Modern English Education

http://dx.doi.org/10.18095/meeso.2019.20.4.31

The Prediction of Writing Scores Using Vocabulary Features in ESL University Students' Essays

Yongkook Won
Seoul National University

ARTICLE INFO

Received: 17 September 2019
Revised: 20 October 2019
Accepted: 8 November 2019

Examples in: English
Applicable Languages: English
Applicable Levels: Secondary/Tertiary

KEYWORDS

vocabulary/
writing/
regression analysis/
어휘/
쓰기/
회귀분석

ABSTRACT

Won, Yongkook. (2019). The prediction of writing scores using vocabulary features in ESL university students' essays. *Modern English Education, 20*(4), 31-40.

The primary purpose of this study is to investigate how well vocabulary features, such as collocation, word family, etc., predict vocabulary scores in writing graded by human raters, and the degree to which they explain holistic scores. Forty-nine writings by first-year non-native English-speaking international students at a mid-western U.S. university were analyzed by using two multiple regression models. Data includes both holistic and analytic vocabulary scores given by four human raters and specific vocabulary features counted by concordance programs. Holistic and analytic vocabulary scores were estimated using many-facet Rasch measurement analysis. It was found that (a) collocation variety and essay length (or tokens) explained 32.0% of variance in the holistic writing scores, and (b) collocation variety and word family accounted for 29.5% of the variance in the vocabulary scores. The results of this study suggest that it is beneficial to teach and learn vocabulary knowledge to improve students' writing proficiency.

I. INTRODUCTION

Vocabulary plays a great role in communication (Bonk, 2001; Breiner-Sanders, Pardee Lowe, Miles, & Swender, 2000; Gitsaki, 1999; Grabe, 1991; Iwashita, Brown, McNamara, & O'Hagan, 2008; Nation & Meara, 2002) and vocabulary knowledge has been regarded as one of the most important components in language competence (Daller & Xue, 2007). The meaningful relationship between vocabulary knowledge and language proficiency is also acknowledged (Grabe, 1991; Nation & Meara, 2002). Several studies have attempted to explain how vocabulary use affects the quality of second language writing (Bonk, 2001; Hsu, 2007; Hsu & Chiu, 2008). However, the studies investigating the relationship between vocabulary knowledge and writing ability generally explain the importance of vocabulary knowledge for writing by presenting correlation indices (Breiner-Sanders et al., 2000; Iwashita et al., 2008). The limitation of these studies is that they have not provided enough information on what elements are involved in determining vocabulary knowledge nor how they are linked to writing ability. Some current studies (Crossley, Salsbury, & McNamara, 2014; McNamara, Crossley, & McCarthy, 2010; McNamara, Graesser, McCarthy, & Cai, 2014) based on computational language processing, human evaluation or both, have started to investigate what linguistic features explain the writing quality of students' essays.

Although it is not possible to predict writing ability based solely on the vocabulary features used in writing, it is still useful to understand the information about the degree of predictability of vocabulary use on writing quality. Because this information can guide teachers in understanding the importance of vocabulary teaching, as well as help them decide what facets of vocabulary should

Yongkook Won (Visiting researcher)
The Center for Educational Research, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, 08826, Korea
Tel: (02) 880-8834 / Email: linguistry@gmail.com

be emphasized in teaching, the role of specific vocabulary features in determining the assessed quality of writing should be studied. This study, therefore, aims to address the shortcomings of correlation-based studies by examining the links between vocabulary knowledge and writing ability and ascertaining how vocabulary features can predict writing scores by human raters.

II. LITERATURE REVIEW

1. Vocabulary Knowledge

Vocabulary knowledge includes many facets of vocabulary. Hartmann and James (1998) noted that vocabulary is the "sum total of the words used in a language, by a speaker, or for dictionary-making" (p. 154). When it comes to the analysis of vocabulary knowledge, however, the sheer number of words a speaker knows does not fully represent all contextual traits associated with vocabulary. Multidimensional aspects of vocabulary, such as form, meaning, collocation, grammatical functions, etc., should be included in the determination of vocabulary knowledge (Folse, 2004; Graves, 2006; Richards, 1976). These aspects are generally categorized into two broad concepts: *breadth* and *depth* of vocabulary knowledge (Nation, 2001; Vermeer, 2001).

The *breadth* of vocabulary knowledge refers to the size of vocabulary one knows. In evaluating the *breadth* of vocabulary knowledge, all words used should be counted. As there are several ways to count words, however, it is also necessary to establish a standard as to how to measure vocabulary size. One way to measure vocabulary size is to count every word form, or *token*, in a sentence. For example, the sentence "all is well that ends well" has six tokens, but there are only five different words, called *types*, because the word "well" appears twice. The tokens and types can be used in calculating the lexical variation of a text by dividing the number of types with the number of different tokens (type-token ratio). Other forms of words are termed *lemmas*, which consist of a headword and its inflected forms. For instance, the words *say*, *says*, *said* are counted as one lemma, because they are comprised of the headword *say* and its inflected forms *says* and *said.* A *word family* includes a lemma and its closely related derived forms. Using word families as indicators of vocabulary richness is useful, because they cover almost the same categories of words that learners encounter in texts (Laufer & Nation, 1995). Vocabulary learning targets could also be used as the evaluation criteria in determining one's vocabulary knowledge. It is recommended that at least 3,000 word families should be acquired for what could be considered a basic vocabulary target and 5,000 word families for specialized purposes, and most native speakers of English use around 2,000 word families in their daily lives (Thornbury, 2002). The 2,000 most frequent word families in English cover more than 80% of the running words in both spoken and written texts, and the 5,000 word families covers around 87% of text in the Brown Corpus of more than one million tokens (Nation, 2006; Read, 2004). These frequency based vocabulary lists can be used to measure the level of vocabulary knowledge because usually high frequency words are shorter (Nation, 2005; Zipf, 1949) and easier to acquire. In addition, the age-of-acquisition for content words (Gilhooly & Logie, 1980) can be a good index to measure the difficulty level of words because the age-of-acquisition index is developed based on human ratings on a scale ranged from 1 (age 0-2 years) to 7 (age 13 years and older) with two-year age bands. The words with higher age-of-acquisition score mean that children are believed to learn them when they become older.

