# Evaluating the Relationship Between VST Score and Lexical Diversity in Written Production Among Asian Learners of English

By Fredrik deBoer

Brooklyn College,

City University of New York

Abstract

Researchers in language learning divide a given learner's passive vocabulary, or words that can be defined on request, from his or her active vocabulary, or words that are utilized in the production of natural language. Active vocabulary is of greater value for communicative competence and of more interest to language testers, but is harder to adequately assess than passive vocabulary. In the present study, the timed essays of writers from China, Korea, and Japan were assessed for their active vocabulary, operationalized as lexical diversity, using three popular metrics for such assessment. These measures were correlated with each writer's score on a test of passive vocabulary, the VST, or vocabulary size test. Regression analysis was conducted for all three metrics as well. Across all lexical diversity metrics and language backgrounds, the correlation with VST score was low, suggesting that there is little direct connection between a writer's passive vocabulary and their active vocabulary as expressed in their writing. This suggests that such tests are of little use for predicting practical vocabulary use in writing and should not be utilized for that purpose.

Keywords: active vocabulary, passive vocabulary, second language writing, lexicon, corpus linguistics

Fredrik deBoer
Brooklyn College
1216 Boylan Hall
2900 Bedford Avenue
Brooklyn, NY 11210
Phone: (860) 336-9931
fredrik.deboer@brooklyn.cuny.edu

## 1. Introduction

### *1.1 Statement of the Problem and Literature Review*

Lexical command is an essential aspect of second language writing. Without a broad vocabulary, second language (L2) writers cannot easily access the words necessary to communicate effectively. In a broad literature review of then-contemporary second language writing research, Silva (1993) identified a lack of lexical resources as a source of difficulty for L2 writers and argues that instructors should work to develop the grammatical and lexical resources of L2 writers (p. 671). Likewise, Grabe and Caplan (1996) argued that development of vocabulary acts as a foundation of language learning, which in turn supports reading, writing, and overall syntactic competence (p. 275). Baba (2009) suggested that L2 writers often identify a lack of lexical resources as a key hindrance in their development as writers, which in turn erodes their confidence in writing in English (p. 192). Beglar (2010) argued that "because of the key role that lexical knowledge plays in reading and listening, it is important that estimates of receptive vocabulary size be available to administrators, teachers, and the learners themselves" (p. 1). The development of an L2 writer's vocabulary, it seems, may be a key aspect of second language writing instruction.

In order to effectively address deficiencies in lexical command, researchers and instructors must be able to effectively assess student lexical resources. Considerable disagreement exists, however, concerning the best methods for how to assess an L2 student's vocabulary. In part, this disagreement stems from a simple question: what constitutes knowing a word? The editors of *Modeling and Assessing Vocabulary Knowledge* (2007) argue in the

collection's introduction that the concept of "knowing" vocabulary depends a great deal on context and definition (p. 4). Scholars have frequently identified a theoretical difference between a language user's passive vocabulary and that user's active vocabulary. Generally, passive vocabulary refers to a language user's ability to correctly identify a particular term's definition when prompted by a test or examiner. Active vocabulary, in contrast, refers to a language user's actual integration of vocabulary into their practical language production. While passive vocabulary knowledge is clearly an important component of linguistic competence, the ability to integrate that knowledge into actual productive linguistic acts is widely believed to be of greater communicative value. Laufer and Paribakht (2002) argue that lexical knowledge depends on the ability of language users to utilize vocabulary in sentences and discourse, and thus measuring that knowledge through multiple-choice testing or similar methods may be ineffective (p. 366).

The assessment of passive vocabulary remains ubiquitous in language testing, however, for a simple reason: such tests are straightforward and practically feasible. Many extant tests, including popular tests commonly used to assess the language readiness of international students such as the Test of English as a Foreign Language (TOEFL) and the (IELTS), count vocabulary among their tested constructs. The measurement of active vocabulary, in contrast, is more practically and theoretically complex. Nation defines the necessary aspects of measuring vocabulary in use as "having the assessment of vocabulary as part of a larger construct such as the ability to read informative texts, taking account of the total vocabulary content of the language use material, and involving the user in having to take account of a range of contextual information" (p. 41). In other words, measuring active vocabulary, or vocabulary in use, requires that the assessed vocabulary be produced as part of a broader communicative act that is context-

dependent and situated. This is a high bar for tests to clear, particularly given the time constraints typical of practical language testing situations.

