https://doi.org/10.17323/jle.2023.11045

Automated Measures of Lexical Sophistication: Predicting Proficiency in an Integrated Academic Writing Task

Randy Appel [1], **Angel Arias** [2]

[1] Waseda University, Tokyo, Japan
[2] Carleton University, Ottawa, Canada

ABSTRACT

Background. Advances in automated analyses of written discourse have made available a wide range of indices that can be used to better understand linguistic features present in language users' discourse and the relationships these metrics hold with human raters' assessments of writing.

Purpose. The present study extends previous research in this area by using the TAALES 2.2 software application to automatically extract 484 single and multi-word metrics of lexical sophistication to examine their relationship with differences in assessed L2 English writing proficiency.

Methods. Using a graded corpus of timed, integrated essays from a major academic English language test, correlations and multiple regressions were used to identify specific metrics that best predict L2 English writing proficiency scores.

Results. The most parsimonious regression model yielded four-predictor variables, with total word count, orthographic neighborhood frequency, lexical decision time, and word naming response time accounting for 36% of total explained variance.

Conclusion. Results emphasize the importance of writing fluency (by way of total word count) in assessments of this kind. Thus, learners looking to improve writing proficiency may find benefit from writing activities aimed at increasing speed of production. Furthermore, despite a substantial amount of variance explained by the final regression model, findings suggest the need for a wider range of metrics that tap into additional aspects of writing proficiency.

KEYWORDS
second language writing, automated analysis, corpus, integrated writing, L2 English

Citation: Appel R., & Arias A. (2023). Automated measures of lexical sophistication: Predicting proficiency in an integrated academic writing task. *Journal of Language and Education, 9*(1), 12-25. https://doi.org/10.17323/jle.2023.11045

Correspondence:
Randy Appel,
r_appel@waseda.jp

Received: July 01, 2020
Accepted: September 20, 2022
Published: March 31, 2023

INTRODUCTION

Over the past three decades, advances in computer aided corpus linguistics have enabled new approaches to discourse analysis that were previously either too time consuming or unreliable to be considered methodologically valid (e.g., Crossley, 2020). Using these relatively new tools, researchers can now apply standardized, and therefore replicable, assessments of a wide range of linguistic metrics to assess various important constructs (e.g., writing proficiency), thus providing important insights into the language produced by first (L1) and second (L2) language users (e.g., Casal & Lee, 2019; Kyle & Crossley, 2017). For instance, Lu (2011) implemented the Lexical Syntactic Complexity Analyzer to examine lexical characteristics of Chinese learners and found several measures of clause complexity and length to be strongly associated with writing ability.

The advent and development of automated text analysis have also inspired research that has examined single and multiword measures in both independent and source-based prompts. In this regard, multiword measures such as n-grams have shown to be important predictors of essay quality in independent tasks but less so in source-based tasks (Kyle & Crossley, 2016). Other text analysis tools such as Coh-metrix (Graesser et al., 2004) and TAALES (Kyle & Crossley, 2015) allow researchers to extract a wide range of linguistic measures that aid in the exploration of texts in unprecedented ways. This new world of computer aided linguistic analysis has led to increased interest in automated forms of assessment that could be used to decrease the burden placed on teachers and increase student independence by creating alternative paths for feedback and evaluation. However, to take advantage of the advances being made and get a more complete picture of how well these metrics can be used to model raters' assessment of L2 discourse, it is important to continue pushing the field of study forward by examining an ever-wider range of writing tasks and linguistic metrics.

In particular, as it relates to source-based, L2 English academic writing (i.e., academic tasks that require writers to integrate material from previously exposed texts and/or listening materials), a very narrow range of task types has thus far been examined, with a frequent focus on summary tasks from the Test of English as a Foreign Language (e.g., Guo, Crossley, & McNamara, 2013; Kim & Crossley, 2018; Kyle, 2017; Kyle & Crossley, 2016; Plakans *et al*., 2019). The TOEFL iBT is an important resource, yet the availability of this data set seems to have created an overreliance on this task and a related lack of research targeting alternative source-based writing. Thus, we lack a full understanding of the effectiveness of automatically extracted linguistic measures in highlighting the features most relevant to raters' assessments of L2 writing proficiency in source-based writing beyond this task.

The present study aimed to extend previous research into the automated analysis of source-based L2 English written discourse by using an unpublished corpus of timed, argumentative, L2 English academic essays collected from a former test of academic English proficiency commonly used to evaluate international students aiming to study at English-medium post-secondary institutions in Canada (Appel & Wood, 2016)[3]. To account for factors that may influence the language being produced, this study controls for essay type, topic, and testing conditions by using a single version (i.e., one writing prompt) of this standardized test. To better understand the underlying discoursal features relevant to human raters' assessments of academic English writing proficiency in this test, the TAALES 2.2 software suit was used to extract 484 single and multi-word candidate measures of lexical sophistication and subsequently examine their potential relationship with holistic scores of this source-based, L2 English writing proficiency test[4].

Empirical Analyses of L2 English Writing Proficiency

The link between linguistic features and writing proficiency dates back to at least the 1970s (Crossley, 2020), with empirical analyses largely taking hold in the 1990s (Wolfe-Quintero *et al*., 1998). The main goal in many of these studies has been to reveal the features most relevant to differences in assessed proficiency, and therefore which elements should receive greater focus during target language instruction. While early examinations of the lexical and grammatical features in L2 discourse indicated specific linguistic elements relevant to differences in assessed proficiency, these studies were often limited to the examination of a single or limited number of metrics due to the time consuming and laborious nature of the manual analysis. Therefore, knowledge of how various lexical factors work in combination were difficult to come by, and reliance on human coding created the potential problem of a lack of consistency in evaluations. However, due to recent advances in computer technology and associated forms of automated analyses, these problems can largely be avoided. In fact, we are now able to consistently and accurately extract a wide range of lexical measures from submitted texts without the need for human intervention.

Automated Analyses of L2 Writing

Automated computer assisted analyses have enhanced our collective understanding regarding the lexical, phrasal, cohesive, and syntactic features most relevant to evaluations of differences in linguistic ability (e.g., Bi & Jiang, 2020; Lei *et al*., 2023; Li *et al*., 2023). For example, Lu (2011) used the automated extraction of 14 syntactic complexity measures to analyze 3,554 independent, argumentative essays from the Written English Corpus of Chinese Learners (Wen, Wang, & Liang, 2005). Findings indicated that a combination of 7 discoursal features (*complex nominals per clause*, *complex nominals per T-unit*, *coordinate phrases per T-unit*, *coordinate phrases per clause*, *mean length of clause*, *mean length of sentence*, *mean length of T-unit*) were the best predictors of differences in perceived writing ability.