The *depth* of vocabulary knowledge means how well and how much one knows vocabulary. For measuring *depth*, *collocations*, or groups of words that frequently co-occur in a natural context (Lewis, 1997; Stubbs, 2001), can be used as units to count. Based on expert native speaker intuition, Benson, Benson and Ilson (1997) categorized collocations into two major groups: *grammatical collocations* with 26 categories and *lexical collocations* with seven categories. This categorization has been used as the criteria for other collocation studies (Bonk, 2001; Gitsaki, 1999) and is used for this study. As a corresponding concept to *word family* in the *breadth* of vocabulary knowledge, *collocation variety*, which refers to the number of non-overlapping collocations, was created for this study. For example, *spend some more time, spend time,* and *time to spend* were counted as one *collocation variety* because they are all derived from the same basic expression *spend time*. N-grams are also used in this study, but they are defined differently from collocations in that n-grams do not consider any grammatical or lexical meaning while collocations do. As is often the case with expanding vocabulary size, it is also recommendable to acquire a huge amount of collocations, which in turn help increase the depth of vocabulary knowledge. Formulaic sequences of words, or collocations, shorten or save the processing resources in the brain because they are automatically recognized as one unit (Boers & Lindstromberg, 2008; Granger, 2011). In addition, collocation use increases socio-interactional functions, such as greeting or thanking, because both speakers (or writers) and listeners (or readers) feel comfortable with the known or shared expressions (Wray, 2000).

2. Previous Studies of Vocabulary Knowledge and Writing Ability

Several studies have investigated the correlations between vocabulary knowledge and writing proficiency of English language learners. Some of these studies verified the positive relationship between the breadth of vocabulary knowledge and writing ability (Breiner-Sanders et al., 2000; Iwashita et al., 2008; Stæhr, 2008), and others presented positive correlations between the depth of vocabulary knowledge and writing proficiency (Hsu, 2007). In these studies, vocabulary knowledge was usually mea-

sured by using vocabulary tests, such as the Vocabulary Size Test (Nation & Beglar, 2007), the Productive Vocabulary Levels Test (Laufer & Nation, 1999), and the Word Associates Test (Read, 1998). These vocabulary knowledge tests are learner-implemented synchronous tests that require test-takers to answer the vocabulary questions. Some other studies have measured vocabulary knowledge by using test-takers' produced discourse as sources of lexical richness (Hsu, 2007; Laufer & Nation, 1995).

Overall, although the studies seemed to suggest that there were meaningful relationships between vocabulary knowledge and writing ability (Hsu, 2007; Hsu & Chiu, 2008; T. Hyun, 2007; McNamara et al., 2010; McNamara, Crossley, Roscoe, Allen, & Dai, 2015; Nation & Meara, 2002; Y. Ryoo, 2017), many of them focused on the relationship between the holistic score of writing and one or two separate facets of their vocabulary knowledge. Some other studies (Ruegg, Fritz, & Holland, 2011) focused on the relationship between vocabulary features and vocabulary scores. Even though many facets of vocabulary have been compared with writing quality and vocabulary scores, these studies with separate vocabulary facets do not guarantee that the vocabulary features used for the evaluation of writing quality reflect both human rater's perceptions of writing quality and vocabulary knowledge. To overcome these weaknesses of the previous studies, this study not only investigates the predictability of vocabulary features for writing quality and vocabulary scores, but also discusses whether the vocabulary features used to predict the writing quality also play the same role in predicting vocabulary scores.

The present study investigates how well the indices of vocabulary use, such as type-token ratio, tokens, word families, n-grams, and collocations, explain holistic writing scores and vocabulary scores graded by human raters. Even though vocabulary knowledge includes different sets of information, such as form, meaning, collocation, connotation, register, cultural accretion, set phrases, grammatical functions, etc., this study focuses on the basic features of vocabulary—*breadth* and *depth*—which teachers can easily apply to their teaching. Research questions for the current study are:

1) What are the most influential vocabulary features (e.g., word family, collocation), for predicting holistic and vocabulary scores in non-native English-speaking students' academic writing?
2) How do different vocabulary features predict a holistic writing score in non-native English-speaking students' academic writing?
3) How do different vocabulary features predict a vocabulary use score in non-native English-speaking students' academic writing?

