Journal of Second Language Writing 34 (2016) 12–24

Contents lists available at ScienceDirect

Journal of Second Language Writing

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

# The relationship between lexical sophistication and independent and source-based writing

Kristopher Kyle, Assistant Professor[a,*], Scott Crossley, Associate Professor[b]

[a] University of Hawaii at Manoa, Honolulu, HI, United States
[b] Georgia State University, Atlanta, GA, United States

ARTICLE INFO

*Article history:*
Received 9 March 2015
Received in revised form 14 September 2016
Accepted 4 October 2016
Available online xxx

*Keywords:*
Lexical sophistication
Independent writing tasks
Source-based writing tasks
Writing assessment
N-grams
Natural language processing

ABSTRACT

Lexical sophistication is an important component of writing proficiency. New lexical indices related to range, n-gram frequency, psycholinguistic word information, academic language, polysemy, and hypernymy have yielded new insights into the construct of lexical sophistication and its relationship with second language (L2) acquisition and writing. For example, recent studies have suggested that range and bigram indices are stronger indicators of lexical sophistication than frequency in the context of L2 acquisition and L2 writing and speaking proficiency. This study explores the relationship between these newly developed indices of lexical sophistication and holistic scores of writing proficiency in both independent and source-based writing tasks. The results suggest that range and bigrams are important predictors of essay quality in independent tasks, but that lexical sophistication indices are not strong predictors of essay quality in source-based tasks. The results also indicate that responses to source-based tasks tend to include more sophisticated lexical items than responses to independent tasks. Implications for second language writing assessment and pedagogy are discussed.



## 1. Introduction

A key measure of academic success is writing proficiency (Kellogg & Raulerson, 2007). Becoming a proficient academic writer is a challenging and multifaceted endeavor, both for first language (L1) and second language (L2) writers (Crossley & McNamara, 2009; National Commission on Writing, 2003). Academic writers must learn to navigate a number of different task types (e.g., Römer & O'Donnell, 2011) that may differ in rhetorical (Cumming et al., 2005; Hyland, 2007) and linguistic (Cumming et al., 2005; Hardy & Römer, 2013) features. Differences in task types may also require writers to employ varied skills (Plakans & Gebril, 2013; Plakans, 2008) and linguistic resources (Guo, Crossley, & McNamara, 2013). These differences make it difficult to assess writing proficiency using a single task (Schoonen, van Gelderen, Stoel, Hulstjin, & de Glopper, 2011), and suggest that the construct of writing proficiency may better be described as a set of writing proficiencies. Accordingly, some high-stakes assessment tools assess writing proficiency using multiple writing tasks. The Test of English as a Foreign Language (TOEFL), for example, includes both independent tasks (i.e., tasks that ask test takers to draw on their personal experience when responding to a prompt) and source-based tasks (i.e., tasks that ask test takers to integrate information from source texts when responding to a prompt) in an effort to better reflect the types of writing tasks

* Corresponding author.
*E-mail addresses:* kristopherkyle1@gmail.com (K. Kyle), sacrossley@gmail.com (S. Crossley).

http://dx.doi.org/10.1016/j.jslw.2016.10.003
1060-3743/

encountered in academic settings (e.g., Chapelle, Enright, & Jamieson, 2008; Cumming et al., 2005). One important question that arises with the inclusion of multiple writing assessment tasks is whether the tasks elicit responses with distinct features (i.e., are assessing different aspects of writing proficiency) (Cumming et al., 2005).

Among the many features of writing proficiency that have been investigated, the role lexical knowledge plays in successful writing is well attested. Lexical knowledge can be considered both a receptive (Baba, 2009; Schoonen et al., 2011) and a productive (Kyle & Crossley, 2015; Laufer & Nation, 1995; Laufer, 1994) trait. Receptive lexical knowledge refers to an individual's ability to understand the meaning of a lexical item that is read or heard, and is often assessed using standardized tests such as the Vocabulary Levels Test (Schmitt, Schmitt, & Clapham, 2001) or the Word Associates Test (Read, 1998). Productive lexical knowledge refers to the words available to an individual when writing or speaking. Productive lexical knowledge is often assessed by examining the lexical sophistication of a speaking or writing sample. Lexical sophistication is generally related to the diversity (e.g., Engber, 1995) and/or the relative difficulty (often based on corpus frequency counts; e.g., Laufer & Nation, 1995) of the lexical items in a text.

Research has demonstrated that L2 writers pay particular attention to lexical concerns as they construct texts (e.g., Cumming, 1990; Leki & Carson, 1994; Manchón, Murphy, & Roca de Larios, 2007), and links have been reported between both receptive and productive lexical knowledge and writing proficiency scores in relation to both independent (Guo et al., 2013; Schoonen et al., 2011) and source-based writing tasks (e.g., Baba, 2009; Guo et al., 2013). One important question that has not been thoroughly addressed, however, is whether productive lexical knowledge (i.e., lexical sophistication) is uniformly important across writing task types. Previous research (Guo et al., 2013) indicates that independent and source-based writing tasks differ in the lexical features that are predictive of writing proficiency scores. Specifically, they found that word familiarity and frequency were predictive of source-based writing quality, while word length and hypernymy were predictive of independent writing. However, Guo et al. did not investigate a number of lexical features theorized to be important components of essay quality, nor did the study consider if source-based and independent tasks led to production differences in lexical output. Such differences may provide a greater understanding of how the two task types differ and provide stronger rationale for the use of the two tasks when assessing writing skills. In addition, differences in the two task types may inform automated essay scoring systems that rely on linguistic features. The current study builds on previous research such as Guo et al. (2013) by exploring a wide range of lexical features across independent and integrated tasks.

### *1.1. Receptive lexical knowledge and writing proficiency*

A number of studies have found links between receptive lexical knowledge and writing proficiency scores in independent (Koda, 1993; Schoonen et al., 2011) and source-based tasks (e.g., Baba, 2009). Koda (1993), for example, found a strong positive relationship ($r = .70$) between receptive language knowledge and holistic scores of writing proficiency for English L1 writers of L2 Japanese. In addition, Schoonen et al. (2011) used scores on a receptive vocabulary task as a predictor of L2 writing proficiency scores, finding moderate, positive correlations ranging from $r = .53$ to $r = .57$. Baba (2009) also found moderate, positive relationships between holistic scores of quality on a summary writing task and vocabulary size ($r = .40$) and vocabulary depth ($r = .34$). Together, these findings suggest that receptive lexical knowledge may be an important indicator of writing proficiency, but do not indicate how receptive lexical knowledge translates into linguistic production.

### *1.2. Productive lexical knowledge and writing proficiency*

Links have also been made between productive vocabulary knowledge (i.e., lexical sophistication) and writing proficiency scores. Lexical sophistication has traditionally been operationalized as the diversity of the words used in a text (e.g., the number of unique words in a text divided by the total number of a text; Engber, 1995) or by the average reference-corpus frequency of words in a text (e.g., Laufer & Nation, 1995).