While the increased speed and accuracy of analysis provided by automated extraction has led to a greater dependence on these tools, their increasing acceptance and application in recent years is likely, at least somewhat, related to their free (e.g., Compleat Lexical Tutor, Coh-metrix, AntConc), or low cost (e.g., WordSmith Tools, LIWC) nature. For example, the 14 lexical measures used by Lu (2011) are now freely available through an online tool, the Web-based L2 Syntactical Complexity Analyzer. Thus, additional researchers are

[3] This study was based on a retired version of the Canadian Academic English Language (CAEL) assessment. The current version of this test maintains the core construct of academic language proficiency yet uses a different structure from the version in the current research.

[4] This study was supported by Paragon Testing Enterprises (now a subsidiary of Prometric). The analyses and conclusions presented in this paper are those of the authors and do not reflect the views or positions of Prometric.

able to make use of these measures to verify the findings of published studies using alternative corpora.

Coh-metrix, another popular and freely available web-based text analysis tool, allows users to submit discourse samples for automatic extraction of over 100 lexical metrics spanning 11 broad categories (*descriptive statistics*, *text easability*, *referential cohesion*, *latent semantic analysis*, *lexical diversity*, *connectives*, *situation model*, *syntactic complexity*, *syntactic pattern density*, *word information*, *readability*). Similarly, Compleat Lexical Tutor offers users a wide range of freely available linguistic tools that can be used to calculate, among others, metrics related to word frequency in reference to a range of previously compiled corpora. While each of these tools offer easily accessible methods to evaluate the language appearing in the corpora under investigation, the range of measures available has remained relatively limited. However, with the release of the Tool for the Automatic Analysis of Lexical Sophistication (TAALES), this is no longer the case.

TAALES

Originally released in 2015, TAALES (Kyle & Crossley, 2015) is a downloadable software application that allows users to batch analyze large collections of text for a wide range of lexical features. As a downloadable application, TAALES holds a distinct advantage over alternative tools (e.g., Coh-metrix, Web-based Syntactic Complexity Analyzer, Compleat Lexical Tutor) in that it can be used offline and does not require each text to be submitted individually, thus allowing for a faster and more efficient form of analysis. Although initially limited to 130 indices, with the release of version 2.2, TAALES now supports the extraction of 484 single and multi-word metrics, thereby making it one of the most comprehensive freely accessible linguistic analysis tools available. With a focus on lexical sophistication, these measures cover 10 broad categories (discussed in greater detail below and fully described at https://www.linguisticanalysistools.org).

Automated assessments of Source-based Writing

With regard to the writing tasks examined in previous automated analyses of L2 English proficiency, an important distinction must be made between independent and source-based (i.e., integrated) tasks. The former is a task type that relies purely on the writers' personal experience to contribute content to the composition. For example, personal narratives (e.g., *Describe a time you had to deal with conflict and how you resolved it*) as well as opinion pieces (e.g., *Do you think it is necessary for students to work part-time while attending university?*) would both fall under this category when composed without reference to outside sources.

These independent tasks are easy to administer since previous exposure to reading and listening materials is not required and the prompt can be provided to writers without any time-consuming preparation. Perhaps because of the relative ease associated with this type of data collection, independent writing has served as the primary focus in the majority of previous studies aiming to automatically extract relevant linguistic features that could be used to model human raters' assessment of L2 English discourse (Guo *et al*., 2013). Findings based on these analyses of independent writing include the identification of text length (e.g., Ferris, 1994) lexical diversity (Crossley & McNamara, 2012), and average length per word (e.g., Grant & Ginther, 2000) as important factors in assessments of L2 English proficiency.

Despite the wide-spread use of independent tasks in automated analyses of L2 English, this type of writing is a poor representation of the real-world, post-secondary assignments students will be required to complete, since 'it is impossible to assign academic writing tasks that don't require preliminary readings' (Johns, 1993, p. 277). Therefore, to better understand how more academically minded tasks are evaluated by human raters, recent years have seen an increasing focus on source-based writing. For instance, Guo et al. (2013) used Coh-metrix to analyze summary tasks from the TOEFL iBT public use data set in relation to holistically assigned proficiency scores. Multiple regressions were used to identify seven indices (*number of words per text, past participle verbs, word familiarity, verbs in 3rd person singular present, semantic similarity, verbs in base form, and word frequency*) that accounted for over 50% of the variance in assessed proficiency on a 5-point scale.

Several studies have also examined the influence of automated measures of text cohesion on writing development. Although these studies initially focused on L1 discourse, an increasing amount of research has adopted an L2 focus (e.g., Casal & Lee, 2019; Guo et al., 2013; Kim & Crossley, 2018). In terms of automated measures of discourse cohesion in relation to L2 English writing quality, Kim and Crossley (2018) found that lexical overlap between paragraphs, a global cohesion metric, was a significant predictor of writing quality for both independent and integrated TOEFL iBT tasks. Also making use of the TOEFL iBT, Guo et al. (2013) revealed that variance in writing quality on integrated tasks could be successfully predicted by local measures of discourse cohesion.

Similar studies to the above have also been conducted by Kyle et al. (2016), Plakans et al. (2019), and Kyle (2017) with a range of lexical factors identified and variance in proficiency scores accounted for. However, one potentially limiting factor in this body of research is that the majority of data comes from a single source: the TOEFL iBT. This source-based task is a quick, written exercise that involves a short lecture and text on the same topic, followed by a 20-minute allotment in which the test taker is asked to produce a summary of the previously exposed to materials. Therefore, although technically source-based, this task type removes any need to present a position and argue it in an effective manner using support from outside sources. Thus, this research approach has continued to neglect source-based argumentative writing, a task type that has been repeatedly identified as a key component of the undergraduate writing genre (e.g., Wingate, 2012).

