



2016-06-01

Rethinking Vocabulary Size Tests: Frequency Versus Item Difficulty

Brett James Hashimoto
Brigham Young University

Follow this and additional works at: <https://scholarsarchive.byu.edu/etd>

 Part of the [Linguistics Commons](#)

BYU ScholarsArchive Citation

Hashimoto, Brett James, "Rethinking Vocabulary Size Tests: Frequency Versus Item Difficulty" (2016). *All Theses and Dissertations*. 5958.
<https://scholarsarchive.byu.edu/etd/5958>

This Thesis is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in All Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu.

Rethinking Vocabulary Size Test Design:
Frequency Versus Item Difficulty

Brett James Hashimoto

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of
Master of Arts

Dan P. Dewey, Chair
Mark Davies
Dee Gardner
Troy Cox

Department of Linguistics and English Language
Brigham Young University

June 2016

Copyright © 2016 Brett James Hashimoto

All Rights Reserved

ABSTRACT

Rethinking Vocabulary Size Test Design: Frequency Versus Item Difficulty

Brett James Hashimoto
Department of Linguistics and English Language, BYU
Master of Arts

For decades, vocabulary size tests have been built upon the idea that if a test-taker knows enough words at a given level of frequency based on a list from corpus, they will also know other words of that approximate frequency as well as all words that are more frequent. However, many vocabulary size tests are based on corpora that are as out-of-date as 70 years old and that may be ill-suited for these tests.

Based on these potentially problematic areas, the following research questions were asked. First, to what degree would a vocabulary size test based on a large, contemporary corpus be reliable and valid? Second, would it be more reliable and valid than previously designed vocabulary size tests? Third, do words across, 1,000-word frequency bands vary in their item difficulty? In order to answer these research questions, 403 ESL learners took the Vocabulary of American English Size Test (VAST). This test was based on a words list generated from the Corpus of Contemporary American English (COCA).

This thesis shows that COCA word list might be better suited for measuring vocabulary size than lists used in previous vocabulary size assessments. As a 450-million-word corpus, it far surpasses any corpus used in previously designed vocabulary size tests in terms of size, balance, and representativeness. The vocabulary size test built from the COCA list was both highly valid and highly reliable according to a Rasch-based analysis. Rasch person reliability and separation was calculated to be 0.96 and 4.62, respectively.

However, the most significant finding of this thesis is that frequency ranking in a word list is actually not as good of a predictor of item difficulty in a vocabulary size assessment as perhaps researchers had previously assumed. A Pearson correlation between frequency ranking in the COCA list and item difficulty for 501 items taken from the first 5,000 most frequent words was 0.474 ($r^2 = 0.225$) meaning that frequency rank only accounted for 22.5% of the variability of item difficulty. The correlation decreased greatly when item difficulty was correlated against bands of 1,000 words to a weak $r = 0.306$, ($r^2 = 0.094$) meaning that 1,000-word bands of frequency only accounts for 9.4% of the variance. Because frequency is not a highly accurate predictor of item difficulty, it is important to reconsider how vocabulary size tests are designed.

Keywords: vocabulary size, vocabulary assessment, vocabulary breadth, vocabulary level, language testing, test design

ACKNOWLEDGEMENTS

I owe an enormous debt to Dr. Dan Dewey. There is no way he could know the profound influence he has been these last two years during my time at BYU in shaping the type of academic and the type of person I would become. Thank you for all of your advice and edits and care and patience. I am also deeply indebted to a wonderful thesis committee: Dr. Mark Davies, Dr. Dee Gardner, and Dr. Troy Cox. They have all have good-naturedly answered all of my frankly witless questions, gave me materials to read, and contributed brilliant and invaluable insight as I went about designing and conducting my study and writing my thesis.

Thank you to all of the coworkers at the Missionary Training Center who made this study possible: Dr. Ray Graham, David Macfarlane, Tim Zeidner, Kenny Adams, Caitlin Jolley, Courtney Dygert, and all of the translators who worked on the VAST.

Thanks to John Nielsen who has been my sounding board for the many, many, many terrible ideas that I have had for thesis topics and research questions and to Kyra Nelson who meticulously edited the insane ramblings of a psychopath and made sense of it all.

Perhaps more indirectly, my loving parents, Jim and Miyo, have always emphasized the importance of higher education. My understanding fiancé, Alisha, has encouraged me and thoughtfully allowed me to devote much of my attention to completing my thesis instead of where it deservedly might have been, constantly on her. I have an army of supportive family, friends, church leaders, coworkers, and classmates whom I also dearly love. Thank you all.

Most of all, I am grateful for God's guiding influence in helping me to choose to attend Brigham Young University and in selecting my thesis topic and committee members. In all ways and in all things, I owe everything that I have accomplished and everything that I am to Him.

TABLE OF CONTENTS

TITLE PAGE	i
ABSTRACT	ii
ACKNOWLEDGEMENTS	iii
LIST OF TABLES	vii
LIST OF FIGURES	viii
Chapter 1: Introduction	1
Chapter 2: Review of Literature	9
The Importance of Word Frequency and Vocabulary Size Testing.....	9
Word frequency and L2 vocabulary acquisition.....	10
Vocabulary size tests for placement, diagnostic, and admissions purposes.....	11
Correlating vocabulary size with other aspects of language.	13
Size and depth.....	15
Testing Vocabulary Knowledge.....	17
Types of word knowledge.	17
Test item type selection for vocabulary size tests	19
Word as a variable.	30
The Construct of a Word.....	31
Type and token.	32
Word family.....	32
Lemma.	38
Homonymy and Polysemy.....	39
Lexeme.	41
Multiword and other lexical items.....	41
Generating Corpora and Word Lists	47
Factors for frequency list creation.	47
Lists used in vocabulary size tests of English.	50
The Corpus of Contemporary American English word frequency list.....	53
Construction and Validation of the Major Vocabulary Size Tests of English	61
The Vocabulary Levels Test.....	62
The Productive Vocabulary Levels Test (Laufer & Nation, 1999).	66
The Computer Adaptive Test of Size and Strength (Laufer & Goldstein, 2004).	70

The Vocabulary Size Test (Nation & Beglar, 2007).	71
The Eurocentres Vocabulary Size Test (Meara & Buxton, 1987; Meara, 1992).	78
Research Questions	83
Chapter 3: Methodology	84
Participants	84
Testing Instrument.....	87
VAST as a yes/no test.....	87
Test format.....	88
Real word selection.	90
Pseudoword selection.	90
Test instructions.....	91
Procedure.....	93
Piloting.....	93
Primary testing.....	93
Scoring.....	94
Data Analysis	94
Chapter 4: Results	95
Rasch-Based Analysis	95
Reliability.	95
Fit.....	96
Non-Rasch-Based Analysis.....	99
Descriptive Statistics.	99
Correlation.	101
Analysis of Variance.	101
Chapter 5: Discussion	103
Research Question 1: To what degree is a vocabulary size test based on the COCA word list reliable and valid?	103
Construct Validity.....	103
Reliability.	104
Separation.....	105
Fit Statistics.	105
ANOVA.....	106

Research Question 2: Is a vocabulary size test based on the COCA word list more reliable and valid than vocabulary size tests based on other word lists?	107
Vocabulary Levels Test	107
Productive Vocabulary Levels Test	108
Vocabulary Size Test	108
Eurocentres Vocabulary Size Test	109
Construct Validity	109
Research Question 3: Do words across 1,000-word frequency bands vary in their item difficulty?	111
Variance and range.	111
Correlations.	111
Chapter 6: Conclusion	113
Implications	113
Limitations	114
Future Research	115
References	119
Appendix A: Vocabulary of American English Size Test Format and Sample Items	134
Appendix B: Item-Person Map	157
Appendix C: Item: Measure Map	158

LIST OF TABLES

Table 1: Frequency versus dispersion.....	7
Table 2: Coverage of Frequent Words.....	11
Table 3: Components of word knowledge.....	18
Table 4: Item factor analysis results.....	29
Table 5: Types of Words Excluded from Previous Vocabulary Size Tests.....	45
Table 7: Comparison of corpora size and published year.....	54
Table 8: COCA composition.....	55
Table 9: Comprehensive comparison between corpora for vocabulary size tests.....	61
Table 10: Rasch item difficulty for VLT.....	65
Table 11: Reliability of the VLT.....	66
Table 12: Reliability of PVLT.....	68
Table 13: Correlations between forms of the PVLT.....	68
Table 14: Breakdown of age of participants.....	85
Table 15: Number of Years Learning English by Participants.....	86
Table 16: L1s of Participants.....	87
Table 17: Pseudowords from EFL Vocabulary Test in COCA.....	91
Table 18: Summary of Rasch-based analysis.....	95
Table 19: Rasch-Based Analysis by Level.....	96
Table 20: Misfitting items from Rasch-based analysis of the VAST.....	98
Table 21: Descriptive statistics from VAST results.....	99
Table 22: ANOVA of 1,000-word levels of VAST by Logits.....	102

LIST OF FIGURES

Figure 1: Dimensions of vocabulary assessment.....	20
Figure 2: Decontextualized versus contextualized items.....	22
Figure 3: Ambiguity with multiple choice items.....	23
Figure 4: Problematic item from VST.....	23
Figure 5: Example of VLT item.....	63
Figure 6: Example of PVLt item.....	66
Figure 7: PVLt example 1.....	69
Figure 8: PVLt example 2.....	69
Figure 9: PVLt example 3.....	70
Figure 10: CATSS example item.....	71
Figure 11: VST example item.....	72
Figure 12: Wright map of VST.....	75
Figure 13: Difficulty by level for VST (Beglar, 2010, p. 109).....	76
Figure 14: Possible responses for yes/no test (Beeckmans et al, 2001, p. 237).....	81
Figure 15: Scoring formula for yes/no tests (Mochida & Harrington, 2006, p. 76).....	82
Figure 16: Sample item from the Vocabulary of American Size Test.....	89
Figure 17: Histogram of real world item logit values.....	100
Figure 18: Boxplot of logit values—WMLE (Measure) or weighted maximum likelihood estimator is also called item logit value.....	100

Chapter 1: Introduction

Understanding how many words an individual knows in a given language can be useful in a variety of ways. For child learners of a language, it allows researchers and educators to understand how much and at what rate vocabulary is being acquired by certain ages. This could inform curricular decisions about how many and how fast new words should be introduced in educational programs and how quickly a child is acquiring vocabulary compared to his or her peers. Non-native speakers would have a means by which they could compare their lexical knowledge with that of a native speaker. It could also inform them about the amount of vocabulary they might need to be successful at a foreign university or to work abroad. This type of information would help second-language program administrators, curriculum designers, and teachers by giving them insights about what types of vocabulary their learners need and when. This type of information could prove useful in other areas such as language learning software and programmatic or proficiency assessments.

Vocabulary size tests, also known as vocabulary breadth tests or vocabulary levels tests, are designed to approximate the number of words an individual knows in a given language. Early published pursuits of measuring an individual's lexicon date back over 100 years (Nation & Waring, 1997), and since that time, the majority have been in English (Beglar & Hunt, 1999; Cameron, 2002; Gyllstad, Vilkaite, & Schmitt, 2015; Nation, 1983; Laufer & Nation, 1999; McLean, Kramer, & Beglar, 2015; Meara, 1988, 1992; Molina, 2009; Webb, 2008).

Paul Nation was the first to generate a modern model of the vocabulary size test. Paul Nation's Vocabulary Levels Test (VLT) resolved the major methodological issue of estimating the breadth of an individual's lexicon in a practical manner (Nation, 1983). It is time and energy consuming to test each entry in a dictionary one-by-one. Rather than this approach, Nation innovated an ingenious methodological design based on a very simple premise: if a word appears

more frequently in a language, it is more likely to be known by a speaker of that language than a less frequent word. Therefore, a learner is not likely to know *thesaurus* if they do not know *dictionary*, and they are not likely to know *dictionary* if they do not know *book* and so forth. Thus, the premise is that by sampling items at various word frequencies, determinations can be made about the approximate vocabulary size of a test-taker.

Certainly, the two most widely used and validated measures of vocabulary size are in the English language and are the Vocabulary Levels Test (Beglar, 2009; Nation, 1983; Nation & Beglar, 2007; Read, 2000) the Eurocentres Vocabulary Size Test (EVSTT—Meara, 1988, 1992, 2005, 2010; Meara & Jones, 1990; Read, 2000). Both of these measures now have online versions which have improved upon previous versions.

Despite the ever-growing body of literature about and interest in vocabulary size tests, a handful of fundamental methodological problems still remain and have yet to be adequately addressed in the published literature. The design of this thesis is to address three of these issues: defining the construct of a word, the selection of words for the test, and the way levels are assigned to test-takers.

Defining what exactly constitutes a “word” is still a major issue that exists, not only in vocabulary size testing but in virtually all vocabulary testing. When testing vocabulary size, the construct of word has been defined in terms of word families (Nation, 1983; Read, 2000). Stated simply, a word family includes the base form of a word plus any word that can be derived from that base form excluding compounding of morphemes. For example, a word family for the word *develop* would include *develop* (verb), *develops* (verb), *developed* (verb and adjective), *developing* (verb and adjective), *developable* (adjective), *undevelopable* (adjective), *developments* (noun), *developmentally* (adverb), *developmentwise* (adjective and adverb),

semideveloped (adjective), *antidevelopment* (noun and adjective), *redevelop* (verb), *predevelopment* (noun or adjective), and many others (Bauer & Nation, 1993). As shown by the example above, a word family in its most general sense can be composed of many words from various parts of speech. Convergent research from diverse studies has shown that for both native and non-native learners of a language, morphology is learned incrementally. Full morphological awareness and mastery may take many years, and some, especially L2 learners, may never fully acquire some derivational morphology (Berko, 1958; Derwing & Baker, 1979; Nagy, Diakidoy, & Anderson, 1993; Schmitt & Meara, 1997; Schmitt & Zimmerman, 2002).

Also, word families include homonymous and polysemous words. Homonymous words have clearly distinct meanings but share the same orthographic representation such as a *bow* as in *to tie a bow* versus *the bow of a boat* or even *a bow and arrow*. The matter becomes even more complex when we consider polysemes. Hatch and Brown explained these concepts thusly.

Polysemes are the many variants of meaning of a word where it is clear that the meanings are truly related. The verb *break* has many different variants which are related in meaning [*He broke his leg; The cup broke; She broke his heart; She broke the world record, etc.*]. The verb *put* also has an array of polysemes. Homonyms (sometimes called homographs [in their written form]), on the other hand, are variants that are spelled alike but which have no obvious commonality in meaning...One question regarding core research is whether or not polysemes such as those shown in the *break* example really are the same word...or whether some should be listed as separate lexical items. (Hatch & Brown, 1995: 49)

Although scholars have addressed the differences and problems in testing homonyms and polysemes (Huertas, Gómez-Ruelas, Juárez-Ramírez, & Plata, 2011), the effect of this issue on vocabulary size testing has yet to be fully considered.

One final difficulty in defining the construct of word, particularly in English, is how to consider multiword lexical units. Gardner (2013) lists the several classes of multiword items as follows:

- Phrasal verbs (e.g., break up, break down, break off);
- Idioms (pop the question, beat around the bush, chip off the old block);
- Open compounds (carbon dioxide, Education Reform Act, sleeping bag);
- Complex discourse markers (in addition to, on the other hand, as a result of);
- Names (George Washington Carver, Henry David Thoreau, William Shakespeare);
- Hyphenations (action-packed, age-specific, mother-in-law);
- Stock phrases (good morning, have a great day, see you later);
- Pre-fabricated strings also known as lexical clusters, lexical bundles, or lexical chunks (the fact that..., the point is..., do you think...); (p. 21-22).

These multiword units behave the same as a single lexical item and have distinct meaning as units separate from the individual components from which they are composed. These, too, do not seem to be addressed anywhere in the literature with regards to how to handle them in a vocabulary size test.

The way in which lists of words that are used in vocabulary size tests have been selected and created is another significant set of issues that need to be addressed. Many previous studies validating vocabulary size tests have examined the validity and usefulness of a particular type of

test item (e.g. c-test, multiple choice, yes/no, etc.) or the scoring method of vocabulary size tests (Beeckmans, Eyckmans, Janssens, Dufranne, & Van de Velde, 2001; Beglar, 2009; Beglar & Hunt, 1999, Gyllstad, Vilkaite, & Schmitt, 2015; Harsch & Hartig, 2015; Huibregtse, Admiraal, & Meara, 2002; Laufer & Nation, 1999; Laufer & Goldstein, 2004; Meara & Jones, 1988; Mochida & Harrington, 2006; Nation, 1983, 1990; Nation & Beglar, 2007; Read, 1993, 2000; Schmitt, Schmitt, & Clapham, 2001; Shillaw, 1996). However, there have been no studies that examine the word lists themselves to determine if they are appropriate and valid for vocabulary size testing. Very little information is ever listed in any study about the nature and origin of the word lists used to generate these tests. Where such information actually is readily available, there appears to be some methodological problems with their usage in contemporary vocabulary size tests. Specifically, problems exist in both the design of the corpora from which the lists were derived and with the generation of the word frequency lists themselves. However, this thesis will primarily address only four types of problems:

1. The corpus is too small in terms of number of tokens and texts.
2. The composition of the corpus is not balanced.
3. The corpus is dated.
4. The word frequency list from the corpus does not include dispersion statistics.