#### *1.2.1. Lexical diversity*

One way that productive vocabulary knowledge has been measured is by using lexical diversity measures, such as the type-token ratio (e.g., Engber, 1995) or more sophisticated measures such as *D* (Jarvis, 2002; Malvern & Richards, 1997). Indices of lexical diversity measure the variety of words used in a text. Generally, positive relationships have been found between lexical diversity and writing proficiency scores. Engber (1995), for example, found a moderate, positive correlation between lexical diversity and holistic scores of L2 writing proficiency with regard to an independent writing task. Essays that included more diverse lexical items tended to earn higher holistic scores. A number of other studies have found similar trends with regard to holistic scores of L2 writing proficiency with independent writing tasks (e.g., Cumming et al., 2005; Grant & Ginther, 2000; Jarvis, 2002). Cumming et al. (2005) also extended these findings to source-based writing tasks, finding a positive relationship between type-token ratios and holistic scores in read-write and listen-write tasks. However, recent research has indicated that lexical diversity is more strongly related to text cohesion than lexical sophistication. Specifically, lexical diversity captures the repetition of words across a text (i.e., lexical overlap) and is thus a measure of lexical cohesion (Crossley, Kyle, & McNamara, 2015). This contrasts with lexical sophistication indices which capture text-external features of words (such as reference corpus frequency).

#### 1.2.2. Word frequency

Lexical sophistication is perhaps most often operationalized using the reference-corpus frequency of the words in a text (e.g., Attali & Burstein, 2006; Crossley, Cobb, & McNamara, 2013; Crossley & McNamara, 2012; Enright & Quinlan, 2010; Laufer & Nation, 1995). Words that occur less frequently are considered sophisticated (e.g., *solidification*, *octogenarians*, *modularized*) while frequent words (e.g., *people*, *place*, *number*) are considered less so (Kyle & Crossley, 2015). Research suggests that more proficient writers will, on average, use less frequent words when writing in response to independent tasks. Laufer and Nation (1995), for example, found that more proficient L2 writers used fewer high frequency words (i.e., words that comprise the most frequent 1000 words in English). Others have found similar results with regard to average reference corpus frequency of words in independent L2 compositions, finding that on average, more proficient L2 writers use less frequent words than less proficient ones (Attali & Burstein, 2006; Crossley & McNamara, 2012; Crossley et al., 2013; Enright & Quinlan, 2010; Guo et al., 2013). Average reference frequency has also been investigated with regard to source-based tasks. Guo et al. (2013), for example, found that frequency was negatively correlated with TOEFL source-based task scores, which is in line with studies that have explored frequency in independent tasks.

### 1.3. Expanding the construct of lexical sophistication

Recent studies have used newer automated indices to measure lexical sophistication in both L1 and L2 contexts. Researchers have, for example, used the average reference corpus word range (i.e., the percentage of texts in a reference corpus that a word occurs in; Kyle & Crossley, 2015), the average reference corpus bigram and trigram (i.e., two and three word sequences), frequency (Crossley, Cai, & McNamara, 2012), the use of academic words and phrases (Kyle & Crossley, 2015), the psycholinguistic properties of words (Crossley & McNamara, 2013; Guo et al., 2013), and the semantic relationships words have (i.e., hypernymy and polysemy; Guo et al., 2013) to measure the relationship between lexical sophistication and holistic scores of both L1 and L2 writing proficiency. These studies have demonstrated that indices beyond word frequency can add to models of lexical sophistication and contribute to our understanding of the relationship between lexical sophistication and L2 writing proficiency. We discuss these newer indices below.

#### 1.3.1. Word range

Range, which is also referred to as *dispersion*, *entropy*, and *contextual diversity*, is a measurement of the number of texts in a reference corpus in which a word occurs (Gries, 2008). Words with high range values (e.g., *people*, *book*, *building*) occur widely throughout a number of different texts and contexts, while words with low range values (e.g., *antifungal*, *lithosphere*, *deictic*) tend to be restricted in use to a smaller number of texts and contexts. The average reference-corpus range has been shown to be negatively correlated with analytic scores of lexical proficiency (Kyle & Crossley, 2015), suggesting that words with a narrower range are more sophisticated than those with a wider range. While links between range and analytic scores of lexical sophistication have been investigated, the link relationship between range and holistic scores of writing proficiency in independent and source-based writing has not.

#### 1.3.2. N-gram frequency

N-grams, or multi-word expressions of *n* words in length, have been of increasing interest in lexical sophistication over the past ten years (e.g., Biber, Conrad, & Cortes, 2004). N-gram frequency has been shown to be correlated with analytic scores of lexical proficiency (Kyle & Crossley, 2015). L2 texts that include more frequent n-grams (e.g., *one of the*, *as well as*, *as a result*) tend to earn higher analytic scores of lexical sophistication than those with infrequent n-grams (e.g., *the former east*, *their hands and*, *of the no*). The relationship between n-gram frequency and independent L1 writing has also been explored with contradictory results (i.e., n-gram frequency was negatively correlated with holistic scores of writing proficiency; Crossley et al., 2012) but has not, to our knowledge, been investigated with relation to independent or source-based L2 writing.

#### 1.3.3. Academic language

Academic language has been defined as words and/or multi-word units (n-grams) that occur frequently in academic texts such as research journal articles, textbooks and academic lectures, but relatively infrequently in general corpora (Coxhead, 2000; Simpson-Vlach & Ellis, 2010; Xue & Nation, 1984). The academic word list (AWL; Coxhead, 2000), for example, includes 570 word families that represent academic language at the word level (e.g., *analyze*, *transmit*, *statistic*). The academic formulas list (AFL; Simpson-Vlach & Ellis, 2010) includes multi-word units that are frequent in academic texts, but less frequent in general corpora (e.g., *in relation to*, *the development of*, *the fact that*). Theoretically, higher proficiency writers are expected to use more academic words and n-grams when responding to academic writing tasks (Morris & Cobb, 2004; Simpson-Vlach & Ellis, 2010). Morris and Cobb (2004), for example, found that independent L2 essays written by higher proficiency learners tended to include more AWL words. Other recent research, however, has suggested academic language (i.e., words from the AWL and AFL) may occur infrequently in some collections of learner writing (Kyle & Crossley, 2015), regardless of proficiency level.

#### 1.3.4. Psycholinguistic properties of words

The psycholinguistic properties of words have been of interest to cognitive scientists for some time (e.g., Coltheart, 1981; Toglia & Battig, 1978), but within the last five years have been connected to the construct of lexical sophistication. Word properties such as concreteness, familiarity, imageability, meaningfulness, and age of acquisition have been linked to analytic scores of lexical proficiency (Kyle & Crossley, 2015) and holistic scores of L2 writing proficiency (Guo et al., 2013). Kyle and Crossley (2015) found that informal L1 and L2 written texts that included less concrete, familiar, imageable, and meaningful words tended to earn higher lexical proficiency scores. Additionally, essays that included words that on average are learned later also tended to earn higher analytic lexical proficiency scores. Guo et al. (2013) found similar relationships between concreteness, familiarity, imageability, and meaningfulness and holistic writing proficiency scores for TOEFL independent and source-based writing tasks. The relationship between word age of acquisition has not, to our knowledge, been used to model holistic scores of L2 writing proficiency with regard to either independent or source-based writing tasks.