Given the practical and theoretical impediments to assessing active vocabulary through test instruments, it may be more fruitful to look for evidence of a broad vocabulary in actual student writing. Such an exploration would have the benefit of increasing the ecological validity of the assessment, as real-world student writing could be used for the task of vocabulary assessment. However, in order to effect such an assessment, it is necessary to understand the relationship between a given language user's passive vocabulary as measured in a typical vocabulary test, and that language user's active vocabulary as found in his or her written texts. Thus far, that relationship is largely unexplored. Koizumi and Yo In'nami (2012) argue that L2 research has thus far not paid adequate attention to the use of vocabulary in language production, despite the obvious importance of this subject (p. 554). Similarly, Skehan (2009) argues that measures of lexis are a vital aspect of adequately assessing language task performance (p. 512). This research is an effort to deepen our understanding in this area, in order to develop more effective tests of vocabulary use and language use generally.

In order to better understand the connection between active and passive vocabulary in written communication, it is necessary to compare a writer's results on a standardized test of vocabulary and the diversity of vocabulary displayed in that writer's actual written production. In other words, a sufficiently large data set could be analyzed for its diversity in vocabulary, and that result correlated with vocabulary test scores for the writers of individual texts. Kojima and Yamashita (2014) state that while attempts have been made in the past to extrapolate active vocabulary performance from the results on tests of passive vocabulary, they have thus far been unsuccessful (p. 24). This difficult likely stems in part from a traditional lack of adequately-sized

data sets and effective tools for assessing their lexical diversity. With contemporary computational linguistics applications and machine-readable corpora of considerable size, these limitations have been overcome. In the following study, a large collection of essays written by Asian learners of English is assessed for lexical diversity using three metrics. Lexical diversity is then compared to performance by students on a popular test of passive vocabulary, to investigate the relationship between the two.

There is a fairly broad range of extant research discussing passive and active vocabulary, but researchers do not agree on the precise definitions of these categories or the existence of subcategories between them. Laufer and Parikbhat (1998) state that a majority of researchers concerned with lexicon now view lexical knowledge as a continuum, rather than through binaries such as known/unknown or active/passive. However, different researchers have proposed differing approaches to understanding vocabulary as a continuum of knowledge (Goulden et al 1990, Meara 1996, Wesche and Paribakht 1996). Whatever the correct approach, researchers generally recognize the necessity of dividing lexical knowledge into divisions such as passive and active, given that most language testers, teachers, and users are more concerned with the vocabulary that a given speaker can use than the vocabulary he or she can define on command. In some research, terms such as "receptive and productive" are substituted. Meara (1990) argued that there is in fact a qualitative difference between the two categories, even beyond differences on a spectrum of understanding, and that studying both, and the relationship between the two, is therefore important. Laufer (1998) demonstrated empirically that active and passive vocabulary develop a different rates under formal instruction, again demonstrating the importance of empirically investigating the relationship between the two constructs.

*1.2 Research Questions*

The present study is designed to quantitatively evaluate the relationship between performance on a standardized vocabulary test and the amount of vocabulary used in essay responses to a given prompt. The research questions for this study are:

- Is there are consistent quantitative relationship between a given second language writer's score on a standardized test of vocabulary and that writer's displayed range of vocabulary in a written response to a standardized essay prompt?
- Are standardized multiple choice tests of vocabulary a valid means to predict a second language writer's ability to display a diverse range of vocabulary in an essay?

## 2. Materials and Methods

This research utilizes a quantitative, computerized approach. A corpus of student essays composed by L2 learners of English from Asian backgrounds was assessed for lexical diversity. Because the best method of assessing lexical diversity remains controversial, three separate measures were derived for each student text: HD-D, MTLD, and Maas Index. These figures were then correlated with each student's performance on the Vocabulary Size Test (VST), a popular test of vocabulary knowledge developed by Nation and Beglar. Correlations were generated for each of the three represented language groups, Chinese, Japanese, and Korean, and for all students combined. Regression analyses were conducted using a combined data set of all three language backgrounds to investigate the ability of VST to predict lexical diversity as operationalized by all three metrics.