Not all of the word lists used in existing vocabulary size tests have all of these problems, although some do. However, all of them do have more than one of problems listed above. Each of these four points will be more fully addressed and explained later in this thesis.

If a corpus is too small, it can become biased by a small group of texts or even a single document that may conflate the frequency of a word that is actually very rare. In the extreme example, we could see that a corpus based on only one book would exclude any words that are

not pertinent to the topic of the book and a large general frequency list would be impossible to derive.

If a corpus is composed of only one genre of texts it may exclude common words that may appear in many other genres. Also, if a corpus is exclusive of particular genres, it may exclude vocabulary unique to particular types of texts. For example, a corpus of only textbooks or scholarly articles will exclude important informal language and a corpus of only spoken language will exclude academic language. Neither alone is sufficient to generate a frequency list which accurately represents general language. Therefore, a corpus composed of a balance of different genres will yield better general frequency information.

Finally, corpora that are outdated have their own set of problems. New words are always entering the lexicon while other words fall out of usage and relevance. Therefore, a corpus of contemporary language is most suitable because it is more likely to include the types of words that learners will encounter and acquire.

Another reason for having multiple genres is it allows a list to incorporate both frequency and dispersion. Very few researchers ever take into account dispersion as an important methodological factor when generating word frequency lists. Put simply, dispersion describes how well a word is spread across a corpus. It answers the questions “Are all of the tokens of this particular word in one text?” “Is it only frequent in one genre?” “Or, does it appear everywhere?” In order to generate good general word frequency lists, these questions need to be addressed (Lyne, 1986). Table 1 from the frequency list generated through the Corpus of Contemporary American English (COCA—Davies, 2011), *acculturation* is over 2,000 ranks lower than *apiece*. Even though the raw frequency count for *acculturation* is one more than *apiece*, it appears in much, much fewer places in the corpus. A frequency list that accounts for both raw frequency

and dispersion will rank words differently than a list based solely on raw frequency, and the disparity can be quite dramatic.

Frequency versus dispersion				
Rank	Word	POS	Frequency	Dispersion
9853	apiece	(Adverb)	1589	0.93
12106	acculturation	(Noun)	1590	0.64

Table 1: Frequency versus dispersion

Because of these various potential issues, it is important to investigate whether taking measures to rectify them improves the nature of a vocabulary size test.

Another issue is that vocabulary size tests have yet to attempt to measure vocabulary size except by grouping words together in 1,000-word units (Beglar & Hunt, 1999; Gyllstad et al, 2015; Meara, 1992; Meara & Jones, 1988; 1990; Nation, 1983; Nation & Beglar, 2007. Nation's Vocabulary Levels Test uses levels that vary in size, with one level representing 5,000 words (Nation, 1983). In other words, the result of a test will indicate that a test-taker knows between 1,000-2,000 or 5,000-6,000 words but cannot give more granular feedback indicating where among those 1,000 words they fall. Nobody has yet addressed whether a vocabulary size assessment can determine the number of words an individual knows down to 500 words or even 100 words. These levels also make the unverified assumption that all words behave similarly and are of equal difficulty across a given level of 1,000 words in a vocabulary size assessment. These tests have presumed that word 1,000 is of equal difficulty in a vocabulary size assessment as word 1,999. No investigation of whether or not this holds true has yet been undertaken.

Because of the limitations of previous vocabulary size tests and the lack of exploration into these important methodological issues, the purpose of this thesis will be to address these varied issues more fully to determine the validity and usefulness of the design of previous

vocabulary size tests. This thesis will also examine possible areas of improvements that can be made to future assessments and future areas for research in vocabulary size testing.

In order to accomplish the aims of this thesis, a new vocabulary size test, the Vocabulary of American English Size Test (VAST), was constructed. This test used words selected from a list generated from a corpus of modern English. This list was lemmatized (grouped by lemma) instead of being grouped into word families and also incorporated a dispersion statistic. Furthermore, it sampled a much larger number of words than previous vocabulary size tests (every tenth word) in order to observe the item difficulty of words across levels of 1,000 words and to investigate if word levels can be fewer than 1,000 words in size. An evaluation of this test will allow us to see if differences in the design of the VAST from previous vocabulary size tests improve its ability to function for its intended purpose of evaluating vocabulary size of the test-taker.

Chapter 2: Review of Literature

In order to fully understand the uninvestigated problems with vocabulary size testing, it is important to first review the published literature discussing the value of testing vocabulary size, the construct of a word, acquisition of vocabulary, methods of vocabulary testing, word lists, creation of vocabulary size tests, and criticisms of vocabulary size tests. Each of these topics will, in turn, shed light on the previously overlooked limitations of former studies and will frame this study's place within the current discussion of vocabulary size tests.

The Importance of Word Frequency and Vocabulary Size Testing

Many studies have shown the importance of and the need for vocabulary size testing. Measurement of word knowledge based on word frequency has been deemed important because of various analyses that have been conducted confirming the importance of word frequency in language acquisition and usage. These analyses definitively show that the number and types of words a learner knows makes an immense difference in his or her ability to function in the L2 (Hazenbergh & Hulstijn, 1996; Laufer, 1992; Nation, 1990, 2006; Schonell, Meddleton, Shaw, Routh, Popham, Gill, Mackrell, & Stephens, 1956; Sutarsyah, Nation, & Kennedy, 1994).

Other studies have shown how vocabulary size tests are useful for placement, diagnostic, or admissions purposes (Laufer, Elder, Hill, & Congdon, 2004; Laufer, 1998; Laufer & Nation, 1999; Meara, 1992; Meara and Jones, 1988; Schmitt, 1994). Still others have correlated vocabulary size with general intelligence (Anderson & Freebody, 1983), academic success (Milton & Treffers-Daller, 2013; Saville-Troike, 1984) reading comprehension (Beglar & Hunt, 1999; Laufer, 1992; Qian, 1999; Stæhr, 2008; Zimmerman, 2004), writing ability (Beglar & Hunt, 1999; Laufer, 1998, Milton, 2010; Stæhr, 2008; Zimmerman, 2004), listening comprehension (Beglar & Hunt, 1999; Stæhr, 2008; Zimmerman, 2004), oral proficiency

(Milton, Wade, & Hopkins, 2010; Zimmerman, 2004), grammatical ability (Zimmerman, 2004), depth of vocabulary knowledge (Shimamoto, 2000; Schmitt, 2014; Vermeer, 2001), and overall general language proficiency (Milton, 2010; Milton & Alexiou, 2009; Tseng & Schmitt, 2008; Zimmerman, 2004). These combined evidences show the worth of vocabulary size testing and the worth of investigating possible ways to improve the validity and usefulness of vocabulary size tests.

Word frequency and L2 vocabulary acquisition.

Research has shown that not all words are created equal. Many studies have shown in different ways that more frequent words are more important for language learners than less frequent words for a number of reasons. For example, more frequent words are much more important for language comprehension and usage by language learners. One study revealed that the 2,000 most frequent word families provide 99% of the words needed for everyday oral communication (Schonell, et al, 1956). Another approximated that these 2,000 word families constitute about 87% of written texts (Nation, 1990). Other studies differ in their estimates in terms of the number of words that L2 learners need in order to read an average written text, but they agree that a certain number of word families are important for the learners to master: 3,000 word families (Laufer, 1992), 3,000-5,000 words word families (Nation & Waring, 1997), and up to 10,000 (Hazenbergh & Hulstijn, 1996). Still other studies assert that 4,000-5,000 word families are needed for understanding academic texts such as textbooks (Sutarsyah, Nation, & Kennedy, 1994), 10,000 word families are needed to study at a university level (Hazenbergh & Hulstijn, 1996), and that in the LOB Corpus, the most frequent 1,000 word families cover 77.86% of written texts, the second most frequent 1,000 covers 8.23 %, and the third most frequent 1,000 covers only 3.7% with each successive level covering fewer and fewer

percentages of the corpus as shown in Table 2 (Nation, 2006). The sum of evidence from these studies shows that word frequency does indeed matter in language acquisition, and therefore, so does measuring vocabulary acquisition based on word frequency.

Coverage of Frequent Words		
Frequency Band	% Coverage added by band	Cumulative %
1,000	77.86	77.86
2,000	8.23	86.09
3,000	3.70	89.16
4,000	1.79	90.95
5,000	1.04	91.99
6,000	0.70	92.69
7,000	0.65	93.34
8,000	0.40	93.74
9,000	0.32	94.06
10,000	0.32	94.38
11,000	0.16	94.54
12,000	0.14	94.68
13,000	0.12	94.80
14,000	0.10	94.90
(Adapted from Nation, 2006, p. 64)		

Table 2: Coverage of Frequent Words

Vocabulary size tests for placement, diagnostic, and admissions purposes.

As mentioned before, vocabulary size tests have primarily been used for placement, diagnostic, and admission purposes. Several studies have been conducted to investigate the effectiveness of these types of tests in achieving those purposes and have concluded that they generally have at least a moderate degree of success in fulfilling their designed function.

Meara and Jones (1988) was the first article to discuss using vocabulary size for placement purposes. The Eurocentres Vocabulary Size Test (EVST) was correlated with a simple programmatic placement test used at the Eurocentres schools in the United Kingdom known as the Joint Entrance Test (JET). The JET was composed of several sections: listening comprehension, grammar, reading, and an oral interview. The JET and EVST were taken by 267

students in the program at two different schools. The scores between the two tests were correlated, and the results yielded an overall correlation coefficient of 0.664 for 109 test-takers at the school in Cambridge and 0.717 for 159 test-takers at the school in London. They determined this to be a moderately strong correlation. The students were then placed in classes based on their JET scores. After one week of classes, the researchers checked to see if the students placed by the test has been placed correctly. Of the group of 109 Cambridge students, five were relocated based on their classroom performance, and for all five students, their relocations were in line with results produced by the EVST. At the London school, a questionnaire was administered to teachers about 14 students whose scores on the JET and EVST yielded the greatest discrepancies between the two tests. Through a survey administered to teachers in the program, of those 14 cases, teachers' judgments agreed with the EVST scores over the JET scores in nine of them. Both the correlations and the teachers' judgments show that the EVST was just as effective in programmatic placement as the JET, if not more so. Thus, in this instance the vocabulary size test outperformed the comprehensive JET test in its placement ability.

Laufer (1998) and Laufer and Nation (1999) also investigated using vocabulary size as a placement tool. In the first study, Laufer examined 10th grade (n=26) and 11th grade (n=22) Israeli English language learners using the Productive Vocabulary Levels Test (PVLTV) and found that on the means of the scores of the 11th graders to be highly statistically significantly ($p < 0.0005$) higher across all levels of vocabulary that they tested. Laufer and Nation (1999) tested EFL learners: 10th graders (n=24), 11th graders (n=23), 12th graders (n=18), and 1st year university students (n=14) who had all studied English since the time they were 5th grade. In this study, the authors again used the PVLTV and discovered again that scores increased according by

grade to a statistically significant degree ($p < 0.0001$). Other studies, too, have shown similar results to the ones summarized here, also validating vocabulary size as a means of placement (Meara, 1992; Read, 2000; Schmitt, 1994).

Read (2000) performed a Guttman scalogram analysis of test scores published in Nation (1990) for the VLT. According to Hatch and Farhady (1982, p. 181), a coefficient of scalability should be well above 0.60 if the scores are to be considered scalable. Read obtained scores of 0.90 and 0.84 from analyses of two sets of scores. This implicational scale shows that the levels of the test are separating fairly well between one another, which the author states shows how this test is valid for both placement and diagnostic purposes.

Laufer et al (2004) created a computer-adaptive test and through their study, they found results suggesting vocabulary size tests to be both useful and valid for diagnostic purposes, especially when combined with measures of strength of knowledge of vocabulary. This same study also suggested, based on the body of literature that exists about vocabulary size tests, that “as vocabulary size is related to success in reading, writing, and general language proficiency as well as to academic achievement [Saville-Troike, 1984; Laufer, 1997], size tests can provide efficient placement and admission [functions] in language teaching programmes” (Laufer et al, 2004, p. 9).

Correlating vocabulary size with other aspects of language.

Although multiple studies have correlated vocabulary size to different aspects of language proficiency, it will suffice in this section to summarize only two of them which are representative of the general trends.

Beglar and Hunt (1999) correlated scores of 496 students on two levels of the Vocabulary Levels Test (VLT) with their overall TOEFL scores as well as with the Reading Comprehension,

Listening Comprehension, and Structure and Written Expression Subsections of the test. Using Hotelling's t-test, the authors reported that the two levels of the VLT had a strong correlation with the overall TOEFL scores ($r = .70$) and the Reading Comprehension Subsection ($r = .69$), a moderately strong correlation with the writing subsection ($r = .65$), and a moderate correlation with the Listening Comprehension Subsection ($r = .43$). These correlations are particularly meaningful because of the high number of participants in the study and because the TOEFL is perhaps the most widely used and validated test of English as a foreign language in the world.

Milton and Alexiou (2009) designed a rather complex study investigating a total of 575 learners: EFL students in Hungary and Greece, French as a foreign-language learners in Britain, Greece, and Spain, and Greek as a second language learners in Greece. In this study, where the scores of a newer version of the EVST known as *X_Lex* were correlated against the Common European Framework of Reference for Languages (CEFR) levels assigned to them by standardized tests across the different schools in the different countries. With high numbers of participants and varied L1s and L2s, the results of the study showed that cross-linguistically, as one moves up through the CEFR levels, they tend to know progressively more vocabulary. Among the different schools, a large amount of the variance of the CEFR level a learner attains can be explained by the single element of vocabulary size: 70% in Spain and Greece and 40% in Britain. The CEFR scales are, of course, designed to measure overall language proficiency in the four critical skills of reading, writing, speaking, and listening and are widely accepted as the most well-designed proficiency scales in existence. It is impressive that a vocabulary size assessment could explain such a large percentage of the variance of a widely-accepted standard of language proficiency, especially in multiple schools, in multiple countries, and in multiple languages.

Size and depth.

Vocabulary size has also been correlated with various measures of vocabulary depth. Vocabulary depth, also known as vocabulary strength, has been conceptualized in a variety of ways. Whereas vocabulary size is fairly straightforward, i.e., counting the number of known words, vocabulary depth is much more complex. Schmitt (2014, p. 922) organized studies of depth of lexical knowledge concisely into seven categories as follows:

1. Receptive versus productive mastery (ability to recognize or understand a word versus ability to use it)
2. Knowledge of multiple word knowledge components (as shown in Table 3 below)
3. Knowledge of polysemous meaning senses (the multiple senses of a single lexical item)
4. Knowledge of derivative forms (word family members)
5. Knowledge of collocation (what words are found near one other frequently)
6. The ability to use lexical items fluently (speed of lexical access)
7. The degree and kind of lexical organization (the manner in which words associate and interact with other words around them both semantically and collocationally)

Studies correlating each one of these constructs against vocabulary size have been undertaken and yielded varied, sometimes contrastive, results. However, despite the lack of consistency amongst the findings of some researchers, close examination of the body of literature taken as a whole, general trends can be found. For example, practically every study undertaken investigating size and depth has found some sort of positive corollary relationship between them (Laufer & Goldstien, 2004; Milton & Hopkins, 2006; Milton, Wade, & Hopkins, 2010;

Shimamoto, 2000; Schmitt & Meara, 1997; Schmitt & Zimmerman, 2002; Vermeer, 2001; Webb, 2008).

In a conceptual review article summarizing all of the major studies of size and depth of vocabulary knowledge, Schmitt (2014) concluded that the correlations between various size and depth measures were stronger for higher frequency words and that “for higher frequency words, there is often little difference between size and depth measures. However, for lower frequency words and for larger vocabulary sizes, there is often a gap between size and depth, as depth measures lag behind the measures of size. Furthermore, some types of word knowledge (e.g., derivate knowledge) seem to have generally lower correlations with size than other types” (p. 941). This section will now briefly summarize some of the more significant findings regarding the relationship between vocabulary size and vocabulary depth.

Melka (1997) estimated that 92% of receptive vocabulary is also known productively. While at least one study estimated that figure to be an even larger percentage (Takala, 1985), other studies approximate it as being much less (Fan, 2000; Laufer, 2005). Milton (2009) found that written and oral vocabulary are strongly linked in Arabic and Greek at ($r = .68$) respectively. The study also found that for vocabulary size of $\leq 2,000$ lemmas, the number of phonologically known items is greater than written, but for greater than that number, more items are known orthographically than orally. Chui (2006) found that for academic words, size measures correlated significantly with various depth measures: for derivatives ($r = .78$), for collocations ($r = .69$), for meaning ($r = .69$), and for word class or part of speech ($r = .53$). Laufer et al (2004) correlated vocabulary size tests measuring different types of word knowledge and determined that “[a]ctive and passive recognition...appear to be indistinguishable from one another in terms of difficulty” (p. 218).