#### 1.3.5. Semantic relationships

Polysemy and hypernymy are types of semantic relationships related to lexical development (Crossley, Salsbury, & McNamara, 2009; Crossley, Salsbury, & McNamara, 2010) and L2 writing proficiency scores(Guo et al., 2013; Reynolds, 1995). Polysemy refers to the number of different (but related) senses (i.e., meanings) a word form has. A word such as *table*, for example, has more senses (i.e., a piece of furniture, geographic feature, etc.) than a word such as *encephalon* (which has only a single sense). Hypernymy refers to the hierarchical relationships between words. Words with high hypernymy values have a large number of superordinate terms, while words with low hypernymy values have few (or no) superordinate terms. For example, *animal* has six hypernymic terms (e.g., organism, animate thing, etc.), while *dog* has thirteen hypernymic terms (e.g., canine, carnivore, mammal, etc.). Recently, Guo et al. (2013) found that L2 independent essays that include words with fewer senses (i.e., lower polysemy scores) tend to earn higher scores, but no links were found between polysemy and source-based writing. Guo et al. (2013) also found a positive relationship between hypernymy (for nouns) and writing proficiency scores for both independent and source-based writing tasks.

### 1.4. Lexical sophistication and automatic essay scoring models

Over the past decade, automatic essay scoring (AES) systems have become increasingly prevalent in standardized writing assessments (Attali & Burstein, 2006; Shermis & Burstein, 2013). Such systems can decrease the time and costs related to essay scoring while also increasing test reliability (Bereiter, 2003; Dikli, 2006; Higgins, Xi, Zechner, & Williamson, 2011). Although AES systems can achieve scoring accuracy that is on par with humans (e.g., Shermis & Hamner, 2013; Weigle, 2010), they do not assess essays in the same manner as humans do and cannot measure some essential aspects of the construct of writing such as argumentation or rhetorical effectiveness (Condon, 2013; Deane, 2013; Herrington & Moran, 2001). An important area of interest has been to increase the construct coverage of AES systems (e.g., Crossley et al., 2015; Enright & Quinlan, 2010). To facilitate this, it is necessary to investigate features that are important to human evaluations of writing proficiency (such as lexical sophistication, among many others) and evaluate ways to measure them automatically (Burstein, Marcu, & Knight, 2003; Enright & Quinlan, 2010; Kyle & Crossley, 2015; Kyle, 2016).

Due to the importance of lexical sophistication in L2 writing (Cumming, 1990; Laufer & Nation, 1995; Schoonen et al., 2011), prominent AES systems include indices of lexical sophistication in their scoring models. Given the proprietary nature of most AES models, however, relatively little is known regarding their use of these indices. Enright and Quinlan (2010) report that e-rater, which is used to score both independent and source-based TOEFL writing tasks, uses two lexical complexity indices (word length and the use of less frequent words). Respectively, these indices account for 7% and 4% of the score produced by e-rater. Foltz, Streeter, Lochbaum, and Landauer (2013) indicate that the Intelligent Essay Assessor (IEA), which is primarily used for assessing L1 writing, includes features related to lexical sophistication such as word maturity and word variety. Foltz et al. do not, however, indicate the weights given to these indices in scoring models.

### 1.5. Current study

While a number of previous studies have explored the relationship between lexical sophistication and holistic scores of writing proficiency, few (if any) studies have investigated the newly developed measures of lexical sophistication outlined in the previous section. Additionally, most of the extant studies regarding the relationship between lexical sophistication and writing proficiency scores have explored independent writing tasks (cf. Cumming et al., 2005; Guo et al., 2013). This has resulted in gaps in our understanding regarding (a) the relationship between lexical sophistication and independent and source-based writing task proficiency scores, and (b) whether independent and source-based tasks require different linguistic resources with regard to lexical sophistication. We address these gaps by examining the relationship between newly developed indices of lexical sophistication and holistic scores of writing proficiency in relation to independent and source-based TOEFL writing tasks.

Specifically, this study is guided by the following research questions:

1. What is the relationship between lexical sophistication and independent writing task proficiency scores?
2. What is the relationship between lexical sophistication and source-based writing task proficiency scores?

**Table 1**
Descriptive statistics for the essays included in the TOEFL public use dataset: mean (standard deviation).

| Corpus | N | Score | Number of words |
|---|---|---|---|
| Independent | 480 | 3.427 (.887) | 315.600 (78.596) |
| Form 1 | 240 | 3.383 (.864) | 321.830 (79.720) |
| Form 2 | 240 | 3.471 (.910) | 309.38 (77.117) |
| Integrated | 480 | 3.151 (1.244) | 200.44 (51.692) |
| Form 1 | 240 | 3.254 (1.179) | 204.840 (52.437) |
| Form 2 | 240 | 3.148 (1.308) | 196.040 (50.662) |

3. Do responses to independent and source-based writing tasks differ with regard to lexical sophistication?

## 2. Method

### 2.1. Corpus

We selected independent and integrated (i.e., source-based) essays written by 480 individuals as part of the TOEFL that comprise the Educational Testing Service (ETS) public data set. The corpus includes responses to two forms of the TOEFL (i.e., includes responses to two independent essay prompts and two integrated essay prompts). The independent prompts ask test takers to write an essay that asserts and defends an opinion on a particular topic based on their own life experience. The integrated prompt asks test takers to read a short passage, listen to a related lecture, and synthesize the information given in the reading and the lecture. Essays were rated by two trained raters employed by ETS using a scale that ranged from 1.0 to 5.0. Any scores that differed by one point or less were averaged. If any two ratings for an essay differed by more than a single point, a third rater evaluated the essay. The holistic rating rubric used to evaluate the independent tasks includes descriptors related to the completion of the task, organization, development of ideas, coherence, word and phrase use, and syntax. The holistic rating rubric used to evaluate integrated tasks includes descriptors mostly related to content (e.g., whether test takers appropriately summarized the two passages and responded to the task).[1] Table 1 comprises descriptive statistics for the corpus including the number of essays, the average score, and the average number of words from each form and task.

### 2.2. Indices of lexical sophistication

We examined a number of indices included in the freely available Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle & Crossley, 2015) that represent a wide range of theoretically important aspects related to lexical sophistication. In order to control for essay length (e.g., Chodorow & Burstein, 2004) we selected the 114 length-normalized indices in TAALES. TAALES calculates indices of lexical sophistication for unigrams (words), bigrams, and trigrams. Included are a number of frequency indices from a variety of both spoken and written corpora (e.g., the British National Corpus), range indices from spoken and written corpora (e.g., SUBTLEXus), academic language indices (e.g., Academic Word List; Coxhead, 2000), and psycholinguistic word information indices (e.g., concreteness; Brysbaert, Warriner, & Kuperman, 2014). We also wrote a Python script that employs the Natural Language Toolkit (NLTK; Bird, Klein, & Loper, 2009) to calculate four indices of polysemy and hypernymy.[2] We present more details of the selected indices below.

#### 2.2.1. Word frequency indices

Word frequency indices are calculated by taking the sum of the frequency values with regard to a particular frequency list (e.g., the BNC) for words in a text and dividing that sum by the number of words in that text. If a word in a target text is not represented in the frequency list, the word is not included in the calculation of the index. TAALES calculates frequency indices based on a number of reference corpora-derived frequency lists. Lemmatized frequency indices are derived from the 4.5-million word Thorndike-Lorge corpus of popular magazine articles (Thorndike & Lorge, 1944), the 1-million word written section of the Brown corpus (Kučera & Francis, 1967), and the 1-million word London-Lund Corpus of English Conversation (Brown, 1984). Un-lemmatized frequency indices are derived from 80-million word written and 10-million word spoken subsets of the British National Corpus (BNC Consortium, 2007) and the 51-million word SUBTLEXus corpus of American subtitles (Brysbaert & New, 2009). For each list, TAALES includes an index for all words (AW), content words (CW), and function words (FW). Additionally, TAALES calculates logarithm-transformed frequency indices.