*2.1 Lexical Diversity Metrics*

The terminology utilized in referring to measures that demonstrate diversity in displayed vocabulary is varied. Some researchers prefer to use the term "lexical richness," while others use that term to refer to a broader understanding of sophistication in vocabulary, including complexity and rarity of displayed words, correct or conventional usage of words, etc. Some researchers have used the term "lexical diversity" to refer to the more specific construct of the displayed range of vocabulary in a given text. This diversity in terms likely reflects the conceptual complexity of the construct. I follow Philip McCarthy and Scott Jarvis (2007) in defining lexical diversity as "the range and variety of vocabulary deployed in a text by either a speaker or a writer" (p. 459). In other words, my consideration here will include the numeric range of displayed words, but will not include the rarity, complexity, or usage of words.

While much work has been done in the past decade to develop better metrics of lexical diversity, to date no one metric can adequately capture the full range of the construct. Current best practices call for assessing lexical diversity with multiple measures, in order to cross-validate each other. For this research, each essay was assessed using multiple measures: hypergeometric distribution of diversity (HD-D), measure of textual lexical diversity (MTLD), and Maas Index (Maas). These three metrics each function in somewhat different ways and thus capture different types of lexical information. This combination was proposed by McCarthy and Jarvis as an appropriate method to assess lexical diversity (2010, p. 391).

HD-D measures the odds that any given unique token will appear in a randomly-drawn sample of a given length from an analyzed text. [REDACTED] MTLD is a sequentially-evaluated LD metric which evaluates changes to type-token ratio (TTR) over the length of a

given language sample that resets when a particular TTR threshold is reached, reducing the impact of text length on the metric. Maas Index is a log-corrected TTR function which has been demonstrated to be more robust to text length effects than other similar measures (Maas 1972). These metrics are highly but imperfectly correlated with each other (Jarvis & McCarthy 2010), indicating that they measure similar but not identical underlying constructs. All are more robust to text length effects than the traditional, now-deprecated metric TTR, but like all extant LD metrics, they are still subject to text length effects (Koizumi & In'nami 2012). All three metrics were generated using the Gramulator, a freely available computerized contrastive corpus analysis tool developed by Philip McCarthy of the University of Memphis.

*2.2 The VST*

The Vocabulary Size Test, or VST, is one of the most popular and most researched tests of vocabulary in use today. Initially developed by Nation and Beglar in the late 1990s, the test has been revised several times and exists in several different forms. The VST utilized in the ICNALE corpus consists of 200 questions designed to measure passive vocabulary knowledge, specifically of 20,000 of the English words most frequently encountered in the British National Corpus. Raw scores are multiplied by 200 for an estimate of vocabulary size; that is, a test taker who scores 50 out of 200 would be estimated to know 50 times 200 words, or 10,000 words. The VST utilizes a multiple-choice format, with all items on a given test having four possible choices (one answer and three distractors). The test is designed to measure receptive vocabulary knowledge – that is, vocabulary knowledge utilized in reading. The VST may therefore be seen as a poor fit for predicting productive vocabulary knowledge. However, the VST and tests like it are often used as generalized vocabulary tests. This research is intended to consider the validity

of inferences drawn in that case. It is not intended as a comment on the validity of the VST as a test of receptive vocabulary nor as a general inquiry into the quality of the test.

One strength of using the VST is that it has been used extensively in other research in language testing and applied linguistics (Schmitt et al. 2001; Cameron 2002; Laufer & Ravenhorst-Kolavski 2010; Qian 1999). Additionally, Beglar (2010) undertook a Rasch validation analysis of the VST. A Rasch analysis is a quantitative validation model that seeks to establish whether a test's ability to differentiate between several different test subjects works independent of which particular test items are used to test that variable. While some variability in this analysis is inevitable, a valid test will assign highly equivalent scores to two different test takers of equal ability even when they have taken different items designed to assess that variable. This alignment in different test items of the same variable is referred to as unidimensionality. Beglar's Rasch analysis of the VST found that the test did indeed test different items of the same variable consistently, with a strong degree of unidimensionality, supporting the validity of the VST and increasing our confidence in its application in this research.