III. METHODS

1. Data

The 49 essay samples used in this study were derived from the English placement test (EPT), a test administered as an instructional placement test for undergraduate students entering a mid-western U.S. university. The EPT is intended to assess non-native English-speaking students' English proficiency and to assign them to an appropriate English program accordingly. The test contains sections assessing listening, reading, and writing skills, but the current study analyzed only the writing samples. Test-takers were given 30 minutes to prepare and write an argumentative academic essay in response to a given prompt (see Appendix 1). The test-takers' essay lengths ranged between 141 and 464 words (M = 306.67, SD = 72.68). All of the test-takers who took the EPT met the university English requirement for admission, having achieved a score of 71 or above on an internet-based Test of English as a Foreign Language (TOEFL iBT®) with minimum score of 17 in the writing section or an overall band score of 6.0 or above in International English Language Testing System (IELTS™) with 5.5 or above in the writing section.

2. Raters

Four raters (n = 4) were recruited and re-graded 49 essays with vocabulary (see Appendix 2) and holistic rating scales (see Appendix 3). All four raters were doctoral students who studied applied linguistics at a U.S. university. They all had taken several language assessment courses and were trained as EPT writing raters at a U.S. university. They were also English as a Second language (ESL) instructors who had taught ESL writing courses for more than one year in a tertiary education institution in the U.S.

3. Rating Scales

The holistic score of the essay was graded using a scale from 1-10 which was modified from the independent writing section of TOEFL iBT® writing rubrics (ETS, 2004); this particular scale was selected, because the writing section on the TOEFL is almost identical to the EPT writing test. The analytic scores were graded on a scale from 1-8 which was modified from Jacobs, Zinkgraf, Wormuth, Hartfiel and Hughey's (1981) ESL composition scoring profile, because this particular scoring profile is applicable to general ESL writing. Each level of the writing rubrics in both holistic and analytic scores was divided into low and high sub-levels to catch the ratings that could be assigned between two adjacent scores in the original rubrics.

4. Procedures

This study was conducted using the following procedures. First, the writing samples were scored with both

a holistic score and a vocabulary score by the raters. For rater training, five sample essays were used to calibrate rating scales among raters. The analytic scoring was conducted at least three days after the holistic scoring to reduce any errors related to the effect of holistic scoring on the second scoring. The current study employed a mixed rating design by having some raters' scorings overlap with other raters' scorings, allowing less scoring burden to raters while keeping enough essays. Figure 1 further illustrates the rating design in the current study. Each mark (x) indicates the score given by each rater. The final score was estimated with the *fair* score in the many-facet Rasch measurement (MFRM) model using the computer program FACETS version 3.80 (Linacre, 2014).

Rater	Essay ID 1-19 (*n* = 19)	Essay ID 20-24 (*n* = 5)	Essay ID 25-30 (*n* = 6)	Essay ID 31-49 (*n* = 19)
	1 2 … 18 19	20 21 22 23 24	25 26 … 29 30	31 32 … 48 49
R1	x x … x x	x x x x x		
R2		x x x x x	x x … x x	x x … x x
R3		x x x x x	x x … x x	x x … x x
R4	x x … x x	x x x x x	x x … x x	

FIGURE 1 Illustration of Rating Design

Second, the scales to measure the features of productive vocabulary use were created. The breadth of vocabulary knowledge was investigated using the number of words used (tokens), the number of different words (types), the relative proportions of types and tokens (or type-token ratio), and the number of word families. These features were counted by using Range program (Heatley, Nation, & Coxhead, 2002), which can process multiple text files and calculate the frequency of word types, tokens, and families. For measuring the depth of vocabulary knowledge, writing samples were further analyzed and checked against the eleven collocation types in Table 1, which were adapted from Gitsaki (1999) based on the *BBI Dictionary of English Word Combinations* (Benson et al., 1997). If a particular collocation in Table 1 was included in the essay, it was given one point for correct use. *Grammatical collocation* and *lexical collocation* were calculated simply by adding every count of collocations used in the text. Collocation variety, however, was calculated by adding only the different types of collocations in the text.

TABLE 1
Collocation Types of Lexical Collocations Used in the Study

	Collocation Type	Example
Grammatical collocations	Noun + Preposition	*argument about*
	Preposition + Noun	*in agony*
	Adjective + Preposition	*angry at*
Lexical collocations	Verb + Noun/Pronoun	*come to an agreement*
	Adjective + Noun (Noun Noun)	*strong tea*
	Noun + Verb	*alarms go off*
	Noun1 of Noun2	*a colony (swarm) of bees*
	Adverb + Adjective	*deeply absorbed*
	Verb + Adverb	*affect deeply*
	Miscellaneous	*in fact*
	Phrasal verb	*to pass on*

(Adapted from Gitsaki, 1999)

In addition, the N-Gram Phrase Extractor (Cobb, 2012) was used to analyze n-grams with three or four words. 3- to 4-gram words were included in the model to understand if repetition of long phrases was related to the vocabulary score or the holistic score. For the measurement of vocabulary difficulty, age-of-acquisition (WRDAOAc) scores of content words were measured by Coh-Metrix (McNamara et al., 2014) based on Gilhooly and Logie's (1980) 1,903 content words list. The WRDAOAc index was scored based on the concept that some words appear earlier than other words in children's language acquisition. Words with higher WRDAOAc index indicate that they are learned later by children (McNamara et al., 2014).