[1] The independent and integrated scoring rubrics are freely available on the ETS website at http://www.ets.org/Media/Tests/TOEFL/pdf/Writing_Rubrics.pdf.
[2] Both this script and TAALES are freely available at http://www.kristopherkyle.com/.

*2.2.2. Range indices*

TAALES includes a number of range indices. Range indices are calculated for AW, CW, and FW. These indices are derived from the 500 texts in the Brown corpus (Kučera & Francis, 1967), 574 spoken and 3083 written texts in the BNC (BNC Consortium, 2007), and the 8388 subtitle texts in SUBTLEXus (Brysbaert & New, 2009). Additionally, TAALES calculates indices based on the 15 text categories in the Brown Corpus (Kučera & Francis, 1967), which can be described roughly as genres (e.g., news reporting, news editorials, academic writing, science fiction, mystery and detective fiction). Scores from these indices calculate the average number of text categories in which the words in a text occur. Words that occur in all categories are general-purpose words, while words that occur in only one category are more restricted in their use.

*2.2.3. N-gram indices*

Crossley et al. (2012) derived bigram and trigram frequency lists from the 80-million word written and the 10-million word spoken subsets of the BNC (BNC Consortium, 2007). TAALES calculates normalized n-gram frequency counts. For each list, normalized counts are calculated using the number of words in the text as the denominator and by using the number of bigrams/trigrams in the text that are also represented in the frequency list as the numerator. Additionally, the proportion of unique bigrams/trigrams that occur in a target text that also occur in the frequency list are calculated.

*2.2.4. Academic list indices*

TAALES includes indices derived from the AWL and the AFL. For the AWL, indices are calculated for the entire AWL and for each of the 10 sublists (see Coxhead, 2000; for more information on the AWL). For the AFL, indices are calculated for the entire AFL, the "core" AFL, the written AFL and the spoken AFL (see Simpson-Vlach & Ellis, 2010; for more information on the AFL).

*2.2.5. Word information indices*

TAALES includes a number of psycholinguistic word information indices based on the MRC psycholinguistic database (Coltheart, 1981) and two newly collected databases (Brysbaert et al., 2014; Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012). Indices are calculated for AW, CW, and FW. Included are familiarity, concreteness (for words and bigrams), imageability, meaningfulness, and age of acquisition.

*2.2.6. Polysemy and hypernymy*

We calculate polysemy and hypernymy indices based on the Wordnet database (Fellbaum, 1998). We calculate polysemy as the mean number of senses contained in content words (nouns, verbs, adjectives, and adverbs). Hypernymy indices comprise the mean number of superordinate terms words in a text have. We calculate hypernymy for nouns, verbs, and the combination of nouns and verbs.

*2.3. Statistical analysis*

In order to determine the relationship between a variety of indices of lexical sophistication and L2 independent and integrated writing assessment scores, we conducted two identical sets of statistical analyses that differed only in terms of the corpora analyzed (i.e., one with the data from the independent essays and one with the data from the integrated essays). Following the procedure outlined below, we ran a stepwise multiple regression to determine the amount of variance in holistic scores that could be explained by indices of lexical sophistication.

We first checked to ensure that each index was normally distributed. Any indices that did not meet the criteria of normality were discarded.[3] We then conducted a multiple analysis of variance (MANOVA) statistic between the data from the two prompts to control for prompt differences (e.g., Crossley, Weston, McLain Sullivan, & McNamara, 2011; Hinkel, 2002). Any indices that were significantly different ($p < .05$) between prompts were removed from further consideration. We then ran a correlation between the remaining indices and holistic essay quality score. Any indices that did not demonstrate a significant ($p < .05$) and meaningful relationship ($r > .1$) with holistic scores were removed from further consideration. We also removed any indices that were strongly correlated ($r \geq .7$) with the number of words in each essay to control for text length, which strongly affects human judgments of quality (Ferris, 1994). We then checked the remaining indices for multicollinearity. Any indices that were very strongly correlated ($r \geq .9$) were flagged and in each collinear set the index with the strongest relationship with holistic scores was kept (Tabachnick & Fidell, 2001). A stepwise multiple regression was then conducted. If the resulting model included any variables with switched signs (i.e., the stepwise model used the inverse of a variable to create an optimal model), the variable was removed from consideration to ensure that the final model reflected the initial correlations and the regression was run again. To ensure that the results of the multiple regression were consistent across the entire data set, a multiple regression with 10-fold cross validation (10-fold CV) was conducted using the indices identified in the initial multiple regression. 10-fold CV is a statistical procedure in which 90% of the data are used to create a

[3] In fine-grained linguistic analyses, normality is often grossly violated due to a high-number of zero-counts (i.e., for rare features). In such cases, variable transformation is not possible. While aggregated features can also be used to obtain normally distributed data, the use of non-aggregated indices tend to produce more accurate and fine-grained results (Crossley et al., 2015).

model, and then that model is tested on the remaining 10% of the data. This procedure is repeated until all of the data have been used as the test set, and then the results from the ten models (or folds) are averaged (e.g., Tabachnick & Fidell, 2001).

To investigate the differences between independent and integrated TOEFL essays with regard to lexical sophistication we conducted MANOVA and discriminant function analysis (DFA) statistics. We first conducted a MANOVA using essay type (independent/integrated) as fixed factors and the predictors identified in the regression analyses above as dependent variables. We then entered the indices that demonstrated significant and meaningful differences into a DFA. DFA is often used to predict group membership of items (e.g., independent and integrated essays) based on predictor variables (e.g., indices of lexical sophistication). We used a stepwise DFA on the entire data set and employed leave one out cross validation (LOOCV) to ensure that the model is generalizable across the data set.

## 3. Results

### 3.1. Independent essays

#### 3.1.1. Assumptions

Assumptions Of the 118 indices considered, 15 violated normality (the majority of these were due to zero counts for AWL and AFL lists) and were removed from further consideration. Of the 103 indices remaining, 26 were not meaningfully correlated (absolute value of $r \geq .1$) with holistic scores. None of the indices were strongly correlated ($r >= .7$) with essay length. Of the remaining 77 indices, 55 demonstrated significant differences between prompts. The remaining 22 indices were analyzed for multicollinearity. After checking for multicollinearity, 14 variables remained. See Table 2 for correlations between the remaining variables and holistic scores.

#### 3.1.2. Regression analysis

The 14 variables were entered into a stepwise multiple regression in order to determine the variance in holistic scores explained by indices of lexical sophistication. After removing variables that switched signs, the regression analysis yielded a significant model that included six indices of lexical sophistication: *BNC written range for all words*, *BNC written bigram frequency logarithm*, *hypernymy (nouns and verbs), and imageability for all words*. The model accounted for 36.8% of the variance in holistic scores (see Table 3 for an overview of the model). A follow up 10-fold CV multiple regression explained 35.4% of the variance, indicating that the model is stable across the dataset. The results indicate that range, bigram frequency, hypernymy and imageability are important predictors of TOEFL independent essay quality.