Research Question

The present study aimed to extend previous research targeting automated measures of writing development in L2 English discourse. Of the three broad constructs commonly featured in research of this kind (i.e., lexis, syntax, and cohesion), we have chosen lexical sophistication as our focus, as this has been argued as presenting the richest measures of writing development (Crossley, 2020). With an emphasis on a single construct of writing quality (i.e., lexical sophistication), we also aim to use a wider range of measures than those commonly applied in this area. Furthermore, while our objective is to expand the number of metrics used to assess lexical sophistication in order to better assess their value in predicting writing quality, we also narrow our focus by targeting a tightly controlled corpus of L2 English writing that shares a common writing prompt, topic, and task, with all samples produced under similar writing conditions.

The composition of this corpus is also important since it serves as an alternative source of integrated, timed, L2 English academic writing designed to better represent the type of task students will find necessary when studying in English medium post-secondary institutions. As a result, a main goal of the present research is to better identify which specific metrics of lexical sophistication are most associated with differences in assessed writing proficiency in academic tasks of this nature. Put differently, the purpose of this study was to explore potential answers to the following research question: Which automatically extracted single and multi-word metrics of lexical sophistication can be used to predict holistic ratings of L2 English writing proficiency on a timed, source-based writing task?

METHOD

Corpus

The corpus analyzed in this study is composed of timed, written responses (*n* = 589) to a single argumentative essay prompt included as part of a four section (reading response, lecture response, oral language response, written response) timed, high-stakes language test. The final portion of this test, the written response, provides the data used in this study. This test aims to assess L2 English learners' ability to use academic English as needed in order to fully participate in English-medium post-secondary institutions. As the final component in a multi-section linguistic evaluation on a single topic, writers are encouraged to make use of previous components of the test (e.g., lecture and reading texts) to help strengthen the position presented in their essays.

This test used a nine-band grading system, with 10 denoting the lowest level of writing proficiency and 90 indicating the highest assessed level (Appendix A). To ensure grade consistency, each essay was assessed by a team of three trained raters using a collaborative, read-aloud protocol. If all three raters agree on the grade the essay should receive, the score is recorded, and the process begins again with the following essay. In cases of disagreement among any of the three raters, discussions are initiated until a consensus is reached. The collaborative read-aloud protocol used to evaluate these essays is viewed as decreasing the focus on surface features, thereby encouraging raters to be more attentive to more substantive aspects of each piece of writing, such as organization, cohesion, and coherence of the presented argument. Thus, differences in essay length may be an important factor in ratings for this timed task, as greater essay length may be seen as providing evaluators with more opportunities to assess the macro features of each text. This emphasis on macro features contrasts with an overt focus on grammatical inaccuracies that may not interfere with the overall structure, organization, and message of each piece of writing[5].

Measures of Lexical Sophistication

In total, 484 indices of lexical sophistication were selected as candidate measures and included in the present study. These measures spanned 10 broad categories, which are summarized below and were selected since "lexical sophistication tends to provide the richest metrics of text quality" (Crossley, 2020, p. 417). For additional details on each category and the individual indices contained within them, see Kyle and Crossley (2015).

Word Frequency

Word frequency scores in TAALES are sourced from metrics related to several major corpora, including the Thorndike-Lorge magazine corpus (Thorndike & Lorge, 1944), Kucera-Francis (Kucera & Francis, 1967), Brown Corpus (Kucera & Francis, 1967), British National Corpus (BNC[6], 2007), Corpus of Contemporary American English (COCA[7], 2008), and SUBTLEXus (Brysbaert & New, 2009). Vocabulary in the submitted texts that does not appear in any of the corresponding corpora is excluded from the final frequency

[5] All essay ratings were provided by Paragon. However, the current version of the CAEL assessment uses a rating method that differs from the one described in this manuscript.

[6] British National Corpus. (2007). British National Corpus, version 3 (BNC XML ed.).

[7] Davies, M. (2008). The Corpus of Contemporary American English: 520 million words, 1990–present. Available online at https://corpus.byu.edu/coca/

counts. Indices from this category include raw frequency, mean frequency, and logarithmic frequency, with options allowing for individual reports or measures related to content words, function words, or all words (i.e., combination of content and function words).

Word Range

Word range information for several of the aforementioned corpora (e.g., Brown, BNC, SUBTLEXus, COCA) is also provided. As with word frequency, these metrics come in raw and logarithmic forms, with the option to restrict scores to content words, function words, or all words.

Contextual Distinctiveness

These values are based on the results from previous research, such as the Edinburgh Associative Thesaurus (Kiss et al., 1973), McDonald's co-occurrence probability (McDonald & Shillcock, 2001), and Latent Semantic Analysis values (Landauer *et al*., 2007).

Psycholinguistic Word Information

Largely based on information from the MRC psycholinguistic database (Coltheart, 1981), a collection of human ratings on more than 26 psychological properties from over 150,000 words, TAALES includes measures of familiarity, imageability, concreteness, and meaningfulness.

Word Neighborhood

Using scores from the English Lexicon Project (Balota et al., 2007), word neighborhood scores are reported for orthographic, phonographic, and phonological neighbors. For example, orthographic neighbors are calculated as the number of words that can be formed by changing only one letter in the target vocabulary item (e.g., *past* and *last*).

Word Recognition

Metrics from this category are calculated based on the norms reported in the English Lexicon Project (Balota et al., 2007). In general, longer word recognition times can be viewed as an indication of increased processing difficulty.

Age of Exposure

Indices from this category are based on data from the TASA corpus (Dascalu et al., 2016). Low age of exposure scores reflect texts with a higher proportion of words from lower grade levels, while texts earning higher scores are indicative of texts with more words from higher grades.

Semantic Relationships

Values in this category include hypernymy and polysemy norms from WordNet (Fellbaum, 1998), which contain information on connections between nouns, verbs, adjectives, and adverbs. Texts with high hypernymy tend to be more specific, have more concrete terms, and have fewer abstract words. High polysemy indicates texts with more sense relationships.

N-Gram Indices

Indices from this category are largely rooted in the bigram and trigram frequencies for the written and spoken sections of the BNC corpus created by Crossley et al. (2012). The proportion of n-grams is a measure of the percentage of unique n-grams present in the text that are also found in the reference corpora. TAALES 2.2 also includes five measures of association strength derived from the COCA: Mutual Information (MI), MI^2, t-score, ΔP, and approximate collexeme score.

Academic Language

Measures from this category aim to identify the prevalence of academic lexis in each submitted text using the Academic Word List (AWL; Coxhead, 2000) and the Academic Formulas List (Simpson-Vlach & Ellis, 2010) as reference points.