In addition to human raters' evaluation scores and the vocabulary use analysis, research questions were answered using correlation analyses and multiple regression analyses among given variables, such as vocabulary features, vocabulary scores, and holistic scores. While correlation analyses help in easily finding the possible variables involved in the overall picture of writing ability and vocabulary features, multiple regression analyses describe predictive or causal relationships between independent variables and dependent variables (Cohen, Cohen, West, & Aiken, 2003). The correlation analysis and the multiple regression analysis were conducted by utilizing the statistical packages IBM® SPSS® 22 (IBM Corp, 2013).

5. Data Analysis

The current study was based on the quantitative approach that emphasizes the frequency of the features of vocabulary used and the numeric scores of the essay. Correlational analysis and two multiple regression analyses were applied to determine how well the use of vocabulary in writing predicts holistic essay scores and vocabulary scores. First, holistic and vocabulary scores were separately estimated using MFRM rating scale analysis of the 49 test-takers by four raters. The MFRM rating scale analysis helped calculate *fair scores*, which are not biased by each rater's different rating severity, of the holistic and vocabulary scores of the test-takers. Second, several vocabulary features for independent (or predictor) variables (IVs) were selected based on a review of previous research (Breiner-Sanders et al., 2000; Gitsaki, 1999; Hsu, 2007; Iwashita et al., 2008; Stæhr, 2008), and their Pearson product moment correlations with holistic and analytic vocabulary scores, or dependent variables (DVs), were analyzed. Finally, two hierarchical multiple regression models that hypothesized relationships among IVs and DVs were tested. The IVs with strong correlations with each other (above .80) were reviewed, and only the most prominent one was used for the regression analysis to prevent multicollinearity among IVs. The cutoffs for multicollinearity were set at $r = .90$ for the correlation (Tabachnick & Fidell, 2013) and the variance inflation factor (VIF) values of around 1.00 (Field, 2009).

IV. RESULTS AND DISCUSSION

1. Selection of Independent Variables

Descriptive statistics for two DVs (holistic and vocabulary scores) and nine potential IVs [the number of word tokens, the number of word types, the number of word family, average age-of-acquisition for content words (WRDAOAc), the number of 4-word gram, and the number of 3-word gram, the number of grammatical collocations, the number of lexical collocations, and the types of collocations (collocation variety)] are reported in Table 2. Table 3 shows the correlations among IVs and DVs. Among the IVs selected for this study, the vocabulary features that had high Pearson correlations with DVs were considered as the final IVs for holistic and vocabulary regression models. Considering the correlations and multicollinearity issues among IVs, the best two variables from both the breadth of vocabulary (*word tokens* and *word family*) and the depth of vocabulary (*lexical collocations* and *collocation variety*) categories were selected as the IVs as shown in Table 3.

TABLE 2
Descriptive Statistics ($n = 49$)

	Variables	*Min*	*Max*	*M*	*SD*
Scores	Holistic score (1-10)	2.19	8.45	5.37	1.31
	Vocabulary score (1-6)	3.54	6.84	4.99	0.95
Breadth of vocabulary	Word tokens	141.00	464.00	306.67	72.68
	Word types	69.00	181.00	132.12	24.03
	Word family	66.00	162.00	119.92	20.82
	WRDAOAc (Mean x 100)	258.65	402.52	327.66	33.62
Depth of vocabulary	4-word gram	0.00	137.00	30.86	33.02
	3-word gram	0.00	199.00	76.12	51.60
	Grammatical collocation	0.00	15.00	2.06	2.50
	Lexical collocation	2.00	20.00	10.65	4.53
	Collocation variety	2.00	20.00	10.73	4.17

For the two IVs from the *breadth* of vocabulary, the Pearson correlations of the holistic score with *word tokens* ($r = .57$, $p < .01$), *word types* ($r = .55$, $p < .01$) and *word family* ($r = .54$, $p < .01$) were first compared, and *word tokens,* which had the highest correlation with the holistic score, was considered as the initial IV for the holistic scores. As the Pearson correlations of *word tokens* with *word types* ($r = .86$, $p < .01$) and *word family* ($r = .84$, $p < .01$) were quite high and that of *word types* and *word family* ($r = .98$, $p < .01$) was not acceptable ($r \geq .90$) (Tabachnick & Fidell, 2013), either *word types* or *word family* should be dropped. As the concept of *word tokens*, the initial IV, overlaps more with *word types* than with *word family*, *word types* was dropped from the final IVs. Even though the correlation of the vocabulary score with *word types* is higher than those with *word tokens* and *word family*, *word tokens* and *word family* were kept for the final IVs because *word types* can be conceptually explained by the combination of *word tokens* and *word family*. For the two IVs from the *depth* of vocaulary, the Pearson correlations of the holistic score with *lexical collocation* ($r = .35$, $p < .05$), *total collocation* ($r = .39$, $p < .01$) and *collocation variety* ($r = .45$, $p < .01$) were first compared. *Total collocation* was dropped from the fintal IVs because it was conceptually and statistically highly correlated with *collocation variety* ($r = .91$, $p < .01$) and could create multicollineanrity issues in the regression model (Berry, 1993).