### 3.2. Integrated essays

#### 3.2.1. Assumptions

Assumptions Of the 118 indices considered, 20 violated normality (the majority of these were due to zero counts for AWL and AFL lists) and were removed from further consideration. Of the 98 indices remaining, 50 were not meaningfully correlated (absolute value of $r \geq .1$) with holistic scores. None of the indices were strongly correlated ($r >= .7$) with essay length. Of the remaining 48 indices, 39 demonstrated significant differences between prompts. The remaining nine indices were analyzed for multicollinearity. After checking for multicollinearity, six variables remained. See Table 4 for correlations between the remaining variables and holistic scores.

**Table 2**
Correlations between indices entered into stepwise multiple regression and holistic score for independent writing.

| Index | N | *r* | *p* |
|---|---|---|---|
| *BNC Written Range AW* | 480 | −0.409 | <.001 |
| *Kuperman age of acquisition CW* | 480 | 0.403 | <.001 |
| *SUBTLEXus Range CW* | 480 | −0.398 | <.001 |
| *Familiarity CW* | 480 | −0.392 | <.001 |
| *Hypernymy (nouns and verbs)* | 480 | 0.372 | <.001 |
| *Kucera-Francis Frequency CW Logarithm* | 480 | −0.361 | <.001 |
| *BNC Spoken Frequency CW* | 480 | −0.337 | <.001 |
| *Kucera-Francis Frequency FW Logarithm* | 480 | 0.274 | <.001 |
| *BNC Written Trigram Proportion* | 480 | 0.216 | <.001 |
| *Hypernymy (verbs)* | 480 | .195 | <.001 |
| *Imageability AW* | 480 | −0.161 | <.001 |
| *BNC Written Bigram Proportion* | 480 | 0.151 | <.001 |
| *BNC Spoken Freq FW* | 480 | 0.109 | <.001 |
| *BNC Written Bigram Frequency Logarithm* | 480 | 0.107 | <.010 |

**Table 3**
Summary of stepwise multiple regression models for independent writing.

| Entry | Predictors included | *r* | $R^2$ | $R^2$ change | *B* | β | *SE* |
|---|---|---|---|---|---|---|---|
| 1 | *BNC Written Range AW* | .409 | .167 | .167 | −.599 | −.196 | .018 |
| 2 | *BNC Written Bigram Frequency Logarithm* | .588 | .345 | .178 | .457 | 2.345 | .243 |
| 3 | *Hypernymy (nouns and verbs)* | .596 | .355 | .010 | .148 | .332 | .104 |
| 4 | *Imageability AW* | .606 | .368 | .013 | −.120 | −.013 | .004 |

*Note*: Estimated constant term = 18.552, β = unstandardized beta, SE = standard error; B = standardized beta.

**Table 4**
Correlations between indices entered into stepwise multiple regression and holistic score for integrated writing.

| Index | N | *r* | *p* |
|---|---|---|---|
| *Hypernymy (nouns)* | 480 | 0.263 | <.001 |
| *Kucera-Francis number of categories AW* | 480 | −0.176 | <.001 |
| *Thorndike-Lorge Frequency AW Logarithm* | 480 | −0.171 | <.001 |
| *BNC Written Bigram Frequency Logarithm* | 480 | −0.122 | .007 |
| *Kuperman age of acquisition FW logarithm* | 480 | 0.109 | .017 |
| *Hypernymy (nouns and verbs)* | 480 | 0.108 | .018 |

**Table 5**
Summary of stepwise multiple regression models for integrated writing.

| Entry | Predictors included | *r* | $R^2$ | $R^2$ change | *B* | β | *SE* |
|---|---|---|---|---|---|---|---|
| 1 | *Hypernymy (nouns)* | .263 | .069 | .069 | .234 | .569 | .110 |
| 2 | *Kucera-Francis number of categories AW* | .288 | .083 | .014 | −.121 | −.448 | .167 |

*Note*: Estimated constant term = 5.628, β = unstandardized beta, SE = standard error; B = standardized beta.

### 3.2.2. Regression analysis

The six variables were entered into a stepwise multiple regression in order to determine the variance in holistic scores explained by indices of lexical sophistication. The stepwise regression resulted in a significant model including two variables (*hypernymy [nouns]* and *Kucera-Francis number of categories*). The model accounted for 8.3% of the variance in holistic scores (see Table 5 for an overview of the model). A follow up 10-fold CV multiple regression explained 7.5% of the variance, indicated that the model is stable across the dataset. The results demonstrate that hypernymy and number of categories are indicators of TOEFL integrated essay quality.

## 3.3. Differences between independent and integrated essays

### 3.3.1. MANOVA

The MANOVA indicated that each index identified in the regression models above demonstrated significant and meaningful differences between independent and integrated essays. These results are summarized in Table 6.

**Table 6**
MANOVA Results.

| Variable | Independent Mean (SD) | Integrated Mean (SD) | *F* (1, 957) | $\eta^2_p$ |
|---|---|---|---|---|
| *BNC Written Range AW* | 80.918(2.712) | 72.858(3.702) | 1480.87 | .607 |
| *BNC Written Bigram Frequency Logarithm* | 1.523(0.173) | 1.252(0.196) | 517.34 | .351 |
| *Hypernymy (nouns)* | 5.422(0.509) | 6.155(0.512) | 495.11 | .341 |
| *Hypernymy (nouns and verbs)* | 3.488(0.395) | 4.292(0.423) | 927.56 | .492 |
| *Imageability AW* | 319.283(8.473) | 337.049(18.005) | 382.59 | .285 |
| *Kucera-Francis number of categories AW* | 14.242(0.262) | 13.355(0.336) | 2076.02 | .684 |

*Note*: For all indices $p < .001$.

**Table 7**
Discriminant function analysis confusion matrix.

| | Predicted Independent | Predicted Integrated | Total |
|---|---|---|---|
| Independent | 460 | 20 | 480 |
| Integrated | 35 | 445 | 480 |
| Accuracy | 95.8% | 92.7% | 94.3% |

*3.3.2. DFA*

Following the MANOVA, the six indices identified in the regression models above were checked for multicollinearity. The two range indices (*BNC written range for all words* and *Kucera-Francis number of categories*) were collinear ($r > .9$). Because the latter demonstrated stronger differences between independent and integrated essays, the former was removed from further consideration. The remaining five indices were entered into a stepwise DFA. The model created by the stepwise DFA achieved a classification accuracy of 94.3% accuracy using three indices (*Kucera-Francis number of categories*, *imageability for all words*, and *hypernymy [nouns]*). This is significantly higher ($df = 1$, $n = 960$, $\chi^2 = 753.340$, $p < .001$) than what would be expected by chance. The reported Kappa = .885, indicates almost perfect agreement between actual and predicted essay type (Landis & Koch, 1977). The stepwise LOOCV DFA also achieved a classification accuracy of 94.3%, suggesting that the predictor model is stable across the dataset. Table 7 comprises the confusion matrix for the stepwise DFA, which shows the number of independent and integrated essays that were correctly predicted by the model. The results indicate that responses to independent and integrated tasks can be accurately distinguished based on indices of lexical sophistication related to range of registers, imageability, and hypernymy.