*2.3 Data and Participants*

This research utilizes the International Corpus Network of Asian Learners of English (ICNALE), a freely accessible research corpus developed by Dr. Shinichiro Ishikawa of Kobe University in Japan. The ICNALE contains both a Written and Spoken portion. This research utilizes the ICNALE Written. The ICNALE Written corpus is comprised of 5600 essays by writers from ten countries. These essays are written under controlled conditions, with each writer composing on a word processor, without the aid of spell check, and responding to two prompts asking the writers to take a position on an issue of controversy. The ICNALE presents a number

of advantages. First is its range of language backgrounds and countries of origin, as well as its sheer size. Additionally, the controlled and consistent conditions under which the ICNALE essays were written and collected helps researchers make generalizable claims about the data set. Finally, and of most direct relevance to this research, the ICNALE contains a great deal of metadata about the writers who contribute texts, including the results of a vocabulary test that writers take as part of data collection. Ishikawa argues in ICNALE supporting materials that for L2 learners VST results are highly correlated with overall language proficiency. The ICNALE data collection process incorporates the VST through an Excel-based application of 50 test items drawn from a 1,000 to 5,000 word version of the VST Monolingual exam. VST scores range from 0-50, with one point scored for each word correctly used in a multiple choice item.

This research utilizes a subcorpus of 800 essays by Chinese L2 writers, 800 Japanese L2 writers, and 600 Korean L2 writers, for a final sample size of 2200 essays. These subpopulations were chosen because they represent the three most common language backgrounds for L2 learners in English-language universities that are represented in the ICNALE. English language samples were excluded because these writers do not take the VST as part of ICNALE data collection. Therefore, this research cannot meaningfully comment on the relationship between VST performance and lexical diversity in writing by L1 students. LD measures from the three tested indices were compared to performance on the VST for all essays. A correlation matrix was developed to examine the relationship between LD metrics and VST for each individual language population and for the data set as a whole. Simple linear regression analyses were performed to assess how predictive VST scores are for each LD measure. These correlation and regression measures were generated using the statistical programming language R, which has recently been proposed as a default method of conducting statistical analysis in applied

linguistics (Mizumoto 2015). Regression scatter plot graphics were generated in the statistical package JASP.

## 3. Results

Descriptive statistics for the overall data set are found in Table 1 below.

**Descriptive Statistics**

| | HD-D | MTLD | Maas | VST |
|---|---|---|---|---|
| **Valid** | 2200 | 2200 | 2200 | 2200 |
| **Missing** | 0 | 0 | 0 | 0 |
| **Mean** | -0.6962 | 68.08 | 99.41 | 31.07 |
| **Std. Deviation** | 1.473 | 18.77 | 16.84 | 8.417 |
| **Minimum** | -11.69 | 28.13 | 54.80 | 10.00 |
| **Maximum** | 3.421 | 166.2 | 274.7 | 50.00 |

Table 1. Descriptive statistics of lexical diversity and VST results from combined data set.

Correlation matrixes were generated to compare each language subgroup's performance on the VST to the lexical diversity of their essays. These matrixes are represented as Table 2, 3, and 4 below. Note that a lower score on Maas Index indicates greater diversity, and as such its correlations with the other two metrics are negative.

| | CHN HD-D | CHN MTLD | CHN Maas | CHN VST |
|---|---|---|---|---|
| CHN HD-D | 1 | | | |
| (*n* = 800) | | | | |
| CHN MTLD | .846** | 1 | | |
| (*n* = 800) | | | | |
| CHN Maas | -.845** | -.801** | 1 | |
| (*n* = 800) | | | | |
| CHN VST | .220** | .205** | -.200** | 1 |
| (*n* = 800) | | | | |

**. Correlation is significant at the 0.01 level (2-tailed).

Table 2. Lexical diversity and VST performance by Chinese L1 Students

| | KOR HD-D | KOR MTLD | KOR Maas | KOR VST |
|---|---|---|---|---|
| KOR HD-D | 1 | | | |
| (*n* = 600) | | | | |
| KOR MTLD | .847** | 1 | | |
| (*n* = 600) | | | | |
| KOR Maas | -.894** | -.761** | 1 | |
| (*n* = 600) | | | | |
| KOR VST | .195** | .189** | -.127** | 1 |
| (*n* = 600) | | | | |