TABLE 3
Inter-correlations for Vocabulary Features and Writing Scores

	Variables	Holistic Score	Vocabulary Score
Breadth of vocabulary	Word tokens #	.57**	.50**
	Word types	.55**	.57**
	Word family #	.54**	.56**
	WRDAOAc	.23	.13
Depth of vocabulary	4-gram	.27	.09
	3-gram	.33*	.18
	Grammatical collocation	.19	.21
	Lexical collocation #	.35*	.29*
	Total collocation	.39**	.35*
	Collocation variety #	.45**	.43**

Note. $^{*}p < .05$, $^{**}p < .01$. # indicates the selected IVs for regression analysis.

In the selection of vocabulary features, this study adopted a theory-based approach equipped with a regression analysis (shown in the following sections), which has the benefit of selecting predictor variables that are theoretically solid, thus making the interpretation of the final prediction model easier. This linear regression analysis approach, however, has the limitation of selecting multiple features to improve the accuracy of the regression model. When the predictor variables, or vocabulary features in this study, are highly correlated with each other, only a few best variables can be used not to produce biased results due to the multicollinearity issues. In contrast, the top-notch machine learning techniques, which have been widely adopted in the current language assessment setting (Shermis, Burstein, & Bursky, 2013), are known to use myriads of language features and have better accuracy of predicting final scores (Foltz, Streeter, Lochbaum, & Landauer, 2013). Even with the benefits of the advanced techniques in language assessment and research contexts, this machine scoring approach also has limitations of interpreting the prediction models (Shermis, Burstein, & Bursky, 2013). This so-called black box model does not reveal how the linguistic features predict the final scores (Elliot & Klobucar, 2013; Latour, 1999), thus making it difficult for teachers and learners to interpret the scores and to plan remedial teaching and learning programs. Thus, for teaching and learning purposes, a theory-based approach with a regression analysis still has more benefits.

2. Prediction of Holistic Writing Score With Productive Vocabulary Use

Even though *word tokens* had the highest correlation with the holistic score, *collocation variety* was first entered to the regression model because *word tokens*, or the length of essays, could suppress the effect of other predictor vari-

ables on the test scores in the given data. The hierarchical multiple regression revealed that in model one, *collocation variety* contributed significantly to the regression model, $F(1, 47) = 11.63$, $p < .01$, and accounted for 18.1% of the variance in the holistic scores as illustrated in Table 4. Adding the number of *tokens* to the regression model explained 13.9% additional variance ($\Delta R^2 = .320 - .181$) in the holistic scores, $F(2, 46) = 12.29$, $p < .01$. The addition of the remaining two variables, *word family* and *lexical collocation*, to the regression model did not show any statistically significant improvement of the model. In addition, because the remaining two variables conceptually overlapped with the *collocation variety* and *word tokens*, they were left out of the model.

The final model, model two, explained 32.0% of variance in the holistic writing scores, and the model shows that as the number of *collocation variety* increases by one standard deviation (4.17 collocation types), holistic scores increase by 0.17 standard deviations (0.17 x 1.31 = scores of 0.22), while holding the effect of the other variable constant. This result suggests that *collocation variety* and *word tokens* play a critical role in explaining the holistic score variance in that just these two vocabulary indices explained a third of the total writing score variance. The prediction of holistic scores with *collocation variety*, however, should be interpreted with caution because the effect of *collocation variety* on holistic scores is not statistically significant ($p = .25$). As the number of *tokens* increases by one standard deviation (72.68), holistic scores increase by 0.48 standard deviations (0.48 x 1.31 = scores of 0.63), while holding the effect of the other variable constant.

TABLE 4
Multiple Regression Analysis:
Vocabulary Features Predicting Holistic Scores

	Model	*R*	R^2	Adjusted R^2	*B*	*SE*	*β*	*t*	*p*
1	(Constant)	.44	.19	.18	3.86	.47		8.18	.00
	Collocation variety				.14	.04	.44	3.41	.00
2	(Constant)	.59	.34	.32	2.17	.68		3.20	.00
	Collocation variety				.05	.05	.17	1.17	.25
	Tokens				.01	.00	.48	3.25	.00

Note. *B* is unstandardized Beta; *SE* is standard error; *β* is standardized Beta.

The results indicate that test-takers need to increase the length of their essays by about 115 words (72.68 tokens / 0.63 scores = 115 words) to have one level higher scores when all other features of the writing are constant. In contrast, test-takers need to use about 19 different more collocations (4.17 collocations / 0.22 scores = 18.95 collocations) to have one level higher scores when all other features of the writing are constant. Considering the ceiling effect of the test-takers with higher scores, acquiring one level higher scores with only vocabulary improvement will not be easily attained. However, test-takers with shorter essays can still benefit from longer essays with more diverse collocation use in their writings.

3. Prediction of Vocabulary Scores with Productive Vocabulary Use

Another hierarchical regression analysis was conducted for the vocabulary score, or the dependent variable, and four predictor variables as shown in Table 3. As was done with the holistic scores regression model, *collocation variety* was first entered to the regression because the predictors in the breadth of vocabulary knowledge category, *word tokens* or *word family*, could suppress the effect of *collocation variety* on the dependent variable. The second predictor for the vocabulary score was *word family*, instead of *word tokens* or *types*, which had slightly higher correlations with the vocabulary score. In addition, *word family* is generally used as a unit to measure language learners' vocabulary size (Nation, 2006).