## 4. Discussion

The results suggest that the production of sophisticated lexis is an important predictor of holistic scores of writing proficiency with regard to independent tasks, but not with regard to source-based tasks. Specifically, the results indicate that indices of word range and bigram frequency, which have not been explored with regard to timed, argumentative L2 writing, are important indicators of holistic writing proficiency scores in independent TOEFL writing samples, but not in TOEFL integrated writing samples. These findings underscore the complexity of the construct(s) of L2 writing proficiency (e.g., Schoonen et al., 2011) and highlight differences with regard to the relationship between receptive and productive lexical knowledge and holistic scores of writing proficiency in integrated tasks. Baba (2009), for example, found a positive relationship between receptive vocabulary knowledge and holistic writing proficiency scores for integrated tasks, while the current study found only a weak relationship between productive lexical knowledge (as measured by indices of lexical sophistication) and holistic integrated writing scores. Furthermore, lexical sophistication differs between responses to independent and integrated tasks in ways that allow for independent and integrated essays to be accurately categorized based solely on indices of lexical sophistication. The findings are important because they suggest that independent and integrated writing tasks lead to the production of different linguistic features supporting the notion that the two tasks measure distinct writing proficiency constructs (Chapelle et al., 2008). Additionally, the findings suggest that range and bigram frequency are stronger predictors of independent essay quality than word frequency, which has been the most common metric of lexical sophistication used in previous writing quality studies (e.g., Crossley et al., 2013; Laufer & Nation, 1995). This finding has important implications for automatic essay scoring and feedback systems along with implications for L2 writing pedagogy. We discuss the importance of these findings below divided by each analysis.

*4.1. Lexical sophistication and independent essay tasks*

A multiple regression model consisting of four indices of lexical sophistication explained 36.8% percent of the variance in holistic independent essay scores (35.4% in the LOOCV). Importantly, the inclusion of two variables, *BNC Written Range AW* and *BNC Written Bigram Frequency Logarithm*, explained a large portion of this variance (16.7% and 17.8%, respectively). Two other indices explained the remaining 2.3% of the variance, including *Hypernymy (nouns and verbs*; 1.0%) and *Imageability AW* (1.3%).

These results indicate that raters tend to assign higher scores to essays that include words with a more restricted range. This suggests that essays that include words that are more specific and language-domain appropriate are likely to receive higher scores. Examples 1 and 2 below highlight differences between sentences with high and low average range scores. Both sentences are taken from essays written on the topic of cooperation.

(1) Low range example (Quality score of 4.5. Range score of 65.96): *Second, the accelerated effect of globalization since the fifties has accentuated the need for cooperation.*

(2) High range example (Quality score of 3. Range score of 79.89): *First, commercial between countries had been expanded and became bigger and that would make countries to think how to deal with other countries more than past days before this revolution of communication and information spreading.*

On average, the words in the low range sample occur in fewer than two-thirds of the texts in the written section of the BNC. Words such as *accelerated, globalization, fifties,* and *accentuated* have particularly low range scores (averaging 7.55), suggesting that they are used in a smaller percentage of texts and are thus more sophisticated. On average, the words in the high range sample occur in almost 80% of the texts in the written section of the BNC. Words such as *revolution, communication*, and *spreading* have relatively low range scores (averaging 31.27), but the majority of the words in the sentence have high range scores, indicating they are less lexically sophisticated. The two sentences communicate similar ideas (i.e., cooperation between countries is becoming increasingly important) but the first employs words that occur in fewer contexts, while the second uses more general words.

These results are novel with regard to independent essay tasks, and suggest that range is an important indicator of holistic L2 writing proficiency scores. Furthermore, in this study and in Kyle & Crossley's (2015) lexical proficiency study, range was a stronger predictor of holistic scores of writing proficiency and written lexical proficiency than frequency. This suggests that predicting productive L2 lexical proficiency is more closely related to the number of texts a word occurs in than the sheer number of times it occurs. This provides some evidence that range indices should be considered in the creation of vocabulary learning lists and in automatic essay scoring models.

Further, these results indicate that raters tend to assign higher scores to essays that include more frequent bigrams. This suggests that using appropriate word combinations may be important for earning higher scores on independent essays. Examples 3 and 4 below illustrate the difference between sentences that earn high and low bigram frequency scores.

(3) High bigram frequency example (total essay quality score 4.5, Bigram score of 1.6): *I think it is true that at schools and companies today group work is valued more than before.*

(4) Low bigram frequency example (total essay quality score 3, Bigram score of 0.7): *The world is developing very fast, the compete is change more hard.*

The bigram frequency database used in TAALES includes the most frequent 50,000 bigrams in a written subset of the BNC. The first example earns a relatively high bigram frequency score because most of the bigrams in the sentence occur in the database and are relatively frequent. Of the 17 bigrams in Example 3, twelve are represented in the database. Interestingly, the sentence is well formed, but *at schools*, *companies today*, *today group*, *is valued*, and *valued more* are not among the 50,000 most frequent bigrams in the BNC. In Example 4, which earns a lower bigram frequency score, only four of the 11 bigrams occur in the database, indicating that they are not among the most frequent 50,000 bigrams in a written subset of the BNC. In the first clause of the sentence, *The world is developing very fast,* which is well-formed, four of the five bigrams are counted, while none of the bigrams in the second clause *the compete is change more hard* are counted. These differences result in an average bigram frequency score that is twice as large in the first example than in the second. The bigram frequency index, therefore, seems to be tapping into both lexical knowledge (e.g., collocational knowledge) and grammatical knowledge, supporting the notion that lexis and grammar are intertwined (Halliday, 1991; Römer, 2009; Sinclair, 1991). These results are novel with regard to timed argumentative independent essays, and add to a growing body of literature that underscores the predictive value of n-gram frequency measures (e.g., Crossley et al., 2012; Kyle & Crossley, 2015). This evidence suggests the importance of n-gram frequency in predicting L2 writing proficiency scores.

Hypernymy (*Hypernymy [nouns and verbs]*) and imageability (*imageability AW*), together explained a small amount (2.3%) of the variance above n-grams and range indices. Essays tended to be assigned higher scores by raters if they included words that had more superordinate terms (i.e., were more specific) and less imageable. These results support previous research in the area of word information and hypernymy in L2 writing in that more essays that contain more specific words are scored higher (e.g., Guo et al., 2013). However, these results contrast with previous research that indicates language learners use words with fewer superordinate terms (i.e., are more abstract) as their proficiency develops (e.g., Crossley et al., 2009). This is likely because human ratings of writing proficiency are more strongly linked to the use of specific textual examples that support a claim whereas the assessment of lexical proficiency is more strongly linked to lexical abstractness.

The results of the regression analysis for the independent essays align well with the TOEFL independent writing rubric. This rubric asks raters to consider features that are strongly related to lexical sophistication such as "appropriate word choice," "range of vocabulary," and "idiomaticity." The findings of this analysis suggest that raters of TOEFL independent essays may either explicitly or implicitly (see Eckes, 2008) attend to these rubric descriptors and give higher scores to essays that contain lexical features such as lower word range, higher bigram frequency, higher hypernymy, and lower imageability.

### *4.2. Integrated essay score*

The relationship between lexical sophistication and integrated writing proficiency scores was much weaker than with independent essays. Overall, correlations between the indices of lexical sophistication investigated and integrated writing proficiency scores were low. Accordingly, the model that included two predictors was able to explain only 8.3% of the variance in integrated essay scores. These indices included *hypernymy (nouns)* and *Kucera-Francis number of categories AW*.