**. Correlation is significant at the 0.01 level (2-tailed).

Table 3. Lexical diversity and VST performance by Korean L1 Students

| | JPN HD-D | JPN MTLD | JPN Maas | JPN VST |
|---|---|---|---|---|
| JPN HD-D | 1 | | | |
| (*n* = 800) | | | | |
| JPN MTLD | .788** | 1 | | |
| (*n* = 800) | | | | |
| JPN Maas | -.891** | -.724** | 1 | |
| (*n* = 800) | | | | |
| JPN VST | .045 | .046 | -.030 | 1 |
| (*n* = 800) | | | | |

**. Correlation is significant at the 0.01 level (2-tailed).

Table 4. Lexical diversity and VST performance Japanese L1 Students

Two trends can be seen in this data: one, the three lexical diversity metrics used in this research are highly but imperfectly correlated with each other, as expected; and two, none of the metrics are highly correlated with student performance on the Vocabulary Size Test, with none higher than the .220 r-value of Chinese student VST results and the HD-D metric. Many of these correlations are highly significant, but this is a function of the large sample size utilized in this study, not of a strong correlation. Japanese student results show a particularly low relationship. A correlation matrix comparing lexical diversity metrics and VST performance for all students combined is printed below as Table 5.

| | COM HD-D | COM MTLD | COM Maas | COM VST |
|---|---|---|---|---|
| COM HD-D | 1 | | | |
| (*n* = 2200) | | | | |
| COM MTLD | .830** | 1 | | |
| (*n* = 2200) | | | | |
| COM Maas | -.887** | -.772** | 1 | |
| (*n* = 2200) | | | | |
| COM VST | .166** | .177** | -.168** | 1 |
| (*n* = 2200) | | | | |

**. Correlation is significant at the 0.01 level (2-tailed).

Table 5. Lexical diversity and VST performance by all students combined

A series of simple linear regressions were carried out using the combined data, showing how well VST scores predicted each lexical diversity measure. Scatterplots of those regressions are shown below as Figures 1, 2, and 3.

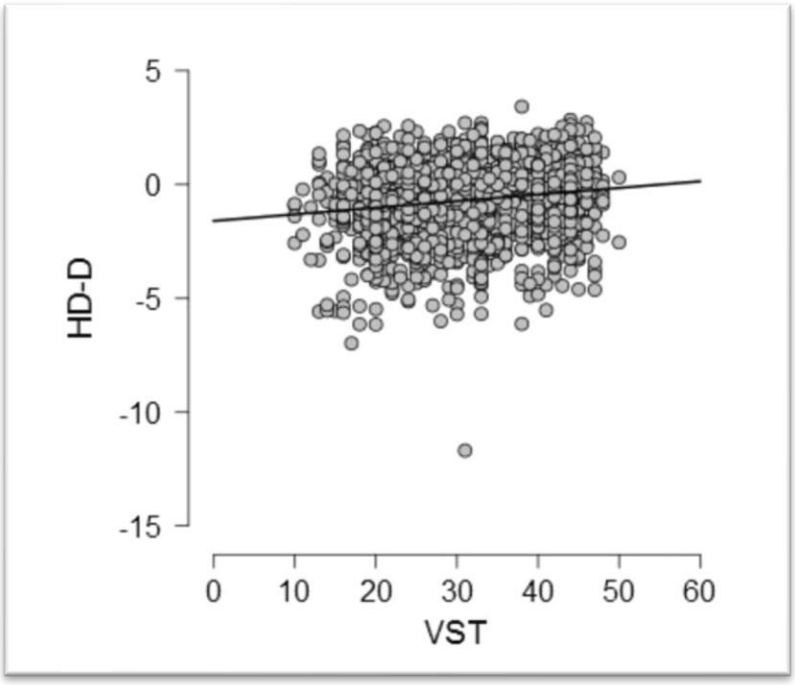


Figure 1. Scatterplot of HD-D regressed on VST.