As shown in Table 5, the hierarchical multiple regression for vocabulary scores revealed that in model one, *collocation variety* contributed significantly to the regression model, $F(1, 47) = 10.59$, $p < .01$, and accounted for 16.7% of the variance in the vocabulary scores. Adding *the number of word family* to the regression model explained 12.8% additional variance ($\Delta R^2 = .295 - .167$) in the vocabulary scores, $F(2, 46) = 11.05$, $p < .01$. The addition of the remaining two variables, *word tokens* and *lexical collocation*, to the regression model did not show any statistically significant improvement of the model, and they were left out of the model.

TABLE 5
Multiple Regression Analysis:
Vocabulary Features Predicting Vocabulary Scores

	Model	*R*	R^2	Adjusted R^2	*B*	*SE*	*β*	*t*	*p*
1	(Constant)	.42	.18	.16	3.94	.34		11.46	.00
	Collocation variety				.09	.03	.42	3.25	.00
2	(Constant)	.57	.32	.29	2.05	.68		2.99	.00
	Collocation variety				.03	.03	.13	0.83	.41
	Word family				.02	.01	.48	3.09	.00

Note. *B* is unstandardized Beta; *SE* is standard error; *β* is standardized Beta.

In the final model, model two, *collocation variety* and *word family* accounted for 29.5% of the variance in vocabulary scores. This result suggests that *collocation variety* and *word family* are important factors in explaining the vocabulary score variance. The final model shows that as the number of *collocation variety* increases by one standard deviation (4.17), vocabulary scores increase by 0.13 standard deviations (0.13 x 0.95 = scores of 0.12), while holding the effect of the *word family* variable constant. Vocabulary scores increase by 0.48 standard deviations (0.48 x 0.95 = scores of 0.45) when the number of *word family* increases by one standard deviation (20.82), while holding the effect of the *collocation variety* variable constant. As the effect of *collocation variety* on vocabulary scores is not statistically significant ($p = .41$), the prediction of vocabulary scores with *collocation variety* should be interpreted with caution.

The results indicate that test-takers need to use more di-

verse vocabulary by about 46 different words (20.82 word family / 0.45 scores = 46.26 word family) to have one level higher vocabulary scores. When it comes to *vocabulary variety*, test-takers need to use about 35 more different collocations (4.17 collocations / 0.12 scores = 34.75) to have one level higher vocabulary scores. Considering the length of the essays, this absurdly high requirement of *collocation variety* only for one vocabulary score level improvement could be due to the fact that the contirbution of *collocation variety* overlapped with that of *word family* in the regression model. According to model one, where *collocation variety* was solely entered, test-takers only need to use about 10 more different collocations (4.17 collocations / 0.39 scores = 10.69) to have one level higher vocabulary scores. This difference indicates that the regression analyses should be interpreted with caution based on the related theories when the sample size is limited.

V. CONCLUSION

The purpose of this study was to measure the predictive value of productive vocabulary use on holistic writing scores and vocabulary scores in writing. This study has demonstrated that the variety of collocations, one of the vocabulary depth indices, in test-takers' writing can be one of the major variables that decide the proficiency level in terms of holistic and vocabulary scores. The findings are consistent with previous studies (Bonk, 2001; Gitsaki, 1999; Hsu, 2007; Lemmouh, 2008), showing that productive collocation use has influences on writing scores in general. It could be because collocations are likely to be less direct in their expressions (Grant & Bauer, 2004) and their usage more difficult to acquire than single word expressions, thus making raters perceive essays with higher *collocation variety* as being better writings (Hsu, 2007). Another explanation could be that test-takers' collocational knowledge might have reduced their cognitive load (Granger, 2011) and provided more time for test-takers to focus on the other aspects of writing (Nation, 2005).

This study has also demonstrated that the breath of vocabulary indices (i.e., *word tokens*, *word family*) were the most meaningful predictor of the writing scores. It has been known that human raters do not explicitly take essay length, or *word tokens*, into account as a strong indicator of essay quality, but good essays have high correlations with essay length (McNamara et al., 2015; Weigle, 2013). However, other studies showed negative correlations between essay length and writing scores (Ruegg et al., 2011). These contradictory findings may indicate that essay length is not a stand-alone predictor for essay scores. Essay topics, disciplines, and other extraneous variables should be taken into account for a more comprehensive understanding. For this study, as the data was derived from timed essay writing, the control of writing time could have affected raters' perception of test-takers' writing proficiency with different essay lengths.

Pedagogical implications could be drawn from this study regarding the importance of vocabulary (including collocational) knowledge in writing. Vocabulary knowledge has been known to play a critical role in second language development (Nation, 2001) and the current study confirms that writing teachers need to provide directions for students about how to acquire vocabulary knowledge. The study results also suggest that, in terms of collocation learning, the typical frequency based teaching and learning of vocabulary (Nation, 2005; Thornbury, 2002) should be carefully applied. Even though collocation, compared to non-collocational words, is typically less frequently used in writing, the results of this study suggests that it is still beneficial to teach collocations to improve students' writing proficiency.