Another category included in this study was text length because "it can be considered especially construct-relevant when it comes to writing in a foreign language" (Fleckenstein et al., 2020, p. 2). In argumentative essays, particularly the timed ones analyzed in this study, using a wider range of linguistic resources can result in longer texts, which in turn can be arguably considered as indicative of higher proficiency.

Analysis

The identification of lexical factors associated with holistic judgments of L2 English writing proficiency on this timed, integrated task, followed a three-step process. First, Pearson correlation coefficients were used to explore the associations between the 484 candidate variables and the target construct of assessed writing proficiency. Second, those lexical metrics with the strongest correlations were assessed for redundancy (i.e., highly correlated predictor variables), with the goal of eliminating redundant lexical variables. Third, after examining redundancy, the remaining lexical measures were included as predictor variables in a stepwise multiple regression to highlight which of the lexical features identified in step two accounted for significant proportions of variance in human raters' assessments of holistic writing ability[8].

[8] Wherever possible, the decision to retain a variable was based on the strength of correlation. In other words, in each pair of collinear variables, the variable holding the strongest relation with assessed proficiency score was retained.

Pearson Correlations

As a first step in assessing the predictive strength of the 484 TAALES measures in accounting for differences in L2 English writing proficiency, using SPSS version 25 (IBM SPSS Statistics, 1989 – 2017), initial Pearson correlations were run between the full range of measures and the score awarded to each essay. Before running any correlations, all data were analyzed to ensure relatively normal distributions, a lack of outliers, and the absence of missing data. Adhering to Plonsky and Oswald's (2014) localized recommendations for meaningful correlations in L2 research (*r* ≥ .250) resulted in a list of variables (*k* = 36) associated with assessed L2 English writing proficiency. This initial group of 36 variables included various metric options within the same category (e.g., Brown Frequency All Words [AW] and Brown Frequency Content Word [CW]). Thus, to avoid statistical dependencies between options "all words", "content words" and "function words" within the same category, only the option sharing the strongest correlation with band score was considered for further analysis. Following previous research and recommendations (e.g., Appel *et al*., 2019; Hinkle *et al*., 2003), to account for the existence of multicollinearity among these significantly correlated predictor variables, in any case where two predictor variables held a minimum correlation of .700 with each other, only one was retained. This helped eliminate the possibility of including multiple overlapping measures in the final step of the analysis (i.e., multiple regression). Following the identification of significantly correlated predictor variables and the elimination of multicollinearity, the final step was to build a multiple regression model to better understand the value of the correlated variables in accounting for differences in L2 English writing proficiency.

Multiple Linear Regression

Given the primary focus of this study was to predict differences in assessed L2 English academic writing proficiency based on single and multi-word metrics, a stepwise multiple regression model was subsequently fit to the data. This type of regression is appropriate for prediction (Keith, 2019) and predictor variables are handled automatically by the software package based on statistical criteria and techniques (Plonsky & Ghanbar, 2018). For this study, the technique of forward selection was used, consisting of first selecting the predictor variable that shared the strongest correlation with the criterion variable and then selecting the next variable with the highest contribution to the model on an iterative basis.

As in other types of multiple linear regression analyses, statistical assumptions were examined and met prior to the application of this statistical procedure to make valid interpretations and inferences from the data. In this regard, the data were screened for influential data points, and after fitting the regression model, the residuals were used to verify their normality and homoscedasticity. The model's residuals were normally distributed, but the homogeneity of variance was slightly violated. However, the variance of Y for each value of X was generally constant and "conditional on the level of each of the predictor variables and on the overall $\hat{Y}$ from the final regression equation" (Howell, 2013, p. 536). Furthermore, the variance inflation factor (≤ 2, see Table 3) indicated no collinearity issues. Additionally, Cook's distances (≤ 0.040) did not raise any concerns regarding outlying measures, thus preserving the model's integrity.

RESULTS

Correlations between TAALES Variables and Assessed Writing Proficiency

As a first step in our results, we provide Pearson correlation coefficients between holistically assessed writing score and the candidate predictive measures of writing proficiency (Table 1). These initial correlations were used to assess the value of each measure as a predictor variable and the amount of collinearity between variables that could be identified. After accounting for multicollinearity (i.e., variables that were too closely related), six predictor variables remained (Table 2). These six predictor variables were subsequently included in the regression model. These variables cover four of the previously mentioned categories (word recognition, word neighborhood information, word concreteness, word frequency), as well as one additional category (sample length).

As can be seen in Table 2, word count, word naming response time, and lexical decision time held positive relationships with assessed writing proficiency. Conversely, orthographic neighborhood frequency, unigram familiarity, and HAL frequency of closest phonological neighbors maintained negative correlations. The strongest correlation with assessed proficiency was for word count with a correlation of .516.

Lexical Predictors of Writing Proficiency

The regression model yielded four predictor variables free of collinear issues (see variance inflation factor): *total word count*, *orthographic neighborhood frequency*, *lexical decision time*, and *word naming response time*, suggesting that the four-predictor model was statistically significant, *F*(4, 584) = 84.668, *p* < .001 and accounted for approximately 36% of the variance in writing proficiency scores (R^2 = .367, adjusted R^2 36.3%) for the integrated writing task. *Total word count* explained the largest portion of the variance associated with writing proficiency, thus exhibiting the strongest weight in the model (β = .452, *t* = 13.428, *p* < .001), followed by *word naming response time* (β = .151, *t* = 3.352, *p* = .001), *orthographic neighborhood frequency* (β = -.125, *t* = -3.22, *p* = .001) and *lexical decision time* (β = .120, *t* = 2.908, *p* = .004). Note, however, that the predictor *orthographic neighborhood frequency* was negatively associated with writing proficiency scores.

