuleers, E., & Balota, D. A. (2015). Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments. *Quarterly Journal of Experimental Psychology, 68*(8), 1457–1468.

Kim, J. Y. (2014). Predicting L2 writing proficiency using linguistic complexity measures: A corpus-based study. *English Teaching, 69*(4), 27–51.

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

Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System, 40*(4), 554–564.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*, 978–990.

Kyle, K. (2019). Measuring lexical richness. *The Routledge handbook of vocabulary studies* (pp. 454–476). Routledge.

Kyle, K. (2020). The relationship between features of source text use and integrated writing quality. *Assessing Writing, 45*, Article 100467.

Kyle, K., & Eguchi, M. (2023). Assessing spoken lexical and lexicogrammatical proficiency using features of word, bigram, and dependency bigram use. *The Modern Language Journal, 107*(2), 531–564.

Kyle, K., & Crossley, S. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49*(4), 757–786.

Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing, 34*, 12–24.

Kyle, K., Crossley, S., & Jarvis, S. (2021). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly, 18*(2), 154–170.

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods, 50*, 1030–1046.

Kyle, K., Sung, H., Eguchi, M., & Zenker, F. (2024). Evaluating evidence for the reliability and validity of lexical diversity indices in L2 oral task responses. *Studies in Second Language Acquisition, 46*(1), 278–299.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*(3), 307–322.

Linnarud, M. (1986). *Lexis in composition: a performance analysis of Swedish learners' written English*. Lund: CWK Gleerup.

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.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal, 96*(2), 190–208.

Maamuujav, U., Olson, C. B., & Chung, H. (2021). Syntactic and lexical features of adolescent L2 students' academic writing. *Journal of Second Language Writing, 53*, Article 100822.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development* (pp. 16–30). UK: Palgrave Macmillan.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: Avalidation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*(2), 381–392.

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech, 44*(3), 295–322.

Meara, P. (2005). Designing vocabulary tests for English, Spanish and other languages. In C. Butler, M. de los Ángeles Gómez González, & S. Doval-Suárez (Eds.), *The dynamics of Language use: Functional and contrastive perspectives* (pp. 271–285). John Benjamins Publishing Company.

Meara, P., & Bell, H. (2001). P-Lex: A simple and effective way of describing the lexical characteristics of short L2 tests. *Prospect, 16*(3), 5–19.

Monteiro, K. R., Crossley, S. A., & Kyle, K. (2020). In search of new benchmarks: Using L2 lexical frequency and contextual diversity indices to assess second language writing. *Applied Linguistics, 41*(2), 280–300.

Nation, I. S. (2005). Teaching and learning vocabulary. *Handbook of research in second language teaching and learning* (pp. 581–595). Routledge.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, Computers, 36*(3), 402–407.

Norris, J., & Ortega, L. (2003). Defining and measuring SLA. *The Handbook of Second Language Acquisition*, 716–761.

O'Loughlin, K. (1995). Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing, 12*(2), 217–237.

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.

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research, 31*(1), 117–134.

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from: ⟨https://www.R-project.org/⟩.

Schleppegrell, M. J. (2004). *The language of schooling: A functional linguistics perspective*. New York, NY: Routledge.

Tan, M. Y. J., & Biswas, R. (2012). The reliability of the akaike information criterion method in cosmological model selection. *Monthly Notices of the Royal Astronomical Society, 419*(4), 3292–3303.

Treffers-Daller, J. (2013). Measuring lexical diversity among L2 learners of French: An exploration of the validity of D, MTLD and HD-D as measures of language ability. *Vocabulary Knowledge* (pp. 79–104). John Benjamins.

Ure, J. (1971). Lexical density and register differentiation. *Applications of Linguistics, 23*(7), 443–452.

Verspoor, M., Schmid, M. S., & Xu, X. (2012). A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing, 21*(3), 239–263.

Wolf, M. K. (2024). Assessment of young language learners. In In. A. J. Kunnan (Ed.), *The Concise Companion to Language Assessment* (pp. 312–325). Wiley-Blackwell.

Wolf, M. K., Oh, S., Wang, Y., & Tsutagawa, F. S. (2018). Young adolescent EFL students' writing skill development: Insights from assessment data. *Language Assessment Quarterly, 15*(4), 311–329.

Woods, K., Hashimoto, B., & Brown, E. K. (2023). A multi-measure approach for lexical diversity in writing assessments: Considerations in measurement and timing. *Assessing Writing, 55*, Article 100688.

Yoon, H. J. (2018). The development of ESL writing quality and lexical proficiency: Suggestions for assessing writing achievement. *Language Assessment Quarterly, 15*(4), 387–405.

Yule, G. U. (1944). *The statistical study of literary vocabulary*. Cambridge: Cambridge University Press.

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing, 47*, Article 100505.

**Hakyung Sung** is an incoming Assistant Professor in the Department of Psychology at the Rochester Institute of Technology. She received her Ph.D. in Linguistics from the University of Oregon. Her research sits at the intersection of applied and computational linguistics, with a particular focus on using natural language processing tools and large language models to examine lexicogrammatical complexity and sophistication in second language production.

**Mikyung Kim Wolf** is a principal research scientist at ETS. She received her Ph.D. in Applied Linguistics specializing in language assessment from UCLA. Her research interests include technology-enhanced language assessments, formative assessment, the use of AI in language education, and validity issues in assessing K-12 English language learners in both U.S. and international contexts.

**Michael Suhan** is a research project manager at ETS. He holds an MA in Teaching English as a Second Language from Northern Arizona University. His work centers on automated scoring and feedback in language education and teacher education, simulated teaching experiences, and AI literacy in education.

**Kristopher Kyle** is an Associate Professor in the Linguistics Department at the University of Oregon. His research interests include second language acquisition, second language writing, corpus linguistics, computational linguistics, and second language assessment. His work involves the development of natural language processing tools used to test and build theories related to second language acquisition and second language writing development.