While this study produced some significant and potentially important findings, there were some limitations. Even though this study investigated several vocabulary features and their relationship with holistic writing scores and vocabulary scores, the normalized collocation features did not show any statistically significant contribution to the test scores. This non-significant contribution might be related to the writing test format, which is timed essay writing, in that more proficient writers tend to write longer. As the essay length variables, such as *word tokens*, *types*, and *families*, had dominant influences on predictability of test scores, those of collocational use were limitedly noticeable. Studies with timed essay writing may not be suitable to comb through for any statistically significant effects of diverse vocabulary features, such as *collocation variety*, when the sample size is limited. In addition, the essay topic and the essay type in this study could have affected the findings. Argumentative essay writing typically requires different writing styles than other essay types (e.g., expository, compare and contrast, cause and effect), and different essay topics could have elicited different expressions. Another limitation of this study is related with the mismatch between the descriptors in the scoring rubrics and the predictor variables in this study. Although the scoring rubric requires word choice and appropriate register, the predictor variables were not arranged considering the context in which the vocabulary items occur. Lastly, there was no elicitation of specific collocations from students; their writing samples do not fully represent the participants' collocation knowledge. The results of collocation use could have been different if the students were asked to use specific collocations.

Further studies need to investigate how vocabulary features can be used to predict writing quality by analyzing larger and more diverse writing samples in terms of their genre, topic, and assessment types. In addition, studies with both linear and nonlinear regression models with more diverse vocabulary feaures extracted using computational textual assessment tools, such as Coh-Metrix (McNamara et al., 2014), need to be conducted to provide more multidimensional aspects of vocabulary use in writing.

REFERENCES

Benson, M., Benson, E., & Ilson, R. (1997). *The BBI dictionary of English word combinations.* Amsterdam: John Benjamins.

Berry, W. D. (1993). *Understanding regression assumption.* Newbury Park, CA: Sage.

Boers, F., & Lindstromberg, S. (2008). *Cognitive linguistic approaches to teaching vocabulary and phraseology* (Vol. 6). Berlin: Walter de Gruyter.

Bonk, W. J. (2001). Testing ESL learners' knowledge of collocations. In T. Hudson & J. D. Brown (Eds.), *A focus on language test development: Expanding the language proficiency construct across a variety of tests* (pp. 113-142). Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.

Breiner-Sanders, K. E., Pardee Lowe, J., Miles, J., & Swender, E. (2000). ACTFL proficiency guidelines-Speaking (revised 1999). *Foreign Language Annals, 33*(1), 13-18.

Cobb, T. (2012). N-gram phrase extractor [Computer software]. Retrieved from http://lextutor.ca/n_gram

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences.* New York: Routledge.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2014). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics, 35*(5), 1-22.

Daller, H., & Xue, H. (2007). Lexical richness and the oral proficiency of Chinese EFL students. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 150-164). Cambridge: Cambridge University Press.

Elliot, N., & Klobucar, A. (2013). Automated essay evaluation and the teaching of writing. In M. D. Shermis & J. Burnstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 16-35). New York: Routledge.

ETS. (2004). *iBT/Next generation TOEFL test: Independent writing rubrics.* Retrieved from https://www.ets.org/Media/Tests/TOEFL/pdf/Writing_Rubrics.pdf

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage Publications.

Folse, K. S. (2004). *Vocabulary myth: Applying second language research to classroom teaching.* Michigan: The University of Michigan Press.

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. Burnstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 68-88). New York: Routledge.

Gilhooly, K. J., & Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation, 12*(4), 395-427.

Gitsaki, C. (1999). *Second language lexical acquisition: A study of the development of collocational knowledge.* San Francisco: International Scholars Publications.

Grabe, W. (1991). Current developments in second language reading research. *TESOL Quarterly, 25*(3), 375-406.

Granger, S. (2011). From phraseology to pedagogy: Challenges and prospects. In T. Herbst, S. Fauhaber, & P. Uhrig (Eds.), *Chunks in the description of language: A tribute to John Sinclair* (pp. 123-146). Boston: Walter de Gruyter.

Grant, L., & Bauer, L. (2004). Criteria for re-defining idioms: Are we barking up the wrong tree? *Applied Linguistics, 25*(1), 38-61.

Graves, M. F. (2006). *The vocabulary book: Learning and instruction.* New York: Teachers College Press.

Hartmann, R. R. K., & James, G. (1998). *Dictionary of lexicography.* London: Routledge.

Heatley, A., Nation, P., & Coxhead, A. (2002). RANGE and FREQUENCY programs [Computer software]. Retrieved from http://www.vuw.ac.nz/lals/staff/Paul_Nation

Hsu, J. (2007). Lexical collocations and their relation to the online writing of Taiwanese college English majors and non-English majors. *Electronic Journal of Foreign Language Teaching, 4*(2), 192-209.

Hsu, J., & Chiu, C. (2008). Lexical collocations and their relation to speaking proficiency of college EFL Learners in Taiwan. *The Asian EFL Journal, 10*(1), 181-204.

Hyun, Taeduck. (2007). The effect of learning collocations on improving English proficiency. *Modern English Education, 8*(1), 191-209.

IBM Corp. (2013). IBM SPSS statistics for Macintosh (Version 22) [Computer software]. Armonk, NY: IBM Corp.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*(1), 24-49.

Jacobs, H., Zinkgraf, S., Wormuth, D., Hartfiel, V., & Hughey, J. (1981). *Testing ESL composition: A practical approach.* Rowley, MA: Newbury House.

Latour, B. (1999). *Pandora's hope: Essays on the reality of science studies.* Cambridge, MA: Harvard University Press.

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

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing, 16*(1), 33-51.