Many other studies have been conducted correlating other aspects of vocabulary size and depth and found similarly moderate or high positive correlations. To the knowledge of the author, no study has found negative correlations or even weak correlations between them.

Testing Vocabulary Knowledge

Testing vocabulary is challenging because defining word knowledge is much more complex than simply testing the link between word form and word meaning. Word knowledge is composed of a variety of factors which overall compose both vocabulary size and depth. In order to understand vocabulary testing, it is important to examine vocabulary knowledge and its many dimensions.

Types of word knowledge.

The table below displays the classification system originally formalized in Richards (1976) and then further developed in Nation (1990). According to this system, the ability to use or produce a word requires additional knowledge beyond simply recognizing it. In other words, there are receptive and productive areas of vocabulary knowledge that mirror one another, and of the two, productive knowledge extends beyond receptive knowledge. This dichotomy has also been described as passive (receptive) and active (productive) knowledge.

Components of Word Knowledge		
<i>Form:</i>		
Spoken Form	R	What does the word sound like?
	P	How is the word pronounced?
Written Form	R	What does the word look like?
	P	How is the word written and spelled?
<i>Position:</i>		
Grammatical Patterns	R	In what patterns does the word occur?
	P	In what patterns must we use the word?
Collocations	R	What words or types of words can be expected before or after the word?
	P	What words or types of words must we use with this word?
<i>Function:</i>		
Frequency	R	How common is the word?
	P	How often should the word be used?
Appropriateness	R	Where would we expect to meet this word?
	P	Where can this word be used?
<i>Meaning:</i>		
Concept	R	What does the word mean?
	P	What word should be used to express this meaning?
Associations	R	What other words does this word make us think of?
	P	What other words could we use instead of the one?
Key: R = Receptive; P = Productive (Nation, 1990, p. 31)		

Table 3: Components of word knowledge

Understanding that vocabulary knowledge is complex and multidimensional is important for vocabulary testing because it informs the decisions one makes when designing test items; different kinds of test item types assess different types aspects of vocabulary knowledge. The studies in the previous section correlating different aspects of vocabulary knowledge becomes more meaningful when taken in this light. Whenever a particular type of vocabulary knowledge is tested for a word, it can give insight about the amount of other types of word knowledge a test-taker might have for that particular word. Schmitt (2014) asserts that based on the body of research from the last few decades, especially for higher frequency vocabulary, there may be

little difference between the number of words learners know receptively/passively and the number of words they know productively/actively.

However, there are certain types of word knowledge that seem to have consistently lower correlations with vocabulary size than other types even though those correlations may be statistically significant. Of all types of word knowledge, word derivations consistently correlated least with measures of vocabulary size (Chui, 2006; Milton, 2009; Noro, 2002; Schmitt & Meara, 1997; Schmitt, 2014). This is an important finding because many of the more popular types of vocabulary size group words together by word families. Word families include derivational morphology, but measures of vocabulary size generally correlate with knowledge of derivational affixes at .50 or lower for L2 learners (Schmitt, 2014). Therefore, researchers should be more careful when testing vocabulary size in assuming that because a learner knows one word in a family, he or she will know the other many words which are morphologically related.

Test item type selection for vocabulary size tests

Just as there are many types of vocabulary knowledge, there are many test item types when it comes to assessing that knowledge. Items can be described as being ‘discrete vs. embedded’, ‘selective vs. comprehensive’, and ‘contextualized vs. decontextualized’ (Read, 2000). Explanations of these three dimensions of vocabulary measures can be seen in the figure below.

<i>Dimensions of Vocabulary Assessment</i>		
<u>Discrete</u>	<----->	<u>Embedded</u>
A measure of vocabulary knowledge or use as an independent construct		A measure of vocabulary which forms part of the assessment of some other, larger construct
<u>Selective</u>	<----->	<u>Comprehensive</u>
A measure in which specific vocabulary items are the focus of the assessment		A measure which takes account all of the vocabulary content of the whole material in reading/listening tasks or the test-taker's response writing/speaking tasks
<u>Context-independent</u>	<----->	<u>Context-dependent</u>
A vocabulary measure in which the test-taker can produce the expected response without referring to any context		A vocabulary measure which assesses the test-taker's ability to take account of contextual information in order to produce the expected response
(Read, 2000, p. 9)		

Figure 1: Dimensions of vocabulary assessment

Understanding the dimensions of vocabulary measures is important because they reflect the construct of vocabulary knowledge being tested. Singleton (1999) argues that sufficient evidence has been uncovered by academic research to suggest that vocabulary knowledge is so integrated into other aspects of language that “the viability of a separate lexical construct has to be seriously questioned” (p. 269). The study suggests that vocabulary testing should be embedded and comprehensive, and vocabulary should be considered to be merely part of grammar, reading, writing, speaking, and listening. However, the work of most leading scholars in vocabulary testing including David Beglar, Batia Laufer, Paul Meara, Norbert Schmitt, James Milton, and Paul Nation contradict this idea. They treat vocabulary knowledge as its own specific and separable component of linguistic knowledge which can be tested independent of grammar, discourse, or high context (Beglar & Nation, 2007; Laufer, 2013; Laufer & Nation, 1999; Meara, 1996; Milton, 2009; Nation, 1990; Schmitt, 2014).

For the most part, vocabulary size tests have used discrete, selective, and decontextualized item types in their construction (Ishii & Schmitt, 2009; Meara & Buxton, 1987;

Nation, 1983; Nation & Beglar, 2007; Vermeer, 2001). However, some size tests have attempted to use context-dependent items (e.g., c-tests in vocabulary size assessments—Harsch & Hartig, 2015; Laufer & Goldstein, 2004; Laufer & Nation, 1999).

Various types of tests span a range of item formats including multiple choice, cloze, c-test, glossing/translating, word association, lexical frequency profiles, yes/no, and others that may combine features of these different formats. Using these diverse formats, educators and researchers are able to test the productive and receptive vocabulary knowledge of form, meaning, function, and appropriate usage in both oral and written contexts. However, only a handful of types of items have yet been used in vocabulary size tests. This section will now briefly discuss both how certain types of items have been used in vocabulary size tests. It will also cover why certain types of items have previously been excluded from these types of tests.

The two major factors determining item type selection for vocabulary size testing have been practicality of test design and isolation of the construct of vocabulary. Simply put, the construct of vocabulary is knowledge of the lexis of a language as a separate skill from reading, writing, speaking, listening, and grammar. Vocabulary size tests typically have many items. Therefore, it is impractical both for the test designers to generate large amounts of long or complex items and for test-takers to answer long batteries of cognitively demanding questions. Also, most previous studies have sought to validate measuring vocabulary size as a construct (Ishii & Schmitt, 2009; Meara & Buxton, 1987; Nation, 1983), validate certain measures or types of vocabulary size (Beeckmans et al, 2001; Beglar, 2010; Beglar & Hunt, 1999; Gyllstad et al, 2015; Huibregtse et al, 2002; Mochida & Harrington, 2006), and/or compare scores on vocabulary size tests with other language related skills or proficiencies (Beglar & Hunt, 1999;

Laufer, 1992; Qian, 1999; Milton, 2010; Milton, Wade, & Hopkins, 2010; Tseng & Schmitt, 2008; Schmitt, 2014; Zimmerman, 2004).

Vocabulary size tests are designed to isolate vocabulary as its own separate construct and thus have generally used discrete (also known as discrete-point) formats. Also, because these tests are designed to test words at specific levels of word frequency, selective measures were used. However, despite the general uniformity among almost all vocabulary size test designers on these first two dimensions of test measurement, there has been somewhat of a division between contextualized and decontextualized items among different vocabulary tests as attempts to measure active vocabulary size and passive vocabulary size respectively. However, as mentioned before, some researchers suggest that there may be little or no difference between active and passive vocabulary size for many learners which may make the distinction a moot point (Melka, 1997; Schmitt, 2014; Takala, 1997).

Usually contextualized items have the target word in a sentence, paragraph, or longer passage whereas completely decontextualized items have words in isolation. Although these two concepts may seem categorical, they, like ‘discrete vs embedded’ and ‘selective vs. comprehensive’, are on more of a continuum (Read & Chapelle, 2001). Examples of these can be seen in Figure 2, both in a multiple choice format.

<u>Decontextualized (Glossing)</u>			
pencil			
a. lapiz	b. papel	c. libro	d. grapadora
<u>Contextualized (Cloze)</u>			
John went to class today, and he got homework. He wrote his name with a _____ at the top of the page, and he started to try to answer the questions.			
a. pencil	b. paper	c. book	d. stapler

Figure 2: Decontextualized versus contextualized items

Despite the fact that most vocabulary size tests utilize decontextualized item types, there are some complications that arise and are challenging to reconcile because the item is not in context. Two such difficulties are the problems of homonymy and polysemy (Read, 2000). Homonymous or homographic words are words that are spelled the same orthographically but have different meaning. For example, if a discrete test question has the English word *match*, a test-taker may know that a match could be ‘a thin piece of wood tipped with materials that can be lit when rubbed against a rough surface’, but be unaware that the string of letters ‘m-a-t-c-h’ also means ‘a sports contest between two opposing sides’. Thus, a test question in the figure below may be confusing to the test-taker who only knows the former definition.

- match

 - a. a sports contest between two opposing sides
 - b. an area where sporting events take place
 - c. a piece of equipment used to play a sport
 - d. a type of person who enjoys sports

Figure 3: Ambiguity with multiple choice items

A similar problem arises with polysemy when words have multiple meanings. A test-taker may know one meaning but may not know the meaning of the word being tested. Consider the actual example below taken from the Paul Nation’s VST (Nation & Beglar, 2007).

10. basis: I don't understand the basis.

 - a. reason
 - b. words
 - c. road signs
 - d. main part

Figure 4: Problematic item from VST

Beglar (2010) reports that this test item turned out to be problematic because 60% of the test takers in his study selected the incorrect (*main part*) as the correct answer over the actual correct answer (*reason*). This is because *basis* is polysemous. A *basis* can be a foundation or

main part upon which other things are built. As something upon which other things are established, it can also be the *reason* for doing something. The reverse discrimination of this test item's distractor demonstrates well how polysemy can be problematic when testing vocabulary.

Another problem with context-independent items is that they are viewed as inauthentic. Actual language use always has context, and vocabulary is not generally needed in complete isolation devoid of a social and linguistic environment. “[D]espite the lack of research evidence on the role of context in vocabulary assessment”, context-independent test items have been favored among test designers over context-dependent tests. (Read, 2000, p.101). It has simply been taken on “faith among both language teachers and testers that vocabulary should always be presented in context” (Read, 2000, p.101). Actually, in perhaps the only formal study to examine the effects of context in vocabulary testing, Stalnaker and Kurath (1935) found that there was virtually no difference in the validity of a context-dependent vocabulary test and a context-independent test. They found that the two tests produced analogous results, which were highly correlated to each other and two other cross-measures. This study reported that there were no real advantages to testing vocabulary in context, despite the fact that there has been a good deal of research investigating the effects of context in learning words.

Examples of contextualized vocabulary test types include cloze and c-tests. The standard cloze test consists of a passage of text where every *n*th word is removed and test-takers are asked to insert a suitable word into the blank created. Modified versions of the cloze test also exist with options such as choosing from a word bank or from multiple choices. Rational or selective-deletion cloze tests deliberately select target words to remove from the passage instead of selecting them randomly. With the cloze test, it is difficult to generate contexts where only specific words will be tested or where multiple words are not viable options. C-tests can help

mitigate this concern by either giving part of the word or some letters at the beginning or end of the word to eliminate other possibilities. However, even in c-tests, it is difficult to create items where one and only one response is possible, which makes scoring these test items difficult. An example of one such case can be found later in this chapter in Figure 7.

Context-dependent tests have a number of confounding variables because test-takers must read, listen to, or produce a passage. For one, it is critical that the test-taker know nearly all of the words in the passage that are not deleted in order to understand much of the context. In fact, one study found that for learners to comprehend the general meaning of a text, they must know at least 98% of the words of that text (Hu & Nation, 2000). It is also important that the deleted words in the early part of a passage do not carry too much of the context. If learners do not know these words, it may hinder their abilities to supply words later in the passage even if they do know those words. Because of these elements, contextualized passages are hard to engineer and even harder to create using authentic texts. Besides all of those factors, contextualized vocabulary tests also involve other skills such as composition writing, grammar, and reading, as well as some content knowledge of the topic in the passage outside of their linguistic knowledge. Therefore, if the goal of the test is to test vocabulary alone, contextualized items are less useful, because scores from these types of tests cannot separate the multiple constructs of which they are composed.

Both embedded and comprehensive tests are also often not used in vocabulary size tests because of this type of issue. Embedded tests examine lexical knowledge as a part of reading, writing, speaking, and/or listening skills tests. These types of tests have the same issues as contextualized tests where the variables of vocabulary and reading, writing, speaking, and/or listening cannot easily be separated. Discrete-point tests, on the other hand, focus on isolating

the target lexical item so that knowledge of an individual word can be examined. This is the type of item generally used in vocabulary size tests and includes glossing, translating, matching, and yes/no item types among others.

Glossing can be simply explained as providing a short explanation or definition in either the L1 or L2. Translation is taking the target L2 item and finding a synonymous word in the L1 or vice versa. If these types of items are left with blank spaces to fill in, they can be time-consuming to score by hand and difficult to program in computerized versions because of the sheer number of possibilities, especially with polysemous and homonymous words. Judgments must also be made by graders as to what is deemed as an acceptable answer, especially when the learner seems to exert partial knowledge of the target item. All of these issues make this item type much less practical for tests that must be automatically scored, especially vocabulary size tests. These types of problems can be reconciled by making the test multiple choice or matching; however, these have their own set of problems that will be discussed later in this section.

Multiple choice and matching items are also popular forms of testing vocabulary, especially in classroom settings. Multiple choice items make test-takers select between options for the best answer and includes true/false, which is merely a two-possibility multiple choice question. Matching items take sets of target items and corresponding definitions, glosses, or translations in a random order and make the test-taker match the two together.

Wesche and Paribakht (1996) lists six difficulties with using multiple choice items on vocabulary tests, which are also largely applicable to matching:

1. They are difficult to construct, and require laborious field-testing, analysis and refinement.
2. The learner may know another meaning for the word, but not the one sought.

3. The learner may choose the right word by a process of elimination, and has in any case a 25 percent chance of guessing the correct answer in a four-alternative format.
4. Items may test students' knowledge of distractors rather than their ability to identify an exact meaning of the target word.
5. The learner may miss an item either for lack of knowledge of words or lack of understanding of syntax in the distractors.
6. This format permits only a very limited sampling of the learner's total vocabulary (e.g., a 25-item multiple-choice test samples one word in 400 from a 10 000-word vocabulary. (p. 17).

Haladayna, Downing, and Rodriguez (2002) confirm the difficulty of writing high-quality multiple choice questions in their survey of 27 textbooks and 27 research studies and reviews on the multiple choice format since 1990. Their study simply reiterates the research of Wesche and Paribahkt (1996).

Yes/no or checklist tests are somewhat different from other types of vocabulary tests. A study by Zimmerman, Broder, Shaughnessy, and Underwood (1977) was the first to utilize this item type in its modern form. The yes/no test has the test-takers indicate either “yes” if they know the word or and “no” if they do not know the word. In the checklist version of this test, the test-takers see a series of words and indicate by checking a box next to all of the words that they claim to know. However, the test-taker is told that not all of the words in the test are real words; some of the words are pseudowords which are also called non-words, non-real words, imaginary words, nonsense words, or nonce words. These false words follow the normal orthographic, phonological, morphological, and other norms and tendencies of the language, but they do not contain any real lexical meaning in the language (Meara, 2012). Examples of pseudowords of

English used in the EVST and the VAST are *oxylate*, *galpin*, *bodelate*, *wallage*, *logam* and *retrogradient*. If test-takers claim to know a pseudoword, it is considered a false alarm because they could not possibly know a fabricated lexical item. Scores are based on a calculation from signal detection theory (Zimmerman et al, 1977) where both the total score of known words and the false alarm rate (the percentage of pseudowords marked as real words) are taken into account. The higher the false alarm rate, the lower the score will be. Different formula for calculating scores have been used by different test-designers, which will be discussed later in this chapter (Beeckmans et al, 2001; Huibregtse et al, 2002; Mochida & Harrington, 2006).