Table 1

Pearson Correlations between Writing Score and Potential Predictive Lexical Measures

Variables	Score	WC	WNRTz CW	WNRT CW	WNRT	WNRTz	LDTsd CW	LDT	LDT CW
WC	0.52								
WNRTz CW	0.37	0.16							
WNRT CW	0.35	0.14	0.98						
WNRT	0.34	0.14	0.94	0.95					
WNRTz	0.33	0.15	0.95	0.94	0.98				
LDTsd CW	0.34	0.19	0.61	0.60	0.58	0.58			
LDT	0.34	0.13	0.84	0.85	0.87	0.85	0.68		
LDT CW	0.34	0.14	0.88	0.88	0.84	0.84	0.71	0.95	
ALD CW	0.32	0.14	0.85	0.85	0.81	0.81	0.57	0.82	0.86
LDTz	0.33	0.11	0.82	0.83	0.86	0.84	0.61	0.98	0.92
MRC AW	-0.34	-0.18	-0.65	-0.64	-0.67	-0.65	-0.47	-0.64	-0.61
HAL CW	-0.34	-0.15	-0.83	-0.82	-0.80	-0.81	-0.57	-0.78	-0.81
HAL	-0.32	-0.14	-0.64	-0.63	-0.70	-0.69	-0.48	-0.66	-0.62
ONF CW	-0.32	-0.16	-0.54	-0.52	-0.53	-0.54	-0.34	-0.52	-0.53

Variables	ALD CW	LDTz	MRC AW	HAL CW	HAL
WC					
WNRTz CW					
WNRT CW					
WNRT					
WNRTz					
LDTsd CW					
LDT					
LDT CW					
ALD CW					
LDTz	0.81				
MRC AW	-0.58	-0.62			
HAL CW	-0.90	-0.76	0.62		
HAL	-0.70	-0.64	0.64	0.79	
ONF CW	-0.48	-0.51	0.40	0.62	0.55

Note: *WC = Word Count; WNRTz CW = Word Name Response Time (z-score) CW; WNRT CW = Word Naming Response Time CW; WNRT = Word Naming Response Time; WNRTz = Word Naming Response Time (z-score); LDTsd CW = Lexical Decision Time (standard deviation) CW; LDT = Lexical Decision Time; LDT CW = Lexical Decision Time CW; ALD CW = Average Levenshtein Distance of closest phonological neighbors CW; LDTz = Lexical Decision Time (z-score); MRC AW = MRC Familiarity CW; HAL CW = Average Log HAL frequency of closest orthographic neighbors CW; HAL = Average Log HAL frequency of closest orthographic neighbors; ONF CW = Orthographic Neighborhood Frequency CW; AW = All words; CW = Content Words.*

DISCUSSION

The current study explored the relationship between a wide range of automatically extracted measures of lexical sophistication and assessments of L2 English proficiency on a timed, integrated, academic writing task. While the initial correlations used to identify metrics associated with differences in perceived L2 English writing proficiency yielded a somewhat unwieldy list of variables, eliminating weak correlations and multi-collinearity helped reduce the number of candidate variables to a more manageable level. The final multiple regression yielded a four-factor model accounting for 36.3% of variance (adjusted R^2) in assessed proficiency. Each of these four metrics is discussed in detail below.

Lexical Correlates Accounting for Significant Variance in Assessed Proficiency Scores

Total Word Count

The importance of *total word count*, which held the strongest correlation with assessed proficiency (r = 0.516) and accounted for the largest portion of variance (26.5%) in test scores, supports findings from previous studies (e.g., Kamimura & Oi, 2001; Sasaki, 2000) regarding developmental trends in L2 English writing proficiency using independent writing tasks. Although each of these studies has suggested a link between length of sample and assessed proficiency in independent tasks, results from the current study extend these findings to integrated, argumentative academic writing tasks produced under testing conditions. Although text length was the main predictor of proficiency, further research should include other text measures to continue exploring the construct complexities of second language writing. More recently, Crossley (2020) decided not to include text length in a systematic review of linguistic features in writing quality, but acknowledged that text length is likely the strongest predictor of writing development quality. Similarly, other studies have found a positive relationship between text length and human ratings, while controlling for language proficiency (McNamara et al., 2015).

The relationship between volume of output and proficiency in this study likely relates to the importance of writing fluency (i.e., speed of production) in assessments of this kind (e.g., timed tasks produced under testing conditions). In timed tasks, lower-level writers may lack sufficient opportunities to plan and compose their essays, thereby leading to reduced output volume. Thus, the ability to quickly produce discourse within a limited time may be interpreted as an important factor that significantly contributes to rater assessments of proficiency in L2 English writing.

The link between longer essays and perceptions of higher proficiency may also relate to the fact that longer texts give assessors more material on which to base their evaluation by increasing writers' opportunities to display more diverse

Table 2

Statistically Significant Correlations after Accounting for Multicollinearity

Category	Variable	Correlation
Sample Length	Word Count	0.516
Word Recognition	Word Naming Response Time – Content Words (z-score)	0.368
	Lexical Decision Time – Content Words (standard deviation)	0.338
Word Neighbor Information	Orthographic Neighborhood Frequency – Content Words	-0.316
Word Concreteness	Unigram Familiarity (mean)	-0.341
Word Frequency	HAL Frequency of Closest phonological neighbors	-0.315

Table 3

Results of the Multiple Regression

Variable	R^2	$R^2(adj)$	β	t	p	VIF
Word Count	0.267	0.265	0.452	13.428	0.000	1.047
Word Naming Response Time (z-score)	0.346	0.344	0.151	3.352	0.001	1.884
Orthographic Neighborhood Frequency (content words)	0.358	0.355	-0.125	-3.22	0.001	1.383
Lexical Decision (content words [standard deviation])	0.367	0.363	0.120	2.908	0.004	1.582

and linguistically advanced structures. In contrast, short samples may not enable writers to produce as wide an array of linguistic elements. For example, low-level writers (band scores of 10-20) are described as using 'restricted language' (Appendix A); without a sufficient volume of writing, it may be challenging for writers at this level to demonstrate the ability to move beyond this limited range of lexical items.

In longer texts, however, raters are more likely to find language instances that match the associated descriptors, regardless of occasional slips or mistakes that may be present. In fact, essays receiving a score of 30 are listed as 'unable to develop ideas' while adept writers (scoring at band 70) are described as using 'the readings and lecture effectively to support the thesis'. Thus, while it is clear that volume of output has a clear relationship with perceptions of L2 ability on this task, this finding may be more broadly linked to the ability of writers to express their ideas using varied lexis and present their position using adequate support from outside sources.

Word Naming Response Time (Z-Score)

Word naming response time contributed approximately 8% additional variance accounted for in the final regression model. This variable is a psycholinguistic measure of the accuracy and time native English speakers require to read aloud words in a given text. To calculate figures for this variable, L1 English speakers were presented with individual word or non-word samples and asked to read each aloud, with their responses being recorded and aggregated (Balota et al., 2007).