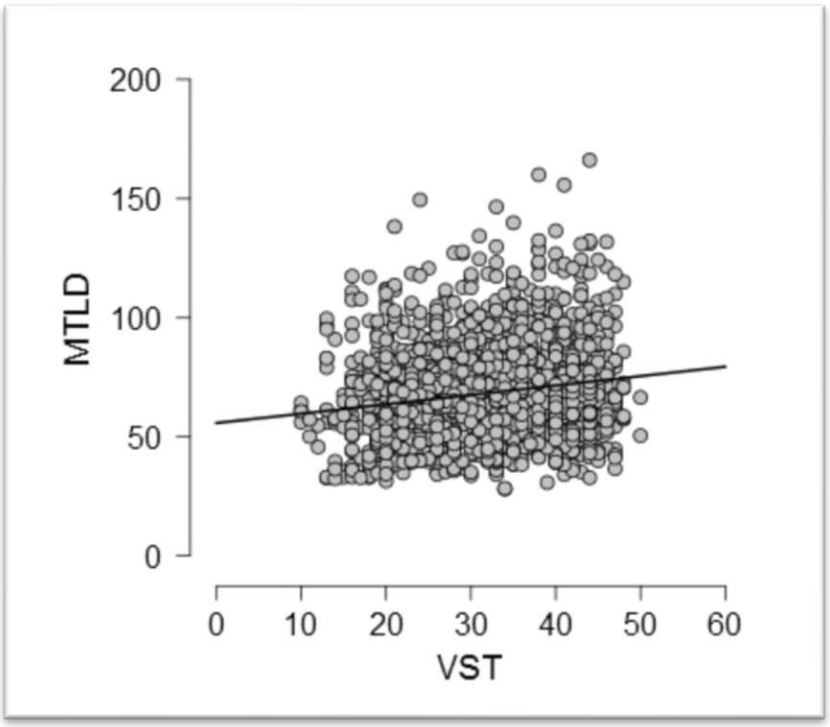


Figure 2. Scatterplot of MTLD regressed on VST.

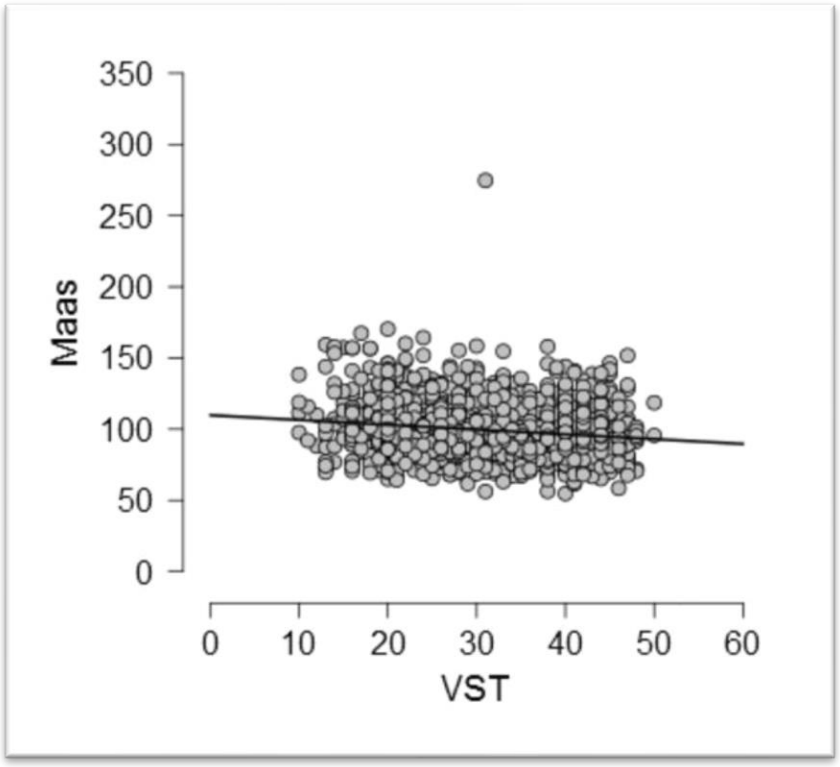


Figure 3. Scatterplot of Maas Index regressed on VST.

Note the presence of a single outlier that falls far outside of the trend. This outlier was removed from data as a quality control and overall trends were not meaningfully different after removal.

Results for the combined student data set demonstrates similar results as found in the correlational data above: high but imperfect correlations between the lexical diversity metrics and significant but very low correlations between each lexical diversity measure and VST score results. VST scores are thus not effective at predicting lexical diversity in student essays. While the evaluation of passive vocabulary knowledge may be of great interest and importance to both researchers and school administrators, tests of passive vocabulary appear to be invalid mechanisms for evaluating active vocabulary. Or, put another way, tests of passive vocabulary

like the VST cannot be used responsibly to make predictions about the range of vocabulary diversity likely to be displayed in written essays.