Because the test-taker is merely indicating if they recognize the target lexical item as a real or pseudoword, they are only demonstrating their written word recognition. Some researchers have expressed different concerns with the validity of this item type as a measure of vocabulary knowledge. For example, Van Zeeland (2013) called into question the relationship between receptive written word knowledge and receptive oral word knowledge. Using an updated version of the EVST and a corresponding oral yes/no test, the researcher tested ELLs from various L1 backgrounds on both versions and correlated the test scores between the two and found a significant correlation of $r = 0.85$. This shows there being a high degree of overlap between written and oral word recognition. Mochida and Harrington (2006) found that the VLT and yes/no test using the same words correlated at $r = 0.88$ (significant at $p < 0.001$, two-tailed) through regression analyses. They also found that the yes/no test accounted for over 75% of the variance in the overall VLT scores. “The results indicate the Yes/No test is a valid measure of the type of L2 vocabulary knowledge assessed by the VLT” (Mochida & Harrington, 2006). This cross-validation study indicates a great deal of construct validity of yes/no tests because the VLT is the most widely used, studied, and validated vocabulary size test in existence.

Shillaw (1996) examined the construct validity and reliability of the yes/no test through a Rasch-based analysis. A total of 201 students took part in this study and took three different forms of Meara's 60-item 1992 version of the EVST. The results were analyzed twice: once with all of the items included and again with only the real words.

In order to test for unidimensionality and reliability, an item level factor analysis was conducted. According to the author, two conditions are important to meet when proving unidimensionality:

1. The item level factor analysis should show that the first factor accounts for at least 20% of the variance of the unrotated factor matrix
2. The eigenvalue of the first factor is significantly higher than that of the next largest factor.

The results of this analysis are shown in Table 4.

In all cases, the eigenvalues of the first factor were significantly higher than any other factors. Shillaw states that based on the results of the factor analysis, unidimensionality can be assumed for all forms of the test. Shillaw also reports that the test shows moderately high reliability based on the KR-20 results (commonly used to measure test reliability) for all versions of the test.

Item Factor Analysis Results		
Version	1st Factor	KR-20
203	27.78%	0.743
207	29.86%	0.680
218	23.18%	0.743

Table 4: Item factor analysis results

Fit statistics were then calculated for persons and items. In this study, an outfit of ± 2 of either person or item was determined to be misfitting. According to this criterion, 15% of the items and 7% of the subjects misfit. Almost all of the misfitting items were pseudowords. The

author asserted that the overall number of misfitting items and subjects was relatively low even though a conservative criterion for misfit was used. Taken all together, the results of this study show construct validity for this type of item.

Word as a variable.

Modern vocabulary size tests all have the underlying assumption that word frequency generally equates to the order in which words are learned. While this supposition has been shown to generally be true by various validations of vocabulary size tests (Laufer & Nation, 1999; Laufer & Goldstein, 2004; Meara, 1992; Nation, 1983; Nation & Beglar, 2007), close examination of each test item reveals that individual words vary in terms of item difficulty even within 1,000-word frequency bands (Beglar, 2010; Schmit et al, 2001; Shillaw, 1996). Variations in item difficulty from word to word will be discussed more later in this chapter.

West (1953) discusses other factors to consider when designing a word list for language learners (which will be discussed in detail later in this chapter) in addition to frequency and dispersion. These factors include ease/difficulty of learning of words, necessity (words that express things that cannot be expressed through other words), cover (learning words so as to cover the most semantic space), and learning semantically neutral words before learning emotionally charged words.

Nation (1986) also lists several principles that affect learnability of words such as regularity (morphological consistency to the general rules of language), frequency in classroom and teaching settings, and language needs (words that are more personally meaningful and more useful for the intended purpose of the language learner). In short, any number of factors may influence which words learner will tend to learn earlier or later for which frequency and dispersion alone cannot fully account.

Also, Meara (2010) notes that English contains many words which are cognates with Romance languages and Greek. Therefore, in vocabulary size assessments, speakers of these languages have an advantage in knowing words which are cognate with their L1, although, he admits that these cognates are usually low-frequency words. Germanic languages including German, Dutch, Danish, Norwegian, and others also share part of their vocabulary with English as well, which gives them an advantages over speakers of other L1s. Meara (2010) also notes that “dividing a large lexicon into equal size chunks of 1000 words is at best a convenient fiction. And when these chunks are derived from frequency lists which do not represent the real difficulty of words for learner then the risk of distortion is extremely high” (p. 3).

To summarize, a variety of different factors that are both difficult to know and to measure affect the words which language learners will acquire. For this reason, levels of words for vocabulary size tests determined by frequency alone may lack complete construct validity. Because of this, it is important to examine the difficulty of items across bands of frequency to determine the exact relationship between word frequency and item difficulty in a vocabulary size test.

The Construct of a Word

One major issue in all vocabulary studies, especially those addressing vocabulary testing, is defining the construct of a word. Upon first blush, the matter may seem rather simple, but upon closer examination, determining what qualifies as a word is actually quite complex. Depending upon the purpose of the research or educational purpose, different definitions might be used. Gardner (2007) surveyed the body of research concerning the construct of a word. The article asserted that three factors are primarily treated in the research, namely “(a) the degree to which learners of various language backgrounds and skill levels can make connections between

morphologically-related words; (b) the impact of homonymy and polysemy; and (c) the impact of multiword items in the lexicon” (p. 213). The purpose of this section will be to define how these three factors affect the construct of a ‘word’ with respect to vocabulary size testing.

Type and token.

The most elementary definition of what might constitute a ‘word’ is a type. A type is a unique contiguous string of characters in written form. Types are separated by spaces or punctuation in written language. Tokens are a count of the total number of occurrences of types. Thus, in the sentence taken from the poem *Sacred Emily* “Rose is a rose is a rose is a rose.” (Stein, 1913), since there are four occurrences of *rose*, three of *is*, and three of *a*, there are three types (*rose*, *is*, and *a*) and a total of ten tokens. In corpus-based studies, it is easy for a computer to count the number of types and tokens because they are based entirely on word form. However, basing the definition of ‘word’ entirely around form alone is highly problematic because it ignores all three of the factors Gardner lists above. In other words, separating each word by type would mean that *boy* would have to be tested separately from *boys*, that the verb *book* would be tested the same as the noun *book*, and that a *light bulb* would have to be tested as two lexical items. Other constructs of word exist that reconciles these factors to a greater or lesser degree such as word family, lemma, and lexeme.

Word family.

Word family—a base word and all its derived and inflected forms—has been the prevailing construct used in vocabulary size tests, especially the most widely used vocabulary size tests such as the VLT (Nation, 1983), the PVLTL (Laufer & Nation, 1999), the Vocabulary Size Test (VST—Nation & Beglar, 2007), the Computer Adaptive Test of Size and Strength (CATSS—Laufer & Goldstein, 2004), and the EVST (Meara & Buxton, 1987). However, using

word family as a construct by which lexical items on a vocabulary test are grouped is somewhat problematic for a number of reasons. First, there are some problems with how word families are constructed and grouped. Also, studies have shown that morphological acquisition takes time both in the L1 (Berko, 1958; Derwing & Baker, 1979; Nagy, Diakidoy, & Anderson, 1993) and in the L2 (Schmitt & Meara, 1997; Schmitt & Zimmerman, 2002). Studies have also shown that morphological knowledge can vary greatly from learner to learner (Freyd & Baron, 1982; Nagy et al, 1993).

Bauer and Nation (1993) is the best attempt to date to address the variables that exist when trying to group morphologically related words together, at least in English. In an analysis of the Lancaster-Oslo-Bergen (LOB) corpus, they were able to break down word family into seven basic levels which are neither absolute nor discrete and have not been further researched and investigated much beyond the original study. The eight criteria determine the level at which a particular affix was placed.

1. Frequency: the number of words in which an affix occurs.
2. Productivity: the likelihood that the affix will be used to form new words.
3. Predictability: the degree of predictability of the meaning of the affix.
4. Regularity of the written form of the base: the predictability of change of the written form of the base when the affix is added.
5. Regularity of the spoken form of the base: the amount of change of the spoken form of the base when the affix is added.
6. Regularity of the spelling of the affix: the predictability of written forms of the affix.
7. Regularity of the spoken form of the affix: the predictability of spoken forms of the affix.

8. Regularity of function: the degree to which the affix attaches to a base of known form-class and produces a word of known form-class. (p. 255-256)

Based on these criteria, Bauer and Nation (1993) defined seven levels of word family morphology. These seven levels are summarized in their article as follows:

Level 1: *Each form is a different word.* The pessimistic view that learners will not recognize any morphological relationship between words. However, even at this lowest level, the concepts of homonymy and polysemy are ignored.

Level 2: *Inflectional suffixes.* Words with the same base and inflections are considered part of the same family. This level assumes that learners can recognize perform “minimal morphographemic analysis in order to recognize regular inflections” (p. 258). In English these would be considered the same as lemma, but for languages with inflectional affixes that are not suffixes they would not. [Plural *-s*, comparative *-er*, and genitive *-’s* would fall at this level.]

Level 3: *The most frequent and regular derivational affixes.* All of the eight criteria above are applied strictly to this level. Only the affixes *-able*, *-er*, *-ish*, *less*, *-ly*, *-ness*, *-th*, *-y*, *non-*, and *un-* are included at this level.

Level 4: *Frequent, orthographically regular affixes.* All eight of the above criteria are still applied at this level. However, frequency is prioritized over productivity and orthography over phonology at this level. Morphemes at this level include *-al*, *-ation*, *-ess*, *-ful*, *-ism*, *-ist*, *-ity*, *-ize*, *-ment*, *-ous*, *in-*. The meanings of these affixes was deemed by authors to be generalizable or easily used on a range of words.

Level 5: *Regular but infrequent affixes.* This adds the rest of the affixes whose form, meaning, and function are regular. Only affixes added to free bases are in this level.

These include 50 affixes, including *-age, -al, -an, -ance, -ite, -let, -ling, -wise, circum-, counter-, and semi-*.

Level 6: *Frequent but irregular affixes*. This level adds affixes that have orthographic allomorphy in their bases or are difficult to segment because of homography. These are *-able, -ee, -ic, -ify, -ion, -ist, -ition, -ive, -th, -y, pre-, and re-*.

Level 7: *Classical roots and affixes*. This level includes all the classical roots which occur as bound roots in English (e.g. embolism) and neo-classical compounds (e.g. photography). Affixes added to bound bases are in this level as well. This level also includes frequent prefixes, such as *ab-, ad-, com-, de-, dis-, ex-, and sub-* (Bauer & Nation, 1993, p. 258-262).

To date, there has been no analysis of these levels to ascertain if they reflect reality for either L1 or L2 learners. Also, no researcher has yet to offer an alternative method for classifying and defining word family. Additionally, there have been no articles yet written about “levels” of word families in languages other than English, what they would look like, or if they would be the same as in English. Because the levels are untested even in English, and only loosely based on morphology acquisition theory, it seems improbable, although possible, that these same definitions of levels of word family would hold up across all languages.

The VST explains that it uses level 6 of these words families, but nowhere in the literature explaining the creation of PVLTL, VLT, CATSS, or EVST does it discuss what level of word family is being used to group together words for frequency counts. However, regardless of whatever level is being used, certain research findings have not been taken into account relating to the acquisition of morphological knowledge. A number of studies have addressed morphological acquisition related to inflectional and derivation morphology in the L1 and L2.

For English L1 learners, Berko (1958) established that acquisition of inflectional morphology is generally underway around first grade (six or seven years old). However, derivational morphology begins sometime before fourth grade (nine or ten years old), continues through middle school (Carlisle, 2000; Nagy, 1989), and proceeds into high school years and beyond (Nagy et al, 1993).

For English L2 learners, too, several studies have also been undertaken to understand inflectional and derivational morphological acquisition. Schmitt and Meara (1997) and Schmitt and Zimmerman (2002) found parallel results indicating that just as with L1 learners, L2 learners acquire inflectional morphology before derivational morphology and that their more than 200 non-native participants showed a more profound awareness and mastery of inflections over derivations as well. The authors concluded that without explicit instruction in morphology and word formation, it is unlikely learners acquire derivational morphology simply by exposure.

Some studies have also noted that morphological knowledge between learners contains observable variations depending on language skill level. Freyd and Baron (1982) observed that advanced fifth graders (nine and ten years old) outperformed average eighth graders (13 years old) in tests of morphological awareness. Studies have also shown that morphological awareness is higher among superior readers (Freyd & Baron, 1982; Nagy et al, 1993; Tyler and Nagy, 1989). Although these studies were done for L1 learners, an assumption can be made that individual differences are possibly a large factor in determining morphological mastery for L2 learners. Future research in this area is needed.

Gardner (2007) also points out several other issues with word families being used as the construct of word for language learners. First, researchers group derivational prefixes and suffixes together despite the fact that these two may present different dilemmas for developing

language learners (Nagy et al, 1993). Second, derivational suffixes, in particular, (e.g. *-ment*, *-ness*, *-ish*) are difficult to define and to grasp. Third, some learners will focus only on the stems of derived forms and will ignore suffixes that they do not understand (Freyd & Baron, 1982). Fourth, some learners will only recognize derived forms as being related to the base form after they learn to recognize the stems of those derived forms (Tyler & Nagy, 1989). Finally, Biemiller and Slonim (2001) found that some learners may acquire derived forms before they acquire their root-form counterparts.

Read (2000) also points out some complexities dealing with word families. For example, the word *society* encompasses a rather large word family: *social*, *socially*, *sociable*, *unsociable*, *sociability*, *socialize*, *socialization*, *socialism*, *socialist*, *socialite*, *sociology*, *sociologist*, *sociological*, *societal*, *sociopath*, *sociopathy*, *sociopathic*, and so on and so on. “These words all share the same *soci-* form and seem to have a common underlying meaning which might be expressed as ‘relating to the organisms in a group or organization’,” but “since the words express quite a range of meanings, we would not want to say that they are members of the same word family” in language testing (p. 19). Yet, that is precisely what the majority of vocabulary size tests have done.

Some studies have investigated the number of individual words that word families actually represent. Brezina and Gablasova (2015) determined that the General Service List’s 2,000 word families represent about 4,100 lemmas. The GSL is the list that is used as the first frequency list for the VLT, PVLt, and CATSS (Nation, 1983; Laufer & Nation, 1999; Laufer & Goldstein, 2004). Therefore, vocabulary size tests up to this point may have drastically misjudged the number of words learners actually know.

Lemma.

Lemmas are not commonly used in studies about vocabulary size, but at least a couple of studies have opted to use lemma as the construct of interest over word families (Ishii & Schmitt, 2009; Milton & Hopkins, 2006). Like a word family, a lemma is another form-based concept of word. A lemma encompasses a base form and its inflection, even irregular inflections, and all the words counted under a lemma must be the same part of speech (Kučera & Francis, 1967).

Using lemma as the construct of word for language learners “poses serious quandaries relating to the psychological validity of such family relationships—namely, that the opaque spelling and phonological connections between the lemma headword and the family members will surely cause more and different learning problems than their more transparent counterparts” in irregular verbs, although this is not necessarily the case with regular verbs (Gardner, 2007, p. 248).

While word family totally ignores the issue of homonymy and polysemy, grouping all morphologically related words under a single umbrella, lemma does not. Because everything grouped within a lemma must be the same part of speech, some cases of homonymy is eliminated when changing from word families to lemmas. For example, *light* as an adjective is categorized under a separate lexical unit than *light* as a noun or even *light* as a verb or an adverb. Even so, according to one semantic count, *light* has 25 senses as an adjective, 15 senses as a noun, seven senses as a verb, and one as an adverb (WordNet, 2003). Just an observation of some of the WordNet meanings of *light* as an adjective reveals a variety of meanings: of comparatively little physical weight (e.g., *a light load*), having a small amount of coloring agent (e.g., *light blue*), characterized by or emitting light (e.g., *the room is light when the shutters are open*), demanding little effort (e.g., *light housework*, *light exercise*), wakeful (e.g. *light sleeper*),

low in calories (e.g., *light beer*), having little importance (e.g., *light banter*, *light matter*), etc.

This issue of lexical ambiguity is addressed further in the next section.

Homonymy and Polysemy.

Vocabulary size assessments are developed based on word counts done by corpus linguists who are somewhat restricted in how they calculate frequency counts based on written word form. Problems arise because the focus of these counts is often only word form and not word meaning. Vocabulary size tests and often other types of vocabulary tests are built under the false notion that written word forms are all the same. One leading scholar notes that “each word form may represent a number of distinct meanings, some of which depend strongly on the reading context, and some of which are quite different from each other in meaning” (Grabe, 1991, p. 392).

Gardner (2007) elaborates upon this point with the example of the English word ‘bear’:

Forms that appear to be related through affixation may actually be homographs in context (e.g. *bear*, the animal, and *bears/bearing*, the verb meaning to carry), or repetitions of the same affixed forms may actually be homographs in context (e.g. *bears*, the plural animal, and *bears*, the verb meaning to carry). The existence of potential polysemes seems to complicate matters even more (e.g. *bear/bears/bearing*, the verb meaning ‘to move while holding up and supporting,’ and *bear/bears/bearing*, the verb meaning ‘to hold in the mind’—Webster’s Ninth New Collegiate Dictionary, 1988). To make matters even worse, the past form of the verb *bear* (either definition above) is *bore*, which bears (additional meaning) no orthographic resemblance to the other forms in the same semantic family. Additionally, the form *bore* itself has much more common meanings that are totally unrelated to *bear* in the senses described above (*to bore a hole*, *the man is*

a bore, I don't want to bore you with this, etc.). A quick search in WordNet (2003) reveals two senses for the noun *bear*, 13 senses for the verb *bear*, two additional senses for the verb *bore*, and four senses for the noun *bore*—a total of 21 different meaning senses for the two forms *bear* and *bore*. This does not even include the two adjective senses of *born* which could also be included in the *bear/bore* family.