The mean number of hypernymic levels for nouns index (*hypernymy [nouns]*) explained 6.7% of the variance in integrated essay scores. Essays that on average included nouns that were more specific (i.e., had more superordinate terms) tended to earn higher scores. The range of registers index (*Kucera-Francis number of categories AW*) added 1.4% to the variance in integrated essay scores explained by the model indicating that integrated included words that occur in fewer registers tended to earn higher scores. These results suggest that the use of specific lexis is an indicator of quality in responses to integrated essay tasks in much the same way that it was predictive of writing quality for the independent tasks.

These results may also highlight an important distinction between receptive and productive vocabulary knowledge. Baba (2009), for example, indicated that vocabulary knowledge (as demonstrated on receptive vocabulary tests of breadth and depth) was moderately correlated ($r = .400$ and $r = .340$, respectively) with scores on a summary writing task. From a productive perspective, as evidenced in this study, however, the strongest correlation between lexical sophistication and integrated writing scores was small ($r = .263$), suggesting that for integrated tasks (which generally involve summary writing), receptive lexical knowledge may be a greater boon than productive lexical knowledge. That is to say that comprehending the reading (and/or listening passage) may be more advantageous than being able to employ sophisticated lexis in writing.

These findings, like the findings for the independent essays, also align well with the expectations found in the TOEFL integrated writing rubric, suggesting that lexical sophistication is not a particularly important factor in TOEFL integrated essays. The integrated scoring rubric generally includes descriptors that focus on the accurate summarization and synthesis of the content included in the reading and listening passage. The rubric contains few references to language use (and none related directly to lexical sophistication) and those that are contained in the rubric focus on language errors that impede accurate summarization and synthesis.

### *4.3. Differences between responses to independent and source-based tasks*

The results of the MANOVA and the DFA provide additional evidence that responses to independent and integrated tasks differ with regard to the occurrence of lexical features related to lexical sophistication. These differences were such that responses could be categorized by task with an accuracy of 94.3%. In general, these results show that independent responses tend to include lexical items that occur widely throughout different texts and context (i.e., are less specific), while integrated responses tend to include lexical items that occurred in a more restricted range of texts and contexts as reflected in the range indices (see Table 6). Responses to independent tasks also contained lexical items that were more sophisticated with regard to bigram frequency and were less imageable and had fewer superordinate terms (i.e., were more abstract) than responses to integrated tasks. Overall, this analysis demonstrated that responses to integrated tasks were more sophisticated with regard to the document and register-level range indices and included lexical items that were more imageable and contained more superordinate terms (i.e., were more specific). All indices demonstrated large effects for the differences between independent and integrated essays.

These results felicitously align with the nature of the two tasks. The independent tasks used in this study ask test takers to discuss their opinions on cooperation and the subjects they would like to study, respectively. The integrated tasks, on the other hand, discuss fish farming and bird migration, respectively. Knowing that test takers tend to repeat specific lexical items from listening and reading passages in their responses to integrated tasks (Crossley, Clevinger, & Kim, 2014), it seems likely that more specific and imageable terms related to fish farming and word migration (i.e., *fish, farm,* and *bird*) will be included in integrated responses. Such responses will also be limited in the range of the words used because the assignment limits the potential topic. In contrast, test takers responding to independent tasks about *opinions* and *school subject* will have much lower imageability and hypernymy scores and will not be limited in the range of words they can use to describe their opinions and experiences.

## 5. Conclusion

Overall, these results suggest that newly developed indices of lexical sophistication are important indicators of writing proficiency with regard to independent tasks. Additionally, the results suggest that indices of lexical sophistication are generally less important indicators of writing proficiency in source-based tasks. The results also suggest that independent and source-based tasks require writers to draw on different linguistic resources related to lexical sophistication.

The results from this study supports the TOEFL validity argument (i.e., the inclusion of both independent and integrated writing tasks; Chapelle et al., 2008) in that, at least with respect to lexical sophistication, the independent and integrated tasks seem to evaluate different language skills (cf. Cumming et al., 2005; Guo et al., 2013). The independent tasks clearly evaluate test takers' productive lexical knowledge. Higher scoring independent essays tend to include words that occur in fewer contexts, are more specific and less imageable, and two-word combinations that are more frequent in reference corpora than lower scoring independent essays. These findings suggest that raters are attending to rubric descriptors related to lexical sophistication. Responses to integrated tasks, on the other hand, seem to evaluate productive lexical knowledge to a much lesser degree, which is also reflected in the rating rubric.

This study also has important implications for automatic essay scoring systems and automatic essay feedback systems. The majority of previous systems have relied on simple word frequency and word length indices to assess lexical sophistication (e.g., Attali & Burstein, 2006; Enright & Quinlan, 2010). The findings from this study indicate that lexical features such as range and bigram information may be better predictors of holistic scores of writing proficiency than simple word frequency. Adding indices related to range and bigram frequency may allow AES systems to more accurately assess the construct of lexical sophistication, and may result in gains in construct coverage and scoring accuracy.

In addition, these findings have implications for L2 writing instruction. For instance, the results related to range indices suggest that L2 learners should be exposed to words within a wide variety of settings, domains, and genres so that they have access to context-general and context-specific lexical items. Furthermore, the results related to bigram frequency suggest

that vocabulary learning may benefit from contextual approaches that look not only at words in isolation but also frequent word neighbors (i.e., collocations) and verbal constructions (e.g., Ellis, O'Donnell, & Römer, 2013; O'Donnell, Römer, & Ellis, 2013). One limitation in our study is the manner in which we dealt with prompt differences (e.g., Crossley et al., 2011; Hinkel, 2002). We opted to remove any indices that demonstrated significant differences between prompts, which eliminated a large number of indices from our analysis. This conservative approach limited the number of indices we could sample, but also ensured that reported model strength was not moderated by prompt difference. In addition, our exploration of writing proficiency assessments only included a limited sampling of writing tasks included in the TOEFL. However, the results from this sample suggest that the construct of writing proficiency is indeed complex (e.g., Schoonen et al., 2011) and likely consists of a number of writing proficiencies. Future research should continue to explore the bounds of these proficiencies to determine which features of writing proficiency are robust and generalizable across writing tasks, and which are domain specific.

## References


Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning, and Assessment 4*(3) . http://ejournals.bc.edu/ojs/index.php/jtla/index.

The British National Corpus, version 3 (BNC XML Edition). (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Retrieved from http://www.natcorp.ox.ac.uk/.

Baba, K. (2009). Aspects of lexical proficiency in writing summaries in a foreign language. *Journal of Second Language Writing, 18*, 191–208.

Bereiter, C. (2003). Foreword. In M. D. Shermis, & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. vii–ix).Mahwah, NJ: Lawrence Erlbaum.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at.: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*, 371–405. http://dx.doi.org/10.1093/applin/25.3.371.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media, Inc..

Brown, G. D. A. (1984). A frequency count of 190 000 words in the London-Lund corpus of English conversation. *Behavior Research Methods Instruments & Computers*, *16*, 502–532. http://dx.doi.org/10.3758/BF03200836.

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*, 977–990. http://dx.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. http://dx.doi.org/10.3758/s13428-013-0403-5.

Burstein, J., Marcu, D., & Knight, K. (2003). Finding the WRITE stuff: Automatic identification of discourse structure in student essays. *Intelligent Systems, IEEE*, *18*(1), 32–39.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). Test score interpretation and use. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language™* (pp. 1–25).New York, NY: Routledge.