Potential reasons for the importance of this measure in assessments of perceived proficiency include the fact that words with longer naming response times and lower accuracy on word naming correctness decisions are generally seen as more sophisticated and thus indicative of more advanced levels of linguistic ability. For instance, both Kim *et al*. (2018) and Kyle *et al*. (2017) found a positive association between word naming response times and holistically scored independent writing tasks. Furthermore, recent research on L2 English speech by Berger et al. (2017) found that more frequent use of words with longer response times was associated with higher perceived proficiency among L2 English users. Findings from the current study echo these findings, thus bringing further support to the notion that the more frequent use of less accurately named words with longer word naming response times is a sign of higher L2 English proficiency in both oral and written communication.

Orthographic Neighborhood Frequency - Content Words

As previously mentioned, this variable provides a summary score for the number of orthographic neighbors for individual words in each sample. When specifically related to content words, as applied here, this metric is likely a truer measure of the number of orthographic neighbors present in each text since function words (e.g., auxiliaries, prepositions, quantifiers, pronouns) are excluded. Importantly, as evidenced by the correlations and multiple regression, this metric holds a negative relationship with assessed proficiency.

This negative association can be explained by the fact that words with fewer orthographic neighbors are generally seen as an indication of more advanced vocabulary. Thus, texts with fewer orthographic neighbors are generally viewed as more linguistically advanced than those texts with a higher number of orthographic neighbors. The relationship between orthographic neighborhood frequency and perceptions of linguistic ability is supported by results from Kim et al. (2018) who also found a negative relationship between the number of orthographic neighbors and perceptions of L2 English proficiency on independent writing tasks. Therefore, as with previous findings in this study, these results allow us to extend the results of previous research targeting independent writing tasks to academic, integrated writing produced under testing conditions.

Lexical decision time – Content Words (Standard Deviation)

Similar to word naming response time, lexical decision time suggests the importance of moving beyond a reliance on the most easily recognized, and often earliest learned vocabulary, in order to incorporate higher level lexis that may be less quickly recognized by native speakers – thereby demonstrating greater vocabulary knowledge (e.g., Appel et al., 2019). With research suggesting that orthographic neighborhood frequency can help account for variance in word naming and lexical decision tasks (e.g., Adelman & Brown, 2007), it seems likely that the combination of *word naming response time*, *orthographic neighborhood frequency*, and *lexical decision time* all point to the same underlying characteristic: increased use of advanced vocabulary that has generally been acquired at a more advanced stage of linguistic development.

Implications

Methodological Implications

Findings suggest that it is necessary to continue expanding the range of variables used in studies aiming to identify underlying features associated with differences in L2 English academic writing ability, particularly in relation to lexical sophistication in integrated tasks. Although a comparatively large range of variables (484) were targeted here, a rather limited number remained after accounting for multicollinearity and weak correlations with the target construct.

Thus, it would seem as though, despite efforts to increase the range of metrics that can be applied in research of this kind, substantial overlap is limiting the usefulness of this widened scope. As a result, it may be necessary to more critically assess the range of metrics being made available to researchers through automated tools in order to better understand how distinct they truly are and what benefit is gained through each new offering.

A critical evaluation of these metrics may also lead to the development of new measures that help highlight alternative aspects of lexical sophistication that are more distinct from current offerings, thereby increasing explanatory power. In close relation to this, although we deliberately focused on one specific construct of writing quality in this study, lexical sophistication, additional constructs should continue to be targeted, with each metric evaluated under the same critical gaze. As 36% of total variation in essay scores was predicted by the metrics used here, further variance remains to be accounted for and additional variables, both established metrics related to other writing constructs (e.g., syntax, cohesion) and new lexical sophistication measures, would prove beneficial in the ongoing goal of accurately modeling human assessments of L2 writing quality on integrated writing tasks.

In the same vein, a vast amount of currently used statistical models (e.g., multiple regression) in this line of research, assume a certain level of independence among variables and breaching these assumptions can lead to the spurious interpretation of the results. Thus, careful attention must be paid to the manipulation of variables, data screening, relevance of the model, and the assumptions underlying statistical models in order to develop and apply metrics that will be of actual value in research of this kind. In relation to this point, it is not only the volume of metrics that must be increased, but also the quality and uniqueness.

With the limited number of multi-word measures currently available (at least in comparison to single-word measures), this category appears to offer a fruitful area for expansion. In fact, with regard to multi-word measures, Paquot (2019) has recently demonstrated that new phraseological metrics can indeed be used to distinguish between levels of the Common European Framework of References for Languages (CEFR). Thus, although multi-word metrics did not factor into the final regression in the present study, this category of variables still holds potential and should not be abandoned.

Pedagogical Implications

The most important finding from a pedagogical perspective is the high amount of variance in writing proficiency that could be explained by total word count. This finding, though not entirely surprising given results from previous studies, is important in that it suggests it is not only text quality that contributes to assessments of L2 writing ability, but also quantity. In light of the importance of total word count in assessed proficiency on this task, and other similarly timed tasks from previous research, it may be beneficial for teachers to incorporate more fluency-based activities to boost speed of written production. For instance, timed activities in which students are encouraged to write as much as possible on a given topic within a short time frame (e.g., 10-15 minutes) could be used as a way of promoting greater writing fluency and prepare students for similar testing conditions.

Although grammatical accuracy and appropriate academic vocabulary should not be ignored, it may be helpful to encourage students to focus more on quantity, as opposed to a strict adherence to accuracy, at least in the early stage of the writing process, in order to increase comfort producing longer texts that offer more linguistic data. This may also help students make use of a wider range of sophisticated lexis, as a fear of making mistakes may be reduced due to the shift in focus to text quantity. As students should already be encouraged to take a process view of writing that involves revision of each draft, an early focus on ‘getting their ideas out’ in the initial stages of the writing process could benefit later drafts as students will have a greater volume of text to revise (Elbow, 1973). Thus, initial drafts should be viewed as a first step that allows students to express their ideas fully without an overly oppressive fear of making mistakes – a factor that may limit production in favor of an emphasis on correctness.

While a focus on writing fluency by way of timed activities is likely to offer some benefit to learners, this recommendation is hedged by the fact that the results from this study, as well as those of many others in this area, are based on exam style writing produced under testing conditions. In this type of task, time constraints are likely a major factor that feeds into the subsequent assessment. In other words, the existence of time constraints may encourage raters to place greater emphasis on text length in the evaluation of each sample since lower-level writers may struggle to spontaneously produce large amounts of text without sufficient planning/drafting time.