Conceivably, a machine-based frequency count of word-family forms could link all of these forms of *bear* and *bore* together (assuming that the researcher had determined to count *bore* under *bear*), but the question remains whether or not a child or other language learner encountering these various forms would make any semantic connections between them based on context. (p. 251-252)

This type of occurrence is particularly apparent in high-frequency words (e.g. *break*—72 senses, *get*—37 senses, *show*—16 sense, etc.—WordNet, 2003). This is notable because vocabulary size tests are built around the most frequent words that exist in languages. Indeed, Ravin & Leacock (2000) asserts that “the most commonly used words tend to be the most polysemous” (p. 1). Gardner (2007) argues that basing findings of corpus research solely on frequency without respect to word meanings would invariably lead to three types of problems: “(a) they will overestimate the true coverage of the word forms; (b) they will underestimate the actual user knowledge required to negotiate word forms; and/or (c) they will underestimate the actual number of meanings inherent in word forms.” These same problems can be assumed in vocabulary size testing where frequency counts in corpora are taken without carefully considering the issues that come with word meaning.

Research into lexical ambiguity has generally concluded that homonymous words have all their senses stored separately in the mental lexicon (Klepousnaitotou, 2002). However, the

way polysemy is catalogued is still under debate and exploration. An explanation of the various theories of the processing of polysemy in the mental lexicon is not necessary here because corpus linguistics does not yet have a systematic way of computing corpora and generating frequency lists based around each individual meaning of a polysemous word (Sinclair, 2004). Suffice it to say that linguists agree that modeling such processes is not a simple endeavor (Klepousnaitou, 2002), and that future vocabulary size tests will need to address this issue in greater depth.

Lexeme.

All of the constructs of word which have thus far been discussed (i.e., type, word family, lemma) have methodological issues if they are to be used with vocabulary testing. Because of differing constructs that are used to group word forms and separating word meanings, vocabulary size estimates have varied. Ideally, vocabulary size tests would test a learner's knowledge of all meanings of words. In other words, future vocabulary size tests should focus on using lexeme as the construct for grouping form and separating meaning (Gardner, 2013; Read, 2000). Biber, Johansson, Leech, Conrad, Finegan, and Quirk (1999) define a lexeme as "a group of word forms that share the same basic meaning (apart from that associated with the inflections that distinguish them) and belong to the same word class (p. 54). Because semantically tagged corpora large enough to generate satisfactory frequency lists are not available, machine-based frequency counts of lemma are currently the best option in terms of mitigating the problems of homonymy and polysemy.

Multiword and other lexical items.

Computerized corpora have allowed linguists to look at language in dynamic ways including the way certain words cluster together in lexical collocations (Sinclair, 1991).

Observation of collocations by corpus linguists have revealed the multiword item which “is a vocabulary item which consists of a sequence of two or more words (a word being simply an orthographic unit). This sequence of words semantically and/or syntactically forms a meaningful and inseparable unit” (Moon, 1997, p. 43). They are also known as formulaic language or formulaic sequences. As mentioned in the introduction, these include compounds, phrasal verbs, idioms, fixed phrases, and prefabs.

These multiword items cover a substantial portion of the English language. In fact, in an analysis of the London Lund Corpus of Spoken English (LLC), the LOB corpus, and two versions of a short fiction text, Erman and Warren (2000) estimated that multiword items account for 58.6 percent of spoken English and 52.3 percent of written English. Gardner (2007) goes so far as to suggest that “[t]here is convergent evidence that the sheer number of different multiword items may exceed the number of individual words in the lexicon” (p. 255). Also, just as other types of lexical units, multiword items can be polysemous (e.g. *broke out* in a cold sweat vs the virus *broke out* across the city) and homonymous (*break in* a new pair of shoes vs. *break in* through the back window). The percentages of coverage of these types of words found by Erman and Warren (2000) make them impossible to ignore as a serious factor in how corpus linguists count, instructors teach, and psychometricians test lexical items. However, of the major tests of vocabulary size of English (i.e. VLT, PVL, VST, CATSS, EVST) only the EVST has used any multiword item in its test, and even then, only a small number of phrasal verbs are used (Nation, 1983; Laufer & Nation, 1999; Nation & Beglar, 2007; Laufer & Goldstein, 2004; Meara, 1992). Read (2000) points out that with multiword items, “the whole unit has a meaning that cannot be worked out just from knowing what the individual words mean. Such multi-word items have long been accepted as part of the vocabulary learning task that students face (p. 21)”.

If realistic approximation of the total vocabulary size of language learners is to be determined, these types of words should be included both in the frequency counts taken from corpora and the tests themselves as ways to count them in corpora become available.

Other types of words that have also been left out of vocabulary size include proper nouns, interjections, abbreviations, and acronyms. No attempt at a theoretical explanation is given as to why these words are excluded; they simply are (Goulden, Nation, & Read, 1990, Meara & Buxton, 1987, Nation, 1983, Laufer & Goldstein, 2004). For example, Laufer (1998) states “proper nouns are omitted (they are not considered to as belonging to the lexis of a given language” (p. 261). But, why not? Surely, learning that *Japan* is *Nippon* or *Nihon* in Japanese or *Ilbon* in Korean or *James* is *Jacques* in French or *Jaime*, *Jacobo*, *Santiago*, or *Iago* in Spanish requires as much from learners as do some other words that are not proper nouns such as cognates. Perhaps the only situation to exclude proper nouns would be in yes/no tests. It would be difficult, perhaps impossible, to create pseudowords to mimic proper nouns. However, formal research needs to be undertaken in this area.

Indeed, several examples of proper nouns, interjections, and acronyms can be found just in the first 5,000 lemma of English taken from the Corpus of Contemporary American English (COCA). Examples of each of these type of words taken from this corpus can be found in the figure below. Perhaps previous studies have excluded these words because the researchers found such lexical items fundamentally different from the words they chose to include in some way. It might well be that there is a critical difference in the way proper nouns are learned. However, for ELLs, there is no existing evidence to support a claim that these types of words are fundamentally different from other types of words in how they are acquired or cognitively stored. Therefore, research needs to be conducting into investigating whether or not these types

of words are learned or stored in the mental lexicon differently from other types of words.

Otherwise, they should not be excluded from vocabulary size tests.

Types of Words Excluded from Previous Vocabulary Size Tests									
<u>Proper Nouns</u>									
Rank	Lemma	POS	Tokens	Dispersion	Rank	Lemma	POS	Tokens	Dispersion
176	American	adjective	214968	0.95	3186	Christian	noun	9751	0.93
544	American	noun	73063	0.95	3189	German	noun	9586	0.94
594	Republican	noun	71611	0.88	3196	God	noun	9694	0.93
638	Congress	noun	62841	0.92	3333	African-American	adjective	9121	0.93
904	Democrat	noun	46905	0.88	3514	T-shirt	noun	8386	0.94
1112	Senate	noun	36809	0.91	3531	Roman	adjective	8299	0.94
1184	British	adjective	32929	0.95	3538	Muslim	noun	8498	0.91
1205	African	adjective	34557	0.89	3539	Hispanic	adjective	8690	0.89
1220	Chinese	adjective	32334	0.94	3546	Korean	adjective	8441	0.92
1250	Soviet	adjective	36193	0.82	3728	European	noun	7688	0.93
1267	European	adjective	31455	0.92	3797	Olympics	noun	8039	0.87
1275	Christian	adjective	30726	0.94	3882	Japanese	noun	7419	0.91
1366	French	adjective	27590	0.96	3910	Israeli	noun	7673	0.87
1440	Indian	adjective	27100	0.92	3988	Arab	noun	7222	0.9
1464	United	adjective	26396	0.93	4026	Russian	noun	6891	0.93
1471	Internet	noun	26983	0.9	4091	Dutch	adjective	6690	0.94
1567	Russian	adjective	23739	0.94	4095	Greek	adjective	6642	0.94
1623	Supreme	adjective	23904	0.9	4223	Cuban	adjective	6601	0.9
1657	Iraqi	adjective	25446	0.82	4505	Thanksgiving	noun	5859	0.92
1726	Japanese	adjective	21800	0.92	4567	Persian	adjective	6340	0.84
1767	Catholic	adjective	20866	0.94					
1792	English	adjective	20235	0.95	<u>Acronyms</u>				
Rank	Lemma	POS	Tokens	Dispersion	Rank	Lemma	POS	Tokens	Dispersion
1805	Jewish	adjective	20196	0.94	825	PM	adverb	54765	0.82
1833	German	adjective	20096	0.93	1194	AM	adverb	34451	0.9
1913	English	noun	18719	0.96	2846	DNA	noun	11580	0.92
2096	Israeli	adjective	17967	0.89	3072	PC	noun	11072	0.86
2215	Arab	adjective	16732	0.88					
2218	Spanish	adjective	15512	0.95					
2262	Mexican	adjective	15514	0.93	<u>Interjections</u>				
Rank	Lemma	POS	Tokens	Dispersion	Rank	Lemma	POS	Tokens	Dispersion
2332	Indian	noun	15021	0.92	258	yes	interjection	157364	0.89
2366	Asian	adjective	14873	0.92	411	oh	interjection	103613	0.89
2542	Latin	adjective	13797	0.9	429	yeah	interjection	103389	0.84
2548	Palestinian	adjective	14008	0.88	914	no	interjection	44951	0.91
2608	Muslim	adjective	13147	0.92	1433	hey	interjection	27659	0.9
2612	Islamic	adjective	13323	0.9	2083	hi	interjection	18910	0.85
2653	Italian	adjective	12384	0.95	2252	hello	interjection	16600	0.87
2654	Canadian	adjective	12820	0.91	2961	mm-hmm	interjection	13755	0.73
2753	Olympic	adjective	13072	0.85	3277	ah	interjection	9788	0.89
2759	Bible	noun	11539	0.96	3804	wow	interjection	8016	0.87
2947	Irish	adjective	10833	0.94	3956	huh	interjection	7563	0.87
3092	French	noun	9845	0.96	4519	uh	interjection	6155	0.87
3121	Catholic	noun	9955	0.94					
*POS=Part of Speech									

Table 5: Types of Words Excluded from Previous Vocabulary Size Tests

In addition to potential issues in words that have been excluded, there is also the potential for issues with words that have been included in vocabulary size tests. Read (2000) points out that function words are seen as belonging more to the grammar of the language than to its vocabulary. He points out that unlike content words, function words have little or no meaning in isolation and serve to link or modify other words rather than carry lexical content. Yet, all vocabulary size tests have included function words in their frequency lists and approximations of learners' total word knowledge counts even though the VLT, PVL, VST, CATSS, and EVST do not include function words as test items. 60 percent of running English speech consists of merely 50 function words (Davies & Gardner, 2010). Other studies confirm that function words are generally high-frequency words in English (Laufer & Nation, 1999; Laufer & Goldstein, 2004; Meara, 1992; Nation, 1983; Nation & Beglar, 2007). However, no study of vocabulary size tests has yet to observe the differences in the items containing function words and the items containing content words to examine if two types of words behave differently. If such evidence were found, perhaps it would be better for future vocabulary lists based on word frequency to exclude such words as vocabulary items and instead to include them in grammar materials.

The research discussed in this section has exposed several important considerations for the construct of word for vocabulary size testing. First, some word forms should and should not be grouped together depending on the demographic of the test-taker. Second, word forms may have numerous meanings which should be tested separately. Third, test makers should explain and test-takers should understand which construct of word is being used in vocabulary size tests in order to appropriately interpret the results. And, lastly, researchers should consider what words are important to include and exclude when measuring the size of a learner's lexicon.

Generating Corpora and Word Lists

The creation and usage of well-designed corpora is essential for the generation of quality word lists which are used in vocabulary size tests. Without understanding the composition of the corpus and the method by which it was tagged and/or parsed, linguists can make incorrect assumptions about lists based on frequency which are generated therefrom. With the advent of the computer, corpus linguistics has allowed researchers to compile lists of words in a fraction of the amount of time it took Thorndike and Lorge (1944) and their colleagues to manually count and tag the 18,000,000 words of their corpus. However, simply gathering huge amounts of text and counting word forms will not generate perfect word frequency lists. There are many factors that go into creating a corpus that is suitable for generating frequency lists, and “[a] frequency [list] is only as good as the corpus on which it is based” (Davies & Gardner, 2010, p. 3).

Factors for frequency list creation.

Nation and Waring (1997) determines six key factors (summarized below) that need to be considered in the development of a list of high-frequency words to be used by language teachers and learners and which will be used to evaluate lists in this chapter:

1. Representativeness: The corpora that the list is based on should adequately represent the wide range of uses of language. In the past, most word lists have been based on written corpora. There needs to be a substantial spoken corpus involved in the development of a general service list. The spoken and written corpora used should also cover a range of representative text types. Biber's (1990) corpus studies have shown how particular language features cluster in particular text types. The corpora used should contain a wide range of useful types so that the biases of a particular text type do not unduly influence the resulting list.

2. Frequency and range [also known as dispersion]: Most frequency studies have given recognition to the importance of range of occurrence. A word should not become part of a general service list because it occurs frequently. It should occur frequently across a wide range of texts. This does not mean that its frequency has to be roughly the same across the different texts, but means that it should occur in some form or other in most of the different texts or groupings of texts.

3. Word families[/construct of word]: The development of a general service list needs to make use of a sensible set of criteria regarding what forms and uses are counted as being members of the same family. Should *governor* be counted as part of the word family represented by *govern*? When making this decision, the purposes of the list and the learners for which it is intended need to be considered. As well as basing the decision on features such as regularity, productivity, and frequency (Bauer & Nation, 1993), the likelihood of learners seeing these relationships needs to be considered (Nagy & Anderson, 1984).

4. Idioms and set expressions[/multiword items]: Some items larger than a word behave like high frequency words. That is, they occur frequently as a unit (*Good morning, Never mind*), and their meaning is not clear from the meaning of the parts (*at once, set out*). If the frequency of such items is high enough to get them into a general service list in direct competition with single words, then perhaps they should be there. Certainly the arguments for idioms are strong, whereas set expressions could be included under one of their constituent words (but see Nagy, this volume, [1997]).

5. Range of information: To be of full use in course design, a list of high frequency words would need to include the following information for each word—the forms and parts of

speech included in a word family, frequency, the underlying meaning of the word, variations of meaning and collocations and the relative frequency of these meanings and uses, and restrictions on the use of the word with regard to politeness, geographical distribution etc. Some dictionaries, notably the revised edition of the COBUILD dictionary, include much of this information, but still do not go far enough. This variety of information needs to be set out in a way that is readily accessible to teachers and learners.

6. Other criteria: West (1953: ix) found that frequency and range alone were not sufficient criteria for deciding what goes into a word list designed for teaching purposes. West made use of ease or difficulty of learning (it is easier to learn another related meaning for a known word than to learn another word), necessity (words that express ideas that cannot be expressed through other words), cover (preference for learning words to span the most semantic space), stylistic level, and emotional words (West saw second language learners as initially needing neutral vocabulary). One of the many interesting findings of the COBUILD project was that different forms of a word often behave in different ways, taking their own set of collocates and expressing different shades of meaning (Sinclair, 1991). Careful consideration would need to be given to these and other criteria in the final stages of making a general service list. (p. 11-12).

These six criteria are all important for generating frequency lists for language learners. However, for vocabulary size tests, the requirements are slightly different. For one thing, because the list is not directly being used for course design or study, range of information is not necessary. For another, because the list is designed for teaching purposes and not for testing purposes, some of the “other criteria” listed above are not as relevant—such as necessity,

stylistic level and emotional words, and consideration of collocation. Also, the factor of representativeness should prioritize contemporary over antiquated language. One criticism of the continued usage of the GSL and other lists in contemporary language teaching/testing is that they are based on corpora that are simply too old (Brezina & Gablasova, 2015; Gardner & Davies, 2014). Finally, the corpus must have enough tokens to obtain counts that are above statistically significant above chance.

Read (2000) agrees with many of these criteria noting that high-frequency word lists should be generated from a large multi-million-word corpus that is representative of a variety of genres (spoken, written, etc.), varieties (British, American, etc.), and registers (narrative, expository, etc.). The list should be lemmatized, words of a single lemma all counted under one headword and take into account dispersion in addition to frequency. He suggests multiword items should be included, and that if possible, semantic analysis should be performed to determine if lemmas really should be grouped into word families.

Using these criteria as a guide, corpora that have been used to generate frequency lists for the VLT, PVLVT, VST, CATSS, and EVST will now be evaluated.

Lists used in vocabulary size tests of English.