Chodorow, M., & Burstein, J. (2004). Beyond essay length: Evaluating e-rater®'s performance on TOEFL® essays. *ETS Research Report Series*, *2004*(1) [i-38. 10.1002/j.2333-8504.2004.tb01931.x].

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology Section A*, *33*(4), 497–505. http://dx.doi.org/10.1080/14640748108400805.

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing*, *18*(1), 100–108.

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

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

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, *35*, 115–135. http://dx.doi.org/10.1111/j.1467-9817.2010.01449.x.

Crossley, S. A., & McNamara, D. S. (2013). Applications of text analysis tools for spoken response grading. *Language Learning & Technology 17*(2), 171–192. http://llt.msu.edu/.

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

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning*, *60*, 573–605. http://dx.doi.org/10.1111/j.1467-9922.2010.00568.x.

Crossley, S. A., Weston, J. L., McLain Sullivan, S. T., & McNamara, D. S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication*, *28*, 282–311. http://dx.doi.org/10.1177/0741088311410188.

Crossley, S. A., Cai, Z., & McNamara, D. S. (2012). Syntagmatic, paradigmatic, and automatic 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).Menlo Park, CA: The AAAI Press.

Crossley, S. A., Cobb, T., & McNamara, D. S. (2013). Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications. *System.*, *41*, 965–981.

Crossley, S. A., 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*, 250–270.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2015). To aggregate or not? Linguistic features in automatic essay scoring and feedback systems. *Journal of Writing Assessment 8*(1) . http://www.journalofwritingassessment.org/index.php.

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.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, *7*, 31–51. http://dx.doi.org/10.1177/026553229000700104.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*(1), 7–24.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment*, *5*(1) .

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*, 155–185.

Ellis, N. C., O'Donnell, M. B., & Römer, U. (2013). Usage-based language: Investigating the latent structures that underpin acquisition [Issue Supplement]. *Language Learning*, *63*(s1), 25–51. http://dx.doi.org/10.1111/j.1467-9922.2012.00736.x.

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

Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing*, *27*, 317–334.

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

Ferris, D. R. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly*, *28*, 414–420. http://dx.doi.org/10.2307/3587446.

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K. (2013). Implementation and applications of the Intelligent Essay Assessor. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 68–88).New York, NY: Routledge.
Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, *9*, 123–145.
Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, *13*, 403–437. http://dx.doi.org/10.1075/ijcl.13.4.02gri.
Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, *18*(3), 218–238.
Halliday, M. A. K. (1991). Corpus studies and probabilistic grammar. In K. Aijmer, & B. Altenberg (Eds.), *English corpus linguistics* (pp. 30–43).New York, NY: Longman.
Hardy, J. A., & Römer, U. (2013). Revealing disciplinary variation in student writing: A multi-dimensional analysis of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora*, *8*, 183–207.
Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? *College English*, *63*, 480–499.
Higgins, D., Xi, X., Zechner, K., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech and Language*, *25*, 282–306.
Hinkel, E. (2002). *Second language writers' text: Linguistic and rhetorical features*. Mahwah, NJ: Lawrence Erlbaum.
Hyland, K. (2007). Genre pedagogy: Language, literacy and L2 writing instruction. *Journal of Second Language Writing*, *16*, 148–164.
Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, *19*, 57–84. http://dx.doi.org/10.1191/0265532202lt220oa.
Kellogg, R. T., & Raulerson, B. A. (2007). Improving the writing skills of college students. *Psychonomic Bulletin & Review*, *14*, 237–242. http://dx.doi.org/10.3758/BF03194058.
Koda, K. (1993). Task-induced variability in FL composition: Language-specific perspectives. *Foreign Language Annals*, *26*, 332–346.
Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.
Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990.
Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings and application. *TESOL Quarterly*, *49*, 757–786.
Kyle, K. (2016). Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication (Doctoral dissertation). Retrieved from http://scholarworks.gsu.edu/alesl_diss/35.
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*. http://dx.doi.org/10.2307/2529310.
Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*, 307–322. http://dx.doi.org/10.1093/applin/16.3.307.
Laufer, B. (1994). The lexical profile of second language writing: Does it change over time? *RELC Journal*, *25*(2), 21–33. http://dx.doi.org/10.1177/003368829402500202.
Leki, I., & Carson, J. G. (1994). Students' perceptions of EAP writing instruction and writing needs across the disciplines. *TESOL Quarterly*, *28*, 81–101.
Malvern, D. D., & Richards, B. J. (1997). A new measure of lexical diversity. *British Studies in Applied Linguistics*, *12*, 58–71.
Manchón, R. M., Murphy, L., & Roca de Larios, J. (2007). Lexical retrieval processes and strategies in second language writing: A synthesis of empirical research. *International Journal of English Studies*, *7*(2), 149–174.
Morris, L., & Cobb, T. (2004). Vocabulary profiles as predictors of the academic performance of Teaching English as a Second Language trainees. *System*, *32*, 75–87.
National Commission on Writing for America's Families, Schools, and Colleges (2003). The neglected R: The need for a writing revolution. New York, NY : College Entrance Examination Board.
O'Donnell, M. B., Römer, U., & Ellis, N. C. (2013). The development of formulaic sequences in first and second language writing: Investigating effects of frequency, association, and native norm. *International Journal of Corpus Linguistics*, *18*, 83–108. http://dx.doi.org/10.1075/ijcl.18.1.07odo.
Plakans, L., & Gebril, A. (2013). Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing*, *22*, 217–230.
Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing*, *13*(2), 111–129.
Römer, U., & O'Donnell, M. B. (2011). From student hard drive to web corpus (part 1): The design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora*, *6*, 159–177.
Römer, U. (2009). The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics*, *7*, 140–162.
Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. Kunnan (Ed.), *Validation in language assessment* (pp. 41–60).Mahwah, NJ: Lawrence Erlbaum.
Reynolds, D. W. (1995). Repetition in non-native speaker writing. *Studies in Second Language Acquisition*, *17*, 185–209.
Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language testing*, *18*, 55–88.
Schoonen, R., van Gelderen, A., Stoel, R. D., Hulstijn, J., & de Glopper, K. (2011). Modeling the development of L1 and EFL writing proficiency of secondary school students. *Language Learning*, *61*, 31–79.
Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. New York, NY: Routledge.
Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 313–346).New York, NY: Routledge.
Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, *31*, 487–512. http://dx.doi.org/10.1093/applin/amp058.
Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.
Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*, 4th ed. Needham Heights, MA: Allyn & Bacon.
Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York, NY: Teachers College Columbia University.
Toglia, M. P., & Battig, W. R. (1978). *Handbook of semantic word norms*. Hillsdale, NJ: Lawrence Erlbaum.
Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing*, *27*, 335–353.
Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication*, *3*(2), 215–229.

**Kristopher Kyle** Kristopher Kyle is an Assistant Professor in department of Second Language Studies at the University of Hawai'i. His research interests include second language writing and speaking, assessment, and second language acquisition. He is especially interested in applying natural language processing (NLP) and corpora to the exploration of these areas.

**Scott Crossley** Scott Crossley is an Associate Professor at Georgia State University. His interests include computational linguistics, corpus linguistics, cognitive science, discourse processing, and discourse analysis. His primary research focuses on the development and application of computational tools in second language learning and text comprehensibility.