However, these time pressures may be less important in assignments included as part of normal course activities (e.g., take-home assignments). This is particularly true in those assignments that make use of minimum/maximum word counts, since all students are expected to produce a similar volume of text. Therefore, while we believe a focus on improving students’ writing fluency is a worthwhile goal that could lead to proficiency gains, it is important to keep in mind that improvements in writing fluency may not equally benefit all types of writing tasks, and fluency is only one aspect of a multi-faceted issue. Thus, time restricted activities may lead to proficiency gains on similar writing tasks; however, these activities may be less effective in achieving comparable gains on compositions that make use of more liberal time constraints (e.g., take-home assignments).

Despite the importance of writing fluency in assessments of this kind, the combination of word naming response time, orthographic neighborhood frequency, and lexical decision time (combined 10% of variance accounted for) reemphasize the importance of not simply producing a large volume of writing but incorporating genre and register appropriate lexis that demonstrates mastery of advanced vocabulary. Thus, although a major pedagogical implication of this study is that it is necessary to incorporate speed of production activities into the language classroom in order to prepare students to be able to write effectively under the time constraints common in testing conditions, appropriate register and genre specific language cannot be ignored. As a result, it would seem important for language instructors to gear their teaching approach to the goals of each specific group of students. For example, for test preparation, fluency activities may prove exceptionally beneficial. However, for students aiming to prepare for other settings (e.g., post-secondary studies), it may be more helpful to focus on activities that could lead to an increase in the level of lexical sophistication present in each student's text.

CONCLUSION

Two main conclusions can be drawn from the current study. First, although substantial advances have been made in the automated analysis of written discourse, further developments are needed to more deeply explore the existing relationship between lexical measures and human scores of L2 proficiency. As previous research also found a relatively low amount of variance accounted for by lexical measures in a similarly designed source-based writing task, there is a need for the development of further metrics to explore this issue. The methodological caveats we encountered can be a limiting factor to the development of automated analyses of written discourse in corpus research and this calls for the development of new measures or techniques that account for such limitations. Second, in light of the identified importance of text length in holistic assessments in the type of integrated writing task targeted here, it may be necessary for writing teachers to offer further opportunities for their L2 English learners to focus on writing fluency if they are to better prepare their students for the type of writing they will be asked to perform in assessments of this kind.

Automated text analysis in applied linguistics is blooming, but more work is needed to further understand the intricacies of linguistic measures in writing ability. New measures need to be developed to overcome the current issues associated with similar measures in different transformations (e.g., logs, standardized, raw), which is by and large the cause of multicollinearity in linear models. Thus, more advanced statistical models need to be explored to overcome this issue. Perhaps machine learning, and multilevel models are a good starting point.

FUNDING

This research was funded by a Waseda University Grant for Special Research Projects (Project # 2017S-180).

DECLARATION OF COMPETITING INTEREST

None declared.

AUTHORS' CONTRIBUTION

Randy Appel: Conceptualization; Funding acquisition; Investigation; Methodology; Project administration; Resources; Supervision; Validation; Writing-review and editing.

Angel Arias: Statistical analyses.

REFERENCES

Adelman, J. S., & Brown, G. D. A. (2007). Phonographic neighbors, not orthographic neighbors, determine word naming latencies. *Psychonomic Bulletin & Review, 14*, 455–459. https://doi.org/10.3758/BF03194088

Appel, R., Trofimovich, P., Saito, K., Isaacs, T., & Webb, S. (2019). Lexical aspects of comprehensibility and nativeness from the perspective of native-speaking English raters. *ITL-International Journal of Applied Linguistics*, *170*, 24-52. https://doi.org/10.1075/itl.17026.app

Appel, R., & Wood, D. (2016). Recurrent word combinations in EAP test-taker writing: Differences between high-and low-proficiency levels. *Language Assessment Quarterly*, *13*, 55-71. https://doi.org/10.1080/15434303.2015.1126718

Balota, D. A, Yap, M. J., Cortese, M. J., Hutchison, K. A, Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39*, 445–459. https://doi.org/10.3758/BF03193014

Berger, C. M., Crossley, S., & Kyle, K. (2017). Using native-speaker psycholinguistic norms to predict lexical proficiency and development in second language production. *Applied Linguistics, 40*(1), 22-42. https://doi.org/10.1093/applin/amx005

Bi, P., & Jiang, J. (2020). Syntactic complexity in assessing young adolescent EFL learners' writings: Syntactic elaboration and diversity. *System*, *91*(4), 1-10. https://doi.org/10.1016/j.system.2020.102248

Brysbaert, M., & New, B. (2009). Moving beyond Kucera 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*, 977–990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*, 904–911. https://doi.org/10.3758/s13428-013-0403-5

Casal, J.E., & Lee, J.J. (2019). Syntactic complexity and writing quality in assessed first-year L2 writing. *Journal of Second Language Writing*, *44*, 51-62. https://doi.org/10.1016/j.jslw.2019.03.005

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology Section A, 33*, 497–505. https://doi.org/10.1080/14640748108400805

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*, 213-238. https://doi.org/10.2307/3587951

Crossley, S. (2020). Lingusitic features in writing quality and development: An overview. *Journal of Writing Research*, *11*(3), 415-443. https://doi.org/10.17239/jowr-2020.11.03.01

Crossley, S., Cai, Z., & McNamara, D. (2012). Syntagmatic, paradigmatic, andautomatic n-gram approaches to assessing essay quality. In P. M. McCarthy & G.M. Youngblood (Eds.), *Proceedings of the 25th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 214–219). AAAI Press.

Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing, 26*, 66–79. https://doi.org/10.1016/j.jslw.2014.09.006

Crossley, S., Kyle, K., McNamara, D. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods, 48*, 1227-1237. https://doi.org/10.3758/s13428-015-0651-7

Crossley, S. A., Roscoe, R. D., McNamara, D. S., & Graesser, A. (2011) Predicting human scores of essay quality using computational indices of linguistic and textual features. In G. Biswas, S. Bull, J. Kay, and A. Mitrovic (Eds.), *Proceedings of the 15th International Conference on Artificial Intelligence in Education* (pp. 438-440). Springer. https://doi.org/10.1007/978-3-642-21869-9_62

Dascalu, M., McNamara, D. S., Crossley, S. A., & Trausan–Matu, S. (2016). Age of exposure: A model of word learning. In *The 30th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence* (pp. 2928–2934). AAAI Press.

Elbow, P. (1973). *Writing without teachers*. Oxford University Press.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. The MIT Press.