The researchers who developed the VLT, PLVT, and CATSS are mostly the same, and thus, these tests share the same word lists (Laufer & Goldstein, 2004; Laufer & Nation, 1999; Nation, 1983). The word frequency data that was employed in these tests came from a number of different resources, but in the generation of the word lists from these corpora, all of them were grouped together according to word family. Thorndike and Lorge's list originally categorized words somewhat based on lemma (Thorndike & Lorge, 1944). However, its frequency data was converted into word families (Nation, 1983).

As far as the words chosen for the test themselves, the VLT chose to exclude proper nouns, multiword units, and any derivational form that had a more frequent form represented in its word family. The words were hand-picked by the researcher so as to be “representative of all the words at that level” (Nation, 1983, p. 14). However, what criteria were used to determine what is “representative” is never fully explained. This is problematic because the decisions about what words are included in these tests then become highly subjective according to the test designer, and it ignores the fact that a word level of 1,000 words might have words that vary vastly in terms of difficulty for test-takers.

The construct of word for these tests, the VLT, and the EVST both fall under the same paradigm of word family. None make any efforts to distinguish between words based on either homonymy or polysemy, including any differentiation based on part of speech. These types of words are grouped together in their frequency data by word family. The most frequent derived form of the word in the word family is the one which shows up in the word list and test (Schmitt, Schmitt, & Clapham, 2001).

The process for creating the composite word list used for the VLT, PVL, and CATSS was quite complex. Three corpora were used to generate four word lists that were then recombined into a single list. The lists used for these tests were derived from lists that were somewhat dated even at the times of their respective creations, including Thorndike and Lorge’s (1944) *The Teacher’s Word Book of 30,000 Words*, the General Service List (West, 1953), Kučera & Francis’ word list (1967), and Campion and Elley’s Academic Vocabulary List (AVL—1971). Thorndike and Lorge’s list was based on an 18-million-word corpus of American English (1944); the General Service List was derived from the *Teacher’s Word Book of 30,000 Words* after some processing of the data (West, 1953); Kučera and Francis (1967) was based on

the 1-million-word BROWN Corpus; the AVL by Campion and Elley was based on a corpus of university textbooks (1971).

Nation (1983; 1990) vaguely describes the process for how the lists were combined to create a single test. Each of these lists were cross-checked against one another to ensure that there was no cross-over or duplicate information. The process for determining which words went at which level was to start with the GSL which was used for the first 2,000 words. Then, based on the combined corpus of *the Teacher's Word Book of 30,000 Words* and the Kučera and Francis' list, words were added according to their frequency if they did not appear in the GSL for the rest of the 10,000 words that composed the frequency list. Finally, the words from the AVL were added to comprise the University Word Level of the test after checking to make sure no words were duplicated across this word list and the other word lists. The University Word Level is known in some publications also as the Academic Word Level.

Nation and Beglar (2007) explains that the first 2,000 words were taken from the GSL. The remaining 12,000 of the 14,000 words for the VST were derived from the 10 million token spoken section of the British National Corpus (BNC). Also, although the VST is a written test, the researchers opted to use spoken language because they deemed the BNC written section to be too formal (p. 11).

Meara (2010) details some of the methodological choices made in the EVST, later rebuilt and retitled as the EFL Vocabulary Tests, X_Lex, and V_YesNo. For the intents of this thesis, these tests will be grouped under the name EVST. The original EVST from 1987 also made use of the *Teacher's Word Book of 30,000 Words* (P. Meara, personal communication, March 1, 2016). The words for 1992 iteration were updated with two sources. The first 2,000 words were updated from Paul Nation's *Vocabulary lists: words, affixes, and stems* (Nation, 1986) which

was partially derived from Champion and Elley's (1971) list and partly from lists from Praninskas (1972), Lynn (1973), and Ghadessy (1979). These lists are all based on academic language mostly from textbooks or annotations in textbooks. Words 3,000-5,000 came from Hindmarsh's *Cambridge English Lexicon* (Hindmarsh, 1980), which was based on a variety of other frequency counts. Table 9 on page 60 compares these lists against the word list used for this thesis, the Corpus of Contemporary American English. Among the weaknesses of the corpora and lists used for previously designed vocabulary size tests are the following:

1. Some are outdated (as dated as 70 years old).
2. Some are too small. (as few as two million tokens).
3. All of them lack range across a wide variety of genres. None include both written and spoken genres.
4. None of them used formal dispersion statistics to generate their respective word lists.

The next section will expand further upon each of these four points and how COCA has strengths in these areas where previously used corpora did not. Again, having an appropriate corpus strengthens the argument of construct validity for the VAST because it more accurately reflects the type of modern and general English that the learners are likely to have encountered while acquiring vocabulary.

The Corpus of Contemporary American English word frequency list.

Considering the nature of lists that have been used for previous tests of vocabulary size in English and the corpora upon which they are based, one must question the validity of previously designed vocabulary size assessment as accurate assessments of modern language. Therefore, one of the foci of this thesis is to compare a list that may perhaps be more suitable for vocabulary size tests against those that have already been used. Therefore, the Corpus of Contemporary

American English and the word list derived therefrom will be compared to those used in previous tests of vocabulary size. This comparison will be based on the same criteria used to evaluate those previously mentioned (i.e., size of corpus, size of list, representativeness, dispersion, and construct of word).

Size of corpus.

Although the size of COCA has recently been expanded to 520 million tokens, at the time of retrieval of the frequency data used for this thesis, the corpus was 450 million tokens (Davies, 2008). The size of this corpus is 20 times the size of the other corpora used in English vocabulary size tests shown in the table below.

Comparison of Corpora Size and Published Year		
<u>Corpus</u>	<u>Published (Year)</u>	<u># of Tokens</u>
Teacher's Word Book of 30,000 Words	1944	18 million
The General Service List	1953	23+ million
Kučera and Francis' Word List	1967	1 million
Campion and Elley's Academic Vocabulary List	1971	300 thousand
British National Corpus Spoken	1995	10 million
Vocabulary Lists: Words, Affixes, and Stems	1986	2 million
Cambridge English Lexicon	1980	18 million
Corpus of Contemporary American English	2008	450 million

Table 6: Comparison of corpora size and published year

Size of list.

The frequency list derived from COCA contains 60,000 words (Davies, 2011).

Representativeness.

Davies (2008; 2009) describe the architecture in great detail included the sources of all 190,000 texts totaling at 450 million tokens. These texts span five genres in equal proportions: spoken (20%), fiction (20%), magazine (20%), newspaper (20%), and academic (20%). The corpus is entirely contemporary language taken from 1990-2015 with each year accounting for

about 20 million tokens. Each genre accounts for about 4 million of the yearly tokens. The chart below shows the breakdown of the composition of the corpus by genre and year. Because the frequency list generated from this corpus only contains data up through 2011, only that information is relevant for the analysis here. The corpus has since been updated to include language up to 2015.

COCA Composition						
Year	Spoken	Fiction	Magazine	Newspaper	Academic	Total
1990	4,332,983	4,176,786	4,061,059	4,072,572	3,943,968	20,587,368
1991	4,275,641	4,152,690	4,170,022	4,075,636	4,011,142	20,685,131
1992	4,493,738	3,862,984	4,359,784	4,060,218	3,988,593	20,765,317
1993	4,449,330	3,936,880	4,318,256	4,117,294	4,109,914	20,931,674
1994	4,416,223	4,128,691	4,360,184	4,116,061	4,008,481	21,029,640
1995	4,506,463	3,925,121	4,355,396	4,086,909	3,978,437	20,852,326
1996	4,060,792	3,938,742	4,348,339	4,062,397	4,070,075	20,480,345
1997	3,874,976	3,750,256	4,330,117	4,114,733	4,378,426	20,448,508
1998	4,424,874	3,754,334	4,353,187	4,096,829	4,070,949	20,700,173
1999	4,417,997	4,130,984	4,353,229	4,079,926	3,983,704	20,965,840
2000	4,414,772	3,925,331	4,353,049	4,034,817	4,053,691	20,781,660
2001	3,987,514	3,869,790	4,262,503	4,066,589	3,924,911	20,111,307
2002	4,329,856	3,745,852	4,279,955	4,085,554	4,014,495	20,455,712
2003	4,404,978	4,094,865	4,295,543	4,022,457	4,007,927	20,825,770
2004	4,330,018	4,076,462	4,300,735	4,084,584	3,974,453	20,766,252
2005	4,396,030	4,075,210	4,328,642	4,089,168	3,890,318	20,779,368
2006	4,304,513	4,081,287	4,279,043	4,085,757	4,028,620	20,779,220
2007	3,882,586	4,028,998	4,185,161	3,975,474	4,267,452	20,339,671
2008	3,635,622	4,155,298	4,205,477	4,031,769	4,015,545	20,043,711
2009	3,969,587	4,143,814	3,855,815	3,971,607	4,144,064	20,084,887
2010	4,095,393	3,929,160	3,806,011	4,258,633	3,816,420	19,905,617
2011	4,033,627	4,166,029	4,199,378	3,982,299	4,064,535	20,445,868
Total	93,037,513	88,049,564	93,360,885	89,571,283	88,746,120	452,765,365

Table 7: COCA composition

The spoken genre contains over 93 million tokens and originates from 150+ different television and radio programs from a wide variety of subgenres in 37,757 texts. Because the corpus designer wanted to have a fifth of the corpus to come from spoken American English, it would have been a nearly impossible task to create a corpus large enough from transcribing tape recorded spoken texts, as was done for the BNC spoken subcorpus. This raises three important questions about the spoken texts used in COCA which are addressed in Davies (2008), namely:

1. Do they faithfully represent the actual conversations?
2. Is the conversation really unscripted?
3. How well does it represent “non-media” varieties of spoken American English?

In response to the first question, the transcripts are done quite comprehensively including everything, including the interruption, false starts, etc. The second question is answered by the nature of the programs from which the texts were derived. All of the spoken texts used for this corpus come from “unscripted programs” such as *All Things Considered* (NPR), *Newshour* (PBS), *Good Morning America* (ABC), *Today Show* (NBC), *60 Minutes* (CBS), *Hannity and Colmes* (Fox), *Jerry Springer*, *Oprah*, etc. Upon close examination of the transcripts found in the corpus, the vast majority of the language is unscripted. However, a small percentage, Davies (2008) estimates about 2-5% of the texts consist of “formulaic/scripted” sentences such as “Welcome to [the name of the program]” or “We’ll be right back to you after a brief commercial break”. The third question about how well the language represents natural, “non-media” language is somewhat difficult to answer empirically. Because people know they are on a broadcasted program, they are likely to alter their speech accordingly. This is manifest in there being “relatively little profanity” and few “highly stigmatized words and phrases such as ‘ain’t got none’”, but “[i]n terms of overall word choice and ‘natural conversation’ (false starts,

interruptions, and so on), though it does seem to represent ‘off the air’ conversation quite nicely” (Davies, 2008). Davies (2008) goes on to make the very valid point that, indeed, “*no spoken corpus* (even those created by linguists with tape records in the early 1990s) will be 100% authentic for real conversation—as long as people know that they’re being recorded” because of the observer’s paradox. Thus, in terms of being authentic spoken texts, while recorded conversations would probably yield better results, the texts for this corpus are still highly valid and, for the most part, accurately reflect contemporary spoken American English.

The 88 million tokens of fiction come from 18,928 diverse text sources, including short stories, plays, books, magazines, and movies. They come from a wide range of fiction subgenres including popular, science, historical, juvenile, fantasy, horror, mystery, romance, etc. This genre heavily represents narrative text types as opposed to the academic subcorpus which is mostly expository texts. Having both is meaningful because of the fundamental and paradigmatic differences between narrative and expository texts, especially in the type of vocabulary each macro-genre employs (Gardner, 2004; Grabe, 2002).

The 93 million tokens of the magazine genre come from about 100 different magazines, with a mix (overall and each year) between subgenres: news and opinion, health, home and gardening, women, financial, religion, sports and outdoors, children, entertainment, food, history, geography, science, technology, fashion, and more. Taken from a total of 50,928 different texts, this genre, together with newspapers, helps span the linguistic space between fiction (highly narrative) genre and the academic (highly expository) genre.

The 89 million tokens of newspaper stem newspaper sources across the United States. These are the *Associated Press*, *Atlanta Journal Constitution*, *Chicago Sun-Times*, *Christian Science Monitor*, *Denver Post*, *Houston Chronicle*, *New York Times*, *Orange County Register*,

Pittsburgh Post-Gazette, San Francisco Chronicle, St. Louis Post-Dispatch, USA Today, and Washington Post. Totaling in 54,824 texts, the researchers selected a mixture of different sections including world, local, opinion, sports, financial, business, health, style, etc.

Lastly, the 88 million tokens of the academic genre are taken from about 100 different peer-reviewed journals. The journals used for this subcorpus were selected to cover the entire range of the 21 categories of the Library of Congress classification system (e.g. B. (philosophy, psychology, and religion), D. (world history), G. (geography, anthropology, and recreation), K. (education), M. (music), N. (fine arts), T. (technology), P. (language and literature), etc. These texts traverse a wide range of academic registers and are characteristic of formal registers and the expository macro-genre.

For each year and for each genre, the texts are balanced between the subgenres that comprise the five genres. This maintains compositional consistency from year to year. The only major genre of language not included in this corpus is internet-based such as emails, listserves, website, blogs, and social media. This was done for two primary reasons. First, in order to maintain compositional consistency back to 1990, certain subgenres would have been difficult or impossible to represent the further back in time one reaches. Second, the corpus is meant to represent only American English for which is difficult to control in the internet genre. All-in-all, however, COCA is perhaps the most balanced and reliable monitor corpus of any variety of the English language (Davies, 2009; Davies, 2010).

Dispersion.

Davies and Gardner (2010) describes how dispersion was calculated and incorporated into the word list from COCA. After the corpus was tagged for part of speech and lemmatized, raw frequency of lemma was computed.

Then Juilland's "D" dispersion index was calculated. This statistical coefficient produces a score between 0 and 1.00. A score of 1.00 means that the word is dispersed perfectly across each section of the corpus. The lower the score, the less an item is spread across other sections. In other words, a low score, of .25 would mean that a word appears frequently in a few sections but appears very infrequently or not at all in other sections. This is calculated as:

$$D = 1 - \frac{V}{\sqrt{n-1}} \quad (1)$$

where n is the number of equally sized sections of a corpus and V is the variation coefficient determined by:

$$V = \frac{\sigma}{\bar{v}} \quad (2)$$

where \bar{v} is the mean frequency of a word in each section and σ is the standard deviation of the frequencies in the sections as:

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (v_i - \bar{v})^2}{n-1}} \quad (3)$$

For the COCA list, 100 equally sized sections of the corpus were taken.

The list generated from this corpus was created by using a simple formula using the dispersion measure of Juilland's D and the total raw frequency of occurrences:

$$score = total\ raw\ frequency \times D \quad (4)$$

Because of research performed by Lyne (1986), Juilland's D is regarded as one of the most reliable measures of lexical dispersion, with some researchers going so far as to say that it is "the most reliable of the various dispersion coefficients available" (Leech, Rayson, & Wilson, 2001, p. 18).

Construct of word.

This word list groups word together by lemma and includes proper nouns, acronyms, and function words. Davies (2009) states that the lemmatizing software used for in this corpus is the CLAWS-7 part-of-speech tagger for English. This tagger was also used for tagging the BNC and consistently achieved 96-97% accuracy in tagging that corpus (Burnard, 2007). Davies also went through the corpus after automatic tagging was complete and manually spent repaired reoccurring errors in tagging (M. Davies, personal communication, December 12, 2015). Multiword items are excluded from this list, although frequent phrasal verbs from COCA have been analyzed (Gardner & Davies, 2007).

All in all, it is clear that COCA and its word list have several theoretically distinct advantages for usage in vocabulary size tests over corpora and lists used in previous tests. First, COCA is ten times larger than other corpora used in other vocabulary size tests in terms of number of tokens. Second, it uses contemporary language where some other lists have largely been based on source texts that are decades or even centuries old. Third, it is based on lemma while all of the other lists are based on word families. Fourth, it represents more genres and subgenres than any of the other corpora in a balanced and systematic way. Fifth, it includes both written and spoken texts which no other corpus used for previous tests does. Sixth and lastly, it incorporates a formal dispersion statistic in addition to raw frequency in order to create a more meaningful word list which was done in only one other corpus. The combination of these six factors certainly make the COCA list far more suitable than any word list that has previously been used in vocabulary size tests and possibly more suitable than any other word list in existence. Table 9 compares COCA with corpora used for previous vocabulary size tests

Comprehensive Comparison Between Corpora for Vocabulary Size Tests							
Corpus	Published	# of Tokens	Word Construct	Source Texts	Written Genre?	Spoken Genre?	Dispersion
Teacher's Word Book of 30,000 Words	1944	18 million	Word Family	magazines, textbooks, fiction, English classics	Yes	No	None
The General Service List	1953	23+ million	Word Family	Teacher's Word Book of 30,000 Words, letters, minutes, newspapers, magazines	Yes	No	None
Kučera and Francis' Word List	1967	1 million	Word Family	15 genres of written texts	Yes	No	None
Campion and Elley's Academic Vocabulary List	1971	300 thousand	Word Family	academic texts	Yes	No	Informal
British National Corpus Spoken	1995	10 million	Word Family	transcribed recordings of speech from a wide-variety of sources	No	Yes	Formal
Vocabulary Lists: Words, Affixes, and Stems	1986	2 million	Word Family	academic texts	Yes	No	None
Cambridge English Lexicon	1980	18 million	Word Family	General Service List, Kučera and Francis' Word List, graded readers	Yes	No	Informal
Corpus of Contemporary American English	2008	450 million	Lemma	a wide variety of spoken, fiction, magazine, newspaper, academic texts	Yes	Yes	Formal

Table 8: Comprehensive comparison between corpora for vocabulary size tests

Construction and Validation of the Major Vocabulary Size Tests of English

The design, composition, and methodology of previous vocabulary size assessments in English will now be addressed. In order to understand what can be improved about vocabulary size assessments as they exist today, it is critical to detail the theoretical backing of the methodological choices made by previous test designers and validators.