Lemmouh, Z. (2008). The relationship between grades and the lexical richness of student essays. *Nordic Journal of English Studies, 7*(3), 163-180.

Lewis, M. (1997). *Implementing the lexical approach:*

Putting theory into practice. Hove, England: Language Teaching Publications.
Linacre, J. M. (2014). *A user's guide to FACETS.* Retrieved from https://www.winsteps.com/a/Facets-ManualPDF.zip
McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication, 27*(1), 57-86.
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.
McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix.* Cambridge: Cambridge University Press.
Nation, P. (2001). *Learning vocabulary in another language.* Cambridge: Cambridge University Press.
Nation, P. (2005). Teaching and learning vocabulary. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 581-595). Mahwah, NJ: Lawrence Erlbaum.
Nation, P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63*(1), 59-82.
Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9-13.
Nation, P., & Meara, P. (2002). Vocabulary. In N. Schmitt (Ed.), *An introduction to applied linguistics* (pp. 35-54). London: Hodder Arnold.
Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. J. Kunnan (Ed.), *Validation in language assessment* (pp. 41-60). Mahwah, NJ: Lawrence Erlbaum.
Read, J. (2004). Research in teaching vocabulary. *Annual Review of Applied Linguistics, 24*, 146-161.
Richards, J. C. (1976). The role of vocabulary teaching. *TESOL Quarterly, 10*(1), 77-89.
Ruegg, R., Fritz, E., & Holland, J. (2011). Rater sensitivity to qualities of lexis in writing. *TESOL Quarterly, 45*(1), 63-80.
Ryoo, Young-sook. (2017). Predictability of the cloze test as a measure of written productive vocabulary. *Modern English Education, 18*(4), 25-45.
Shermis, M. D., Burstein, J., & Bursky, S. A. (2013). Introduction to automated essay evaluation. In M. D. Shermis & J. Burnstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 1-15). New York: Routledge.
Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal, 36*(2), 139-152.
Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics.* Oxford: Blackwell Publishers.
Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Pearson Education Limited.
Thornbury, S. (2002). *How to teach vocabulary.* Harlow, UK: Pearson Education.
Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics, 22*(2), 217-234.
Weigle, S. C. (2013). English as a second language writing and automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 36-54). New York: Routledge.
Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics, 21*(4), 463-489.
Zipf, G. K. (1949). *Human behavior and the principle of least effort.* Oxford: Addison-Wesley Press.

APPENDIX 1

Essay Writing Prompt for the Study

Modern conveniences such as fast food, automated teller machines, and labor-saving appliance promise to make life easier. Do these products and services actually make our lives more convenient or do they simply create new problems? Explain your position with reasons and examples from your own experiences, observations or reading.

APPENDIX 2

Analytic Scoring Rubric for Academic Writing

Level	Score	Description
Excellent	8	• sophisticated range • effective word/idiom choice and usage
Very good	7	• word form mastery • appropriate register
Good	6	• adequate range
Average	5	• occasional errors of word/idiom form, choice, usage but meaning not obscured
Fair	4	• limited range
Poor	3	• frequent errors of word/idiom form, choice, usage • meaning confused or obscured
Very poor	2	• essentially translation
Not gradable	1	• little knowledge of English vocabulary, idioms, word form

(Adapted from Jacobs et al., 1981)

APPENDIX 3

Holistic Scoring Rubric for Academic Writing

Level	Score	Description
5-High	10	**An essay at this level largely accomplishes all of the following:** • Effectively addresses the topic and task • Is well organized and well developed, using clearly appropriate explanations, exemplifications, and/or details
5-Low	9	• Displays unity, progression, and coherence • Displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice, and idiomaticity, though it may have minor lexical or grammatical errors
4-High	8	**An essay at this level largely accomplishes all of the following:** • Addresses the topic and task well, though some points may not be fully elaborated • Is generally well organized and well developed, using appropriate and sufficient explanations, exemplifications, and/or details
4-Low	7	• Displays unity, progression, and coherence, though it may contain occasional redundancy, digression, or unclear connections • Displays facility in the use of language, demonstrating syntactic variety and range of vocabulary, though it will probably have occasional noticeable minor errors in structure, word form, or us of idiomatic language that do not interfere with meaning
3-High	6	**An essay at this level is marked by one or more of the following:** • Addresses the topic and task using somewhat developed explanations, exemplifications, and/or details • Displays unity, progression, and coherence, though connection of ideas may be occasionally obscured
3-Low	5	• May demonstrate inconsistent facility in sentence formation and word choice that may result in lack of clarity and occasionally obscure meaning • May display accurate but limited range of syntactic structures and vocabulary
2-High	4	**An essay at this level may reveal one or more of the following weaknesses:** • Limited development in response to the topic and task • Inadequate organization or connection of ideas
2-Low	3	• Inappropriate or insufficient exemplifications, explanations, or details to support or illustrate generalizations in response to the task • A noticeable inappropriate choice of words or word forms • An accumulation of errors in sentence structure and/or usage
1-High	2	**An essay at this level is seriously flawed by one or more of the following weaknesses:** • Serious disorganization or underdevelopment
1-Low	1	• Little or no detail, or irrelevant specifics, or questionable responsiveness to the task • Serious and frequent errors in sentence structure or usage

(Adapted from ETS, 2004)