Fitzgerald, J., & Spiegel, D. L. (1986). Textual cohesion and coherence in children's writing.*Research in the Teaching of English*, *20*, 263-80.

Fleckenstein, J., Meyer, J., Jansen, T., Keller, S., & Köller, O. (2020). Is a long essay always a good essay? The effect of text length on writing assessment. *Frontiers in Psychology*, *11*, 562462. https://doi.org/10.3389/fpsyg.2020.562462

Guo, L., Crossley, S., & McNamara, D. (2013). Predicting human judgements of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing, 18*, 213-238. https://doi.org/10.1016/j.asw.2013.05.002

Hinkle, D., Wiersma, W., & Jurs, S. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Houghton Mifflin.

Howell, D. C. (2013). *Statistical methods for psychology* (8th ed.). Wadsworth.

Johns, A. M. (1993). Reading and writing tasks in English for academic purposes classes: Products, processes and resources. In J. G. Carson & I. Leki (Eds.), *Reading in the composition classroom: Second language perspectives* (pp. 274–289). Heinle & Heinle.

Kamimura, T., & Oi, K. (2001). The effects of differences in point of view on the story production of Japanese EFL students. *Foreign Language Annals, 34(2)*, 118 - 128. https://doi.org/10.1111/j.1944-9720.2001.tb02817.x

Keith, T. Z. (2019). *Multiple regression and beyond*: *An introduction to multiple regression and structural equation modeling* (3rd ed.). Routledge.

Kim, M., Crossley, S. (2018). Modelling second language writing quality: A structural equation investigation of lexical, syntactic, and cohesive features in source-based and independent writing. *Assessing Writing, 37*, 39-56. https://doi.org/10.1016/j.asw.2018.03.002

Kim, M., Crossley, S., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal, 102*, 120-141. https://doi.org/10.1111/modl.12447

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In A. J. Aitkin, R. W. Bailey, & N. Hamilton–Smith (Eds.), *The computer and literary studies* (pp. 153–165). Edinburgh University Press.

Kyle, K., & Crossley, S. (2015) – Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49*, 757-786. https://doi.org/10.1002/tesq.194

Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and sources-based writing. *Journal of Second Language Writing, 34*, 12-24. https://doi.org/10.1016/j.jslw.2016.10.003

Kyle, K., & Crossley, S. (2017). Assessing syntactic sophistication in L2 writing: a usage-based approach. *Language Testing, 34(4)*, 513-535. https://doi.org/10.1177/0265532217712554

Kyle, K., Crossley, S., & Berger, C. (2017). The tool for the automatic analysis of lexical sophistication. *Behavior Research Methods, 50*, 1040-1046. https://doi.org/10.3758/s13428-017-0924-4

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English. English*. Brown University Press.

Kuperman, V., Stadthagen–Gonzales, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30 thousand English words. *Behavior Research Methods, 44*, 978–990. https://doi.org/10.3758/s13428-012-0210-4

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). *Handbook of latent semantic analysis*. Lawrence Erlbaum.

Lei, L., Wen, J., & Yang, X. (2023). A large-scale longitudinal study of syntactic complexity development in EFL writing: A mixed effects model approach. *Journal of Second Language Writing*, *59*, 1-14. https://doi.org/10.1016/j.jslw.2022.100962

Li, Y., Lin, S., Liu, Y., & Lu, Z. (2023). The predictive powers of fine-grained syntactic complexity indices for letter writing proficiency and their relationship to pragmatic appropriateness. *Assessing Writing*, *56*, 1-15. https://doi.org/10.1016/j.asw.2023.100707

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly 45(1)*, 36–62. https://doi.org/10.5054/tq.2011.240859

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *Modern Language Journal, 96*, 190–208. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x

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*, 295–322. https://doi.org/10.1177/00238309010440030101

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing*, *23*, 35–59. https://doi.org/10.1016/j.asw.2014.09.002

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research, 35*, 121-145. https://doi.org/10.1177/0267658317694221

Plonsky, L., & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R^2 values. *The Modern Language Journal, 102*, 713-731. https://doi.org/10.1111/modl.12509

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878-912. https://doi.org/10.1111/lang.12079

Sasaki, M. (2000). Toward an empirical model of EFL writing processes: an exploratory study. *Journal of Second Language Writing, 9(3)*, 259-291. https://doi.org/10.1016/S1060-3743(00)00028-X

Simpson–Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31*, 487–512. https://doi.org/10.1093/applin/amp058

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. Bureau of Publications, Teachers College.

Wen, Q., Wang, L., & Liang, M. (2005). *Spoken and written English corpus of Chinese learners*. Foreign Language Teaching and Research Press.

Wolfe-Quintero, K., Inagaki, S., & Kim, H-Y (1998). *Second language development in writing: measures of fluency, accuracy, & complexity*. University of Hawaii Press.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press.

APPENDIX A

Writing Performance Band Score Criteria

Score	Criteria
10-20	***Very Limited:*** Generally unable to express ideas effectively Very restricted and/or ungrammatical language Uses words randomly and without overall coherence
30	***Limited:*** Attempts to write something which is related to the topic but the writing is not predictable Restricted and/or ungrammatical language Seems to understand the topic, but is unable to develop ideas because language constrains or distorts expression
40	***Marginally Competent:*** Makes links among ideas and addresses the topic but the writing lacks clarity and cohesiveness Displays elements of control in the writing (e.g. a thesis statement, an introduction and conclusion) but internal coherence is lacking Uses little or no support (i.e., quotations, examples, etc.) to develop the thesis
50	***Competent but Limited:*** Addresses the topic to a degree but with somewhat limited clarity and cohesiveness Uses some support to develop the thesis Control of the argument is limited by poor comprehension of the readings and lecture, and/or poor understanding of the requirements of academic writing
60	***Competent:*** Develops a thesis using a range of support Uses language that is generally accurate but is constrained by a somewhat limited vocabulary Demonstrates general understand of the requirements of academic writing
70	***Adept:*** Responds readily to the demands of the topic and presents information clearly and logically Uses the readings and lecture effectively to support the thesis Demonstrated understand of the requirements of academic writing
80-90	***Expert:*** Demonstrates mastery of appropriate, concise, and persuasive academic writing Writes with authority and style

Note. *The scale presented above was used to evaluate all essays included in this study (i.e., a retired version of the CAEL Assessment). However, this rating scale is no longer in use and has been replaced with an alternative version.*