This section will describe the design and validation of the most widely used, reviewed, and validated vocabulary size tests in English or in any language: the VLT, PVL, CATSS, VST, and EVST which was later updated to the EFL Vocabulary Test in 1992, X_Lex in 2003 and V_YesNo v1.0 in 2015 (Beeckmans, Eyckmans, Janssens, Dufranne, & Van de Velde, 2001; Beglar, 2009; Beglar & Hunt, 1999; Nation, 1983; Nation & Beglar, 2007; Meara, 1888; Meara, 1992; Meara, 2005; Meara & Jones, 1990; Meara & Milton, 2003; Meara & Miralpeix, 2015; Read, 2000). Almost all of these have versions freely available online, and although some of the most current online versions are not documented much in published literature, they are all similar, if not, the same as those discussed in academic journals.

The Vocabulary Levels Test.

This test was devised primarily by Paul Nation in the early 1980s. He originally developed the test at Victoria University of Wellington in New Zealand as a simple classroom assessment of L2 English vocabulary size, but over time, it has become more of a diagnostic/placement vocabulary test (Nation, 1983; Read, 2000). In fact, its usage is so widespread that one leading scholar in vocabulary assessment has called it the “nearest thing we have to a standard test in vocabulary” (Meara, 1996, p. 38).

In the VLT, the test is divided into five parts. Each part of the test represents a different “level” of word frequency: the first 2,000 words, 3,000 words, 5,000 words, the University word level (beyond 5,000), and 10,000 words. Each of the five levels of Nation’s test correspond to a different learning objective for English language learners. According to Nation (1990, p. 261), the 2,000- and 3,000-word levels are the high-frequency general words of English that all learners need to know for normal functionality in the language. The 5,000-word level is the upper limit of general high-frequency vocabulary. The 10,000-word level encompasses the “more common lower-frequency words of the language” (Read, 2000, p. 119). Finally, the University Word Level (UWL) is needful for understanding academic material such as textbooks.

The item type chosen for this test is a word-definition matching format. For each question, there are six available words and three available definitions. The test-takers must match words to definitions making the definition the test item rather than the word. All of the words in the definition must be less frequent than the target words (Schmitt, Schmitt, & Clapham, 2001). An example can be seen in the figure below.

Example of VLT Item	
1 blend	_____ hold tightly in your arms
2 devise	_____ plan or invent
3 embroider	_____ mix
4 hug	
5 imply	
6 paste	

Figure 5: Example of VLT item

This type of item makes it clear that this is a test of receptive vocabulary knowledge rather than productive. At each level there are 36 words and 18 definitions. The format is designed to minimize reading and the ability of the test-taker to guess. All of the words in each item belong to the same part of speech and need to be semantically distant from one another. With a sampling rate of 18 words per level, the sampling rate at each level are as follows: 0.9% for the 2000 level, 1.8% for the 3000 level, 0.9 % for the 5000 level, 0.36% for the 10,000 level, and 0.6% for the UWL (3,000 words).

The VLT has been evaluated by several researchers in various ways. Read (1988) administered a pre- and post-test to an unspecified number of participants in a three-month English proficiency course. He found reliability coefficients (formula not reported) at 0.94 in the pre-test and 0.91 in the post-test. Beglar and Hunt (1999) also describes an unpublished study involving the VLT at Japan's Temple University where a version of the VLT was found to be reliable at 0.95 using Cronbach's α and at 0.97 using IRT reliability statistics. All of these coefficients estimate the reliability of the VLT to be $r > 0.9$, which indicates high reliability.

Beglar and Hunt (1999) also do their own investigation into the validity of two levels of the VLT. In this study, the authors used four forms of the 2000 Word Level and four forms of the University Word Level of the VLT. These tests were administered to students at three high schools, two junior colleges and two universities in the Kansai region of Japan ranging from ages 15 to 23. Four hundred and six participants completed the 2000 Word Level Test and 464

participants completed the University Word Level. The means and standard deviations for both levels were found to differ. At the 2000 Word Level, both a repeated-measures ANOVA and post-hoc Sheffé's test revealed that the four forms differed significantly from one another, $F(3, 1485) = 167.14, p < .001$ and $p < .004$ respectively. The same tests were performed at the University Word Level, and they showed significant differences between the tests at that level as well, $F(3, 1389) = 386.34, p < .001$ and $p < .001$ respectively. These disparities between forms of the test indicate that the words as variables play an enormous factor with the reliability of vocabulary size tests, and it is an issue that this thesis hopes to address.

This variability of the behavior of words as test items is further shown in this study's analysis of item discrimination. Using item-total correlations (r_{pb}), the 144 total items of these tests were evaluated and 12 items (16.7%) of the 2000 Word Level Tests and 23 items (31.9%) of the University Word Level Tests, produced correlations below .30. According to Beglar and Hunt (1999), these 35 items (24.3%) of the 144 did not meet the required criteria. This is concerning because for tests where item discrimination has not been determined for all items. Where items are only selected based on their frequency, approximately one quarter of the items could be invalid if the correlations hold to be the same throughout all items.

After evaluating the initial versions of the test, the authors revised the two equivalent forms of both the 2000 and University Word Level Tests to include only acceptable items and reanalyzed it using a one-parameter Rasch analysis. The analysis revealed a wide range of item difficulty for both tests. For the 2000 Word Level Tests, item difficulties ranged from -2.49 to 1.97, and for the University Word Level Test, items ranged from -2.79 to 3.77 with many items indicating a potential area of further examination which this thesis hopes to investigate further. This variability in item difficulty strongly supports the idea that frequency is perhaps not as good

of a predictor of item difficulty as vocabulary size test designers have assumed it to be. The item difficulties of the test items can be seen in the table below.

Rasch item difficulty estimates for the 2000 Word Frequency Test						Rasch item difficulty estimates for the University Word Level Test					
Form A			Form B			Form A			Form B		
Item	Word	Difficulty	Item	Word	Difficulty	Item	Word	Difficulty	Item	Word	Difficulty
1	roar	1.48	1	victory	-1.60	1	deficiency	-0.78	1	hypothesis	-2.19
2	debt	1.19	2	birth	-2.18	2	prestige	0.1	2	episode	0.12
3	pride	-1.22	3	root	0.95	3	affluence	-0.85	3	region	3.09
4	temperature	0.14	4	opportunity	0.13	4	configuration	-0.33	4	instance	3.77
5	flesh	1.15	5	dozen	-0.16	5	adult	-0.60	5	consent	-1.33
6	salary	-0.54	6	tax	0.04	6	equilibrium	-0.04	6	geometry	-0.92
7	wage	0.23	7	wealth	0.47	7	trend	-0.52	7	clinic	0.12
8	skirt	-1.05	8	pupil	0.64	8	diagram	-0.55	8	text	-0.60
9	justice	0.14	9	clerk	-0.65	9	philosophy	1.01	9	motive	-0.33
10	cream	-2.49	10	education	-1.12	10	doctrine	-0.52	10	research	-0.47
11	motor	-2.22	11	scale	0.62	11	intimacy	1.01	11	democracy	-0.83
12	copy	-0.65	12	journey	-0.01	12	volume	0.95	12	project	-1.38
13	treasure	0.73	13	speed	-2.18	13	vision	0.25	13	sequence	-0.63
14	charm	1.19	14	castle	-1.43	14	anomaly	-0.18	14	intellect	-0.45
15	lack	0.87	15	rail	-1.88	15	sex	0.43	15	crisis	2
16	earn	0.62	16	stretch	0.83	16	rely	-0.23	16	supplement	-0.66
17	wander	1.37	17	introduce	-0.08	17	evaluate	-0.60	17	revise	0.73
18	limit	0.9	18	admire	1.56	18	attain	-2.79	18	ensure	1.71
19	manufacture	-0.58	19	burst	0.79	19	expose	-0.04	19	inspect	0.59
20	elect	-0.36	20	improve	0.39	20	publish	-0.01	20	accumulate	0.78
21	melt	-0.12	21	deliver	1.39	21	absorb	-0.76	21	saturate	0.39
22	hide	0	22	develop	-0.30	22	restrict	0.83	22	subside	0.13
23	spoil	0.85	23	arrange	1.97	23	transfer	-0.86	23	indicate	0.68
24	surround	0.37	24	prefer	-0.02	24	assume	1.01	24	participate	-1.01
25	original	0.65	25	usual	0.62	25	subsequent	1.01	25	valid	-1.76
26	private	-0.10	26	ancient	0.14	26	minimum	1.84	26	civic	-0.06
27	total	-1.81	27	brave	0.36	27	inherent	0.15	27	implicit	-0.38

Table 9: Rasch item difficulty for VLT

Beglar and Hunt (1999) go on to criticize the VLT based on several other factors. The article suggests that a larger sample size of words would be more accurate in determining vocabulary size. Also, it cites a limitation as being that there is no attempt at all to account for polysemy. Another area it criticizes is that it uses word families to group words together and not lemma. Finally, the test item type of six words and three definitions has yet to be thoroughly examined. The degree to which items in sets interact to see if item independence holds true has yet to be clarified.

Schmitt, Schmitt, and Clapham (2001) was another validation study of the VLT. Upon examining two versions of the test, a distractor analysis was performed. Any item with a distractor attracting more than 10% was deemed potentially problematic. 18 of the 300 distractor items (6%) met this criterion. It seems, although it is not entirely clear, that no distractor attracted better than chance (1/6) which is positive evidence that test-takers are not guessing and

not being fooled by attractive distractors. An analysis of variance with Scheffé's test was performed between the means of the frequency levels showing a statistically significant ($p < .001$) between all levels. They also performed a Guttman scalability analysis yielding results of 0.993 and 0.995 for the Coefficient of Reproducibility and 0.971 and 0.978 for Scalability which both indicate a very high degree of scalability. Finally, Cronbach's α was used as a reliability index across the levels yielding results greater than .91 as indicated in the table below.

	<i>Reliability of the levels sections (Cronbach's α)</i>		
Level	Items per version	Version 1	Version 2
2000	30	0.920	0.922
3000	30	0.929	0.927
5000	30	0.927	0.927
10000	30	0.915	0.924
Academic	30	0.958	0.960

Table 10: Reliability of the VLT

Webb and Sasao (2013) have also criticized the VLT for not having a 1000 Word Level because of the value of high frequency words in vocabulary tests and because the lists used for this test are rather old. Many of these studies pertain to the PVLT and CATSS as they are related tests that were largely based on the VLT and its design/composition.

The Productive Vocabulary Levels Test (Laufer & Nation, 1999).

The format of this test is modelled to mirror the VLT with 18 test items at each of the 2000, 3000, 5000, UWL, and 10000 Word Levels. Three parallel versions of the test were created. The test items are c-test item types, but they unlike the classical c-test, they are discrete and selective, meaning that each one is a single sentence designed to test specific words. Here is an example used to elicit the word *sermon*.

On Sunday, in his last se_____ in Church, the priest spoke against child abuse (p. 49).

Figure 6: Example of PVLT item

It also differs from other c-tests because instead of using the whole first half of a word, the minimal number of letters possible were used. Because the test is design to measure productive vocabulary knowledge, additional letters were only added to eliminate multiple possible answers. In this test, both spelling and grammatical mistakes are ignored if they can still be understood.

Laufer (1999) examines the reliability and validity of this test. Briefly mentioned in the earlier in this chapter, this study tested 79 EFL students: 10th graders (n=24), 11th graders (n=23), 12th graders (n=18), and 1st year university students (n=14). In addition to the results reported earlier in this chapter, a second part to the study was also conducted. Four groups of learners took four levels of the test: 2000, 3000, 5000, and UWL. One group took all four forms of the test at each level. The levels were matched to the students according to judgments by the researchers and the 10000 Word Level was excluded because it was deemed too difficult for this particular set of students.

Reliability was calculated using Kuder-Richardson Formula 21 (KR-21) which is a common method of determining internal test reliability. The results of these calculations are shown in the table below. The authors point out that the reliabilities of the 5000 Word Levels were low because of the small number of subjects and homogeneity of the group. However, even if the group were very homogenous, numbers higher than 0.02 and 0.04 would probably have been expected. Overall, the reliability figures are moderate, but the wide discrepancies between forms (e.g. 2000 Level Form A = 0.51 and Form C = 0.80) and the low reliabilities of several of the tests (e.g. 3000 Level Form B = 0.39 and three of the forms at the 5000 Word Level < 0.4) again indicate the variability of the word in terms of item difficulty within 1,000-word levels.

<i>Reliabilities for the levels in each of the four test versions</i>				
Level	Form A	Form B	Form C	Form D
2000 Level	0.51	0.67	0.80	0.67
3000 Level	0.50	0.39	0.47	0.56
UWL	0.72	0.63	0.61	0.78
5000 Level	0.61	0.38	0.04	0.02

Table 11: Reliability of PVL

The study also reported Pearson correlations between the different forms of the test. The correlations are mostly moderate or high and are statistically significant. However, the 5000 Word Level, again, shows less impressive results than the other levels. The authors attribute these lower correlations to fewer subjects and the “patchy, unsystematic knowledge at this level” (p. 43). Again, this result indicates the variable of word playing a significant role in determining the validity of vocabulary size tests.

<i>Correlations between four forms of the PVL</i>						
Level	A/B	A/C	A/D	B/C	B/D	C/D
2000 Level (<i>n</i> = 45)	.82*	.82*	.78*	.83*	.81*	.77*
3000 Level (<i>n</i> = 36)	.71*	.70*	.82*	.82*	.71*	.80*
UWL (<i>n</i> = 33)	.75*	.80*	.84*	.83*	.76*	.80*
5000 Level (<i>n</i> = 18)	.72 (<i>p</i> = .004)	.83*	.69 (<i>p</i> = .003)	.49 (<i>p</i> = .1)	.77 (<i>p</i> = .003)	.67 (<i>p</i> = .006)

Table 12: Correlations between forms of the PVL

Laufer (1999) is the only study that attempts to validate the PVL. However, upon closer examination of the tests, a few concerns obviate themselves. First, some of the c-test items could be answered with a variety of possible responses despite the efforts of the authors to eliminate this problem. For example, in a 2000 Word Level Test the following item appears.

You must have been very br_____ to participate in such a dangerous operation. (p. 49).

Figure 7: PVLT example 1

Although the intended response for this question is *brave*, there are a number of possible responses that are highly valid as well, including *brilliant*, *bright*, *brainy*, *brainless*, *brash*, *breathless*, *brotherly*, *brutal*, and *brutish* to name only a few. Because there is only limited context, a wide-variety of responses that are feasible. There are numerous items in both forms of the test that are similar to the example above where any number of alternate words might be used to complete the blank to create sensible sentences.

Another issue is that there are items where words used in the context-giving sentence are less frequent than the target word. While the VLT carefully designed word definitions to contain no words less frequent than the target word, the PVLT takes no such precaution. If the sentence is providing the context for the target word, it is paramount that the test-taker understand every other word that sentence contains or else two words are being tested in the sentence instead of only the target item. In a 3000 Word Level Test of one, the following example occurs.

Anthropologists study the struc_____ of ancient societies. (p. 49)

Figure 8: PVLT example 2

A simple query in COCA reveals that the lemma *anthropologist* occurs only 4,356 times, while the lemma *structure* as a noun has 55,693 tokens, 12.78 times more (Davies, 2008). The word list generated from COCA records *structure* as a noun as Rank 951 while falls far below at Rank 6366 (Davies, 2011).

Finally, idiomatic usages of words occur. One of the 5000 Word Levels Test contains the following example.

Some people find it difficult to become independent. Instead they prefer to be tied to their mother's ap_____ strings. (p. 49).

Figure 9: PVLT example 3

If a student taking the test knows that an apron is an article of clothing people wear in front over clothing to keep them from getting stained or dirty, there is nothing in the test item above to imply that definition. *Tied to one's mother's apron strings* is an idiomatic expression meaning to be controlled by one's mother. Therefore, what this test item is really testing is a learner's knowledge of that multiword idiomatic expression and not the intended target word, *apron*. Other issues with this test exist that are the same as the VLT because they share the same word list and similar test format design.