## 4. Discussion

There are several potential limitations of this study. First, as noted, the data analyzed here concerns only L2 learners of English from Asian backgrounds. While it would seem likely that a similar lack of correlation would be found among L1 writers and writers from other geographical, cultural, and language backgrounds, this is supposition and cannot be responsibly concluded from this data. Second, while the essay prompts utilized for the ICNALE are modeled on common writing tasks for college students, these texts are generated for the specific purpose of inclusion in the ICNALE and thus may not be considered authentic student texts. In particular, the lack of grading or other direct stakes for student writers may cause them to invest less effort in their writing. A third limitation of this study concerns the VST. While the VST is a well-known and respected test of vocabulary knowledge, whatever flaws and limitations in its validity and reliability might exist would undermine our confidence in this study's results. For example, researchers have suggested that the VST overestimates student vocabulary knowledge thanks to the possibility of guessing inherent to any multiple choice test (Stewart 2014). Such concerns about the VST must be considered when evaluating the claims of this research.

## 5. Conclusion

The very low correlations between VST results and lexical diversity, as measured using a variety of metrics, suggests that the relationship between active and passive vocabulary is not simple or direct. In particular, we cannot reasonably conclude that a student with a large passive

vocabulary as measured by the VST will necessarily produce an essay with similarly large range in vocabulary. This lack of a consistent relationship between a student's VST score and the lexical diversity of that student's essay demonstrates again the difficulty in measuring active vocabulary. If the VST or a similar test is used based on the assumption that the test will provide pragmatically useful information about how diverse a given student's vocabulary use will be in their actual writing, this practice is likely invalid, as the present study provides evidence that no such predictive inference can be made. As tests similar to the VST are regularly used to provide evidence for the overall communicative competence of test takers seeking entry into educational and professional systems, the role of tests of passive vocabulary in the larger language testing industry is called into question.

**References**


Baba, K. (2009). Aspects of lexical proficiency in writing summaries in a foreign language. Journal of Second Language Writing, 18(3) 191-208.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. Language Testing, 27(1) 101-118.

Daller, H., Milton, J., & Treffers-Daller, J. (Eds.). (2007). Modelling and assessing vocabulary knowledge. Cambridge: Cambridge University Press.

Grabe, W., & Kaplan, R. B. (2014). Theory and practice of writing: An applied linguistic perspective. London: Routledge.

Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? Applied linguistics, 11(4) 341-363.

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

Kojima, M., & Yamashita, J. (2014). Reliability of lexical richness measures based on word lists in short second language productions. System, 42 23-33.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: same or different?. Applied linguistics, 19(2) 255-271.

Laufer, B., & Paribakht, T. S. (1998). The relationship between passive and active vocabularies: Effects of language learning context. Language learning, 48(3) 365-391.

Laufer, B., and Ravenhorst-Kalovski, C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. Reading in a foreign language 22(1) 15-30.

Mass, H. D. (1972). Über den Zusammenhang zwischen Wortschatzumfang und Länge eines Textes. Zeitschrift für Literaturwissenschaft und Linguistik, 2(8) 73.

McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. Language Testing, 24(4) 459-488.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. Behavior research methods, 42(2) 381-392.

Meara, P. (1990). A note on passive vocabulary. Interlanguage studies bulletin, 6(2) 150-154.

Mizumoto, A., & Plonsky, L. (2015). R as a Lingua Franca: Advantages of Using R for Quantitative Research in Applied Linguistics. Applied Linguistics, 37 (2) 284-291.

Nation, P. (2001). Learning vocabulary in another language. Cambridge: Cambridge University Press.

Qian, D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. Canadian modern language review, 56(2) 282-308.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. Language testing, 18(1), 55-88.

Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. Tesol Quarterly, 27(4), 657-677.

Skehan, P. (2009). Lexical performance by native and non-native speakers on language-learning tasks. In Richards, B., Daller, H., Malvern, D., Meara, P., Milton, J., & Treffers-Daller, J., editors, Vocabulary studies in first and second language acquisition: The interface between theory and application, (pp 107-124). New York: Springer.

Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST?. Language Assessment Quarterly, 11(3) 271-282.

Wesche, M., & Paribakht, T. S. (1996). Assessing Second Language Vocabulary Knowledge: Depth Versus Breadth. Canadian Modern Language Review, 53(1) 13-40.