The Computer Adaptive Test of Size and Strength (Laufer & Goldstein, 2004).

This test is also modeled after the VLT in its design of levels. Like the VLT, it contains five levels: 2000, 3000, 5000, 10000 and UWL. Like the VLT and PVLT, items were randomly selected from the frequency lists which were used for the VLT. A total of 150 words are tested by CATSS, and 30 words are tested at each level. This test differs from other vocabulary size tests because four modalities of items exist at each level, which are scaled in order of difficulty: active recall is hardest, then passive recall, followed by active recognition, and the easiest being passive recognition. Examples of the four types of items for L1 Hebrew ELLs can be seen in the figure below.

<u>Active Recall</u> (retrieval of form/supply the L2 word)			
a. _____	שפע		
<u>Passive Recall</u> (retrieval of meaning/supply the L1 word)			
Affluence	_____ ש		
<u>Active Recognition</u> (retrieval of form/select the L2 word)			
שפע			
a. precision	b. affluence	c. axis	d. episode
<u>Passive Recognition</u> (retrieval of meaning/select the L1 word)			
affluence			
דיוק	פרק	ציר	שפע

Figure 10: CATSS example item

Test-takers start with the most difficult modality of an item for a word and continue until they have correctly responded or exhausted all of the modalities. Differences between each of these modalities was found to be highly significant ($p < .001$) using Tukey's post hoc test. No report about the validity of the levels of this test were reported or examined because the topic of interest for the study was the scalability of the four modalities. In order to determine this, a sampling of 16 randomly selected items were taken and Guttman's coefficient for reproducibility was calculated. The authors determined that a coefficient above .90 would be considered valid, and all 16 items had coefficients above this criterion.

The Vocabulary Size Test (Nation & Beglar, 2007).

This test was created in response to the desire to fill in the gaps of the VLT. The VST has 14 levels of 1,000 words, which not only fills in the gaps of the VLT but also goes beyond it by 4,000 words. This is the only vocabulary size test that explicitly states by which level word families are grouped according to Nation and Bauer's (1993) scale. For this test, level 6 word families are used, which includes almost all derivational affixes. This test samples 10 words from each level, which works out to 1% of the words. Because each item in the test essentially

represents 100 words, the total number of questions a subject answered correctly is multiplied by 100 to give a final score. The target word is placed in a non-defining context to help orient the test-taker toward the correct part of speech. Also, the most frequent form of a word in a word family was used as the target item. It is designed in a four-choice multiple choice format that looks like the example below taken from the 5th 1000-word level (p. 11).

1. miniature: It is a miniature.
 - a. a very small thing of its kind
 - b. an instrument for looking at very small objects
 - c. a very small living creature
 - d. a small line to join letters in handwriting

Figure 11: VST example item

The first and second levels of the test used West's (1953) GSL and the other 12 levels were taken from the BNC Spoken subcorpus. The authors tried to design items so that no words in the definitions were lower frequency than the target word, although there were a few exceptions. The authors do not seem to believe that this is a substantive issue; however, if the whole idea motivating the test is to test words based on frequency, confounding factors due to frequency should be considered paradigmatically problematic.

Beglar (2010) is the only other to evaluate the VST outside of the original study. In this article, Beglar performed a Rasch-based validation of the test, which showed reasonable validation in a number of ways. 197 participants from four groups took part in the study: adult native speakers of English studying in a Masters or Doctoral program at a major American university (n = 19), advanced English proficiency (TOEFL 560-617) Japanese ELLs (n = 29), intermediate English proficiency Japanese ELLs (n = 53), and low English proficiency Japanese ELLs (n = 96).

Figure 12 below shows the relationship between the Rasch calibrations for the test-takers and the 140 items of the test. The far left side shows the Rasch logit. The right side shows the item difficulty. The items are labeled according to their word frequency level and by their item number on the test form (e.g., 13,000-4 means the thirteenth 1,000-word level, item 4). While the test items fit the model fairly well, close observation shows a mixture of items from different word frequency levels across the map. Two 2,000-word level items have scores above 50 while two 9,000-word level items have below 50. This seems to confirm the notion that individual words vary greatly in their item difficulty, and frequency is only part of what makes vocabulary items easier or harder for learners to acquire. Despite this, however, the authors interpreted the results from Figure 1 to indicate that ten items per level is more than enough to estimate test-takers' lexical knowledge with a "high degree of precision" (p. 107).

Fit statistics were taken as another measure of validation. For both mean-square and standardized values > 2.00 was used as the criterion for determining infit and five items were determined to be misfitting: (1000-10, *basis* (Infit Mnsq = 2.05, Infit Zstd = 6.4), 3000-4, *scrub* (Infit Mnsq = 1.19, Infit Zstd = 2.7), 3000-9, *rove* (Infit Mnsq = 1.50, Infit Zstd = 4.3), 11000-10 *hessian* (Infit Mnsq = 1.48, Infit Zstd = 2.1), and 14,000-8, *erythrocyte* (Infit Mnsq = 1.36, Infit Zstd = 2.1). Misfits were either caused by a small number of examinees whose responses resulted in large residuals or poorly designed distractors. These were only a small percentage of the total number of items.

Wright map of person measures and item calibrations									
More Able Persons				More Difficult Items					
80	*								
					13,000–				
	*				4				
					14,000–				
	#				9				
	*#			T					
					11,000–	12,000–			
70	##	T			3	6	14,000–10		
					11,000–	11,000–	14,000–	9000–	9000–
	*##				10	9	8	4	9
					10,000–	12,000–	12,000–	12,000	13,000
					1	10	5	–9	–7
					9000–				
					10				
					11,000–	12,000–	13,000–	14,000	6000–
	*#				6	8	2	–4	8
					10,000–	10,000–	13,000–	13,000	14,000
	*			S	10	9	10	–9	–1
					14,000–				
					6				
					10,000–	10,000–	11,000–	11,000	11,000
	*			S	6	8	1	–5	–8
					13,000–	13,000–	13,000–	5000–	6000–
					3	5	6	5	10
					9000–7				
					10,000–	13,000–	14,000–	14,000	5000–
60	###				4	1	3	–5	9
					7000–5	9000–1	9000–6		
					10,000–	1000–	12,000–	12,000	7000–
	*				2	10	2	–3	2
					8000–7				
					10,000–	11,000–		4000–	4000–
	*##				3	7	2000–6	10	3
								7000–	9000–
					6000–5	6000–7	7000–4	7	5
					10,000–	11,000–	12,000–	12,000	14,000
	##			M	5	2	1	–4	–7
								4000–	5000–
					2000–9	3000–9	4000–1	4	1
						7000–		7000–	7000–
					5000–3	10	7000–3	6	8
					8000–1	8000–9	9000–8		

50	#####	M		10,000–7	11,000–4	5000–4	6000–2	6000–3
				7000–9	8000–2	8000–3	8000–6	
	*#####			14,000–2	2000–3	2000–4	2000–5	3000–2
				3000–3	3000–7	6000–9	8000–10	8000–5
				8000–8				
	*#####			1000–4	12,000–7	13,000–8	4000–2	4000–9
				5000–7	6000–1	6000–4	7000–1	
	*#####			1000–3	2000–1	3000–4	3000–6	5000–2
	#####	S		6000–6	2000–8	5000–8	8000–4	
	#####	S		2000–10	3000–10	3000–5	4000–7	5000–10
40	##			9000–2	9000–3			
	##			1000–9	3000–1	4000–6		
	##			1000–7	2000–2	4000–8	5000–6	
	*			1000–8	2000–7			
	*	T		1000–5	3000–8			
				1000–1				
30				1000–2	4000–5			
				1000–6				
Less Able Persons				Less Difficult Items				
<i>Note:</i> Each # represent approximately 3 persons. Each * represents approximately 1 person. M = the mean of the person or item estimates. S = standard deviation from the mean. T = 2 standard deviations from the mean.								

Figure 12: Wright map of VST

Figures 13 shows the mean ensemble difficulties of the frequency levels. As expected, the difficulty of the items have a general upward trend relating to frequency. However, the trend is not perfect. As shown in the figure below, the 8000-word level was greatly affected by one easy item, 8000-4 *kindergarten*. This, again, shows how items vary in difficulty across frequency

levels. The difference in difficulty between the first two levels is great, but last few levels are nearly the same in terms of their difficulty, which indicates that the relationship between frequency and difficult is not linear. Rather, it seems to a variagating curve that was not addressed in this article and has not been addressed anywhere else in the published literature. In fact, upon close examination, there are three levels (3000, 8000, and 12000) which have lower mean difficulty estimates than the previous level. This indicates issues of scalability and, therefore, validity for accurate placement by this type of test.

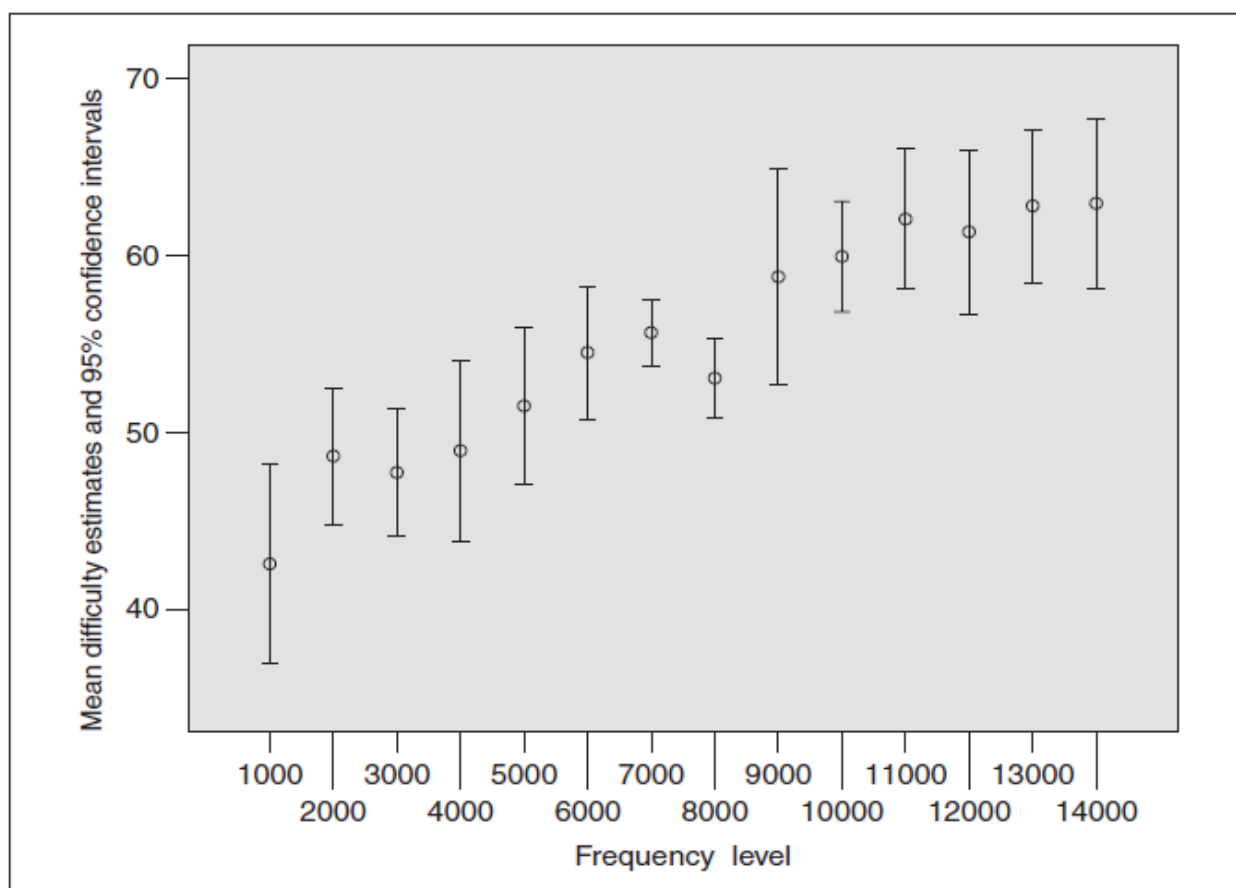


Figure 13: Difficulty by level for VST (Beglar, 2010, p. 109)

The statistics displayed a high degree of unidimensionality which accounted for 85.6% of the total variance. The authors stated that this amount of unidimensionality was far beyond their criterion for the primary latent construct. Four components other than the primary construct

were found, which all accounted for between 0.4% and 0.6% of the data. These data were interpreted by the researchers to mean that no secondary meaningful dimension existed in the test—only vocabulary knowledge. These are promising statistics and give further evidence that lexical knowledge can be tested as a construct separate from other aspects of linguistic knowledge.

Internal split-halves Pearson correlation yielded a coefficient of 0.98 significant at $p = .01$ (2-tailed test), indicating high internal reliability of the test. Another Pearson correlation was conducted between the positive and negative residual loading person measures resulting in correlations of $r = 0.84$ (disattenuated correlation = 0.91) at $p < .01$. Additionally, various different combinations of test items still indicated Rasch reliability indices > 0.96 , which indicates high reliability. Taken together, these statistics show a high degree of reliability for this test.

Person strata statistics were also taken which indicate the number of statistically distinct levels of person ability, which are separated by at least three standard errors of measurement. These statistics shows that for the 197 participants, there were seven statistically distinct levels of receptive vocabulary knowledge in the sample. Perhaps rather than determining level based on frequency, person strata statistical indices could be used to determine a person's lexical proficiency, as these are modeled to fit results based on person abilities and item difficulties.

All-in-all, this analysis showed the VST to be a valid instrument but with some potential issues that could use further attention. This is the most in-depth look at the validation of a vocabulary size test yet performed and is the type of analysis that every measure of vocabulary size should undergo before it is deemed valid. For this reason, Rasch-based analysis will be performed on the results attained from the VAST.

The Eurocentres Vocabulary Size Test (Meara & Buxton, 1987; Meara, 1992).

Meara (1990) and Meara and Jones (1988) describe how the EVST was originally designed in 1986-1987 as a computer-based assessment to replace the battery of placement tests administered at the Eurocentres schools in the United Kingdom but this test and future iterations of this test have been used for diagnostic purposes as well (Read, 2000). It measures the first 10,000 words of English in ten 1,000-word frequency bands. In 1992, a paper version was created and called the EFL Vocabulary Tests, which tested only the first 5000 words of English. In 2010, small changes were made to the test, based on criticisms from Beeckmans et al (2001), Huibregtse et al (2002), and Mochida and Harrington (2006). The latest version of the test is known as V_YesNo and can be found on Paul Meara's website (Meara, n.d.). Meara and Buxton (1987) was the first vocabulary size test to use the yes/no format and, to date, is the most widely used and studied yes/no vocabulary test in existence.

In the earliest computerized version of the test, the test-taker receives a random sample of 20 words from the 1000-word band; if the criterion level of performance is achieved, the test-taker will then move on to the next level and so forth until the criterion is not met. At this point, the program presents another 50 words from that specific level to estimate the learner's vocabulary size more precisely. The pen-and-paper version of the test, presents the test-taker with 60 items at each level. Multiple versions of the test are available at each level. Test-takers are to follow a protocol similar to the computerized version where they start at the lowest level and move up to take more tests if they attain a high enough score. The latest version of the test is documented in Meara and Miralpiex (2015). It states that this version of the test is about 200 items long.

The pseudowords for this test were created by Meara and his associates and piloting of the test revealed some problems with the words depending on the L1 of the test-taker. Initial testing revealed that for some words a “cognate effect” is created where certain pseudowords are similar to words in learners’ L1s and can fool them into thinking that it is a real word (Meara, 1992). For whatever reason, “some of the imaginary words are easier to handle than others: some can be rejected instantaneously while others cause even native speakers of English to puzzle for a long time” (Meara, 1988. p. 85-86). Any pseudoword that was deemed to be too much like a real word/too distracting or not enough like a real word/not distracting enough were eventually eliminated as the test was updated. The exact criteria for how the researchers went about eliminating words is never discussed in any of their publications.

There have been particular problems with certain language backgrounds with students of French, in particular, guessing more than any other language. For whatever reason, “French speakers seem to be more willing to accept imaginary words than speakers of other languages” and “the tests seem to underestimate their real vocabulary knowledge” (Meara, 2010, p. 11). Meara admits “we don’t know why this should be, but it is a problem that should be borne in mind by anyone using these test” (p. 11). Learners with other L1 backgrounds have not been found to have such problems, although, it is probable that learners from certain backgrounds simply have different response style tendencies.

The EVST has had much positive research which show its usefulness and validity. As mentioned above, Meara and Jones (1988) validated the original test by showing moderate correlations with a placement test, and other researchers have confirmed the validity of the test by various other means as well. Also, as was mentioned earlier in this chapter, Van Zeeland (2013) correlated the EVST with a test of oral vocabulary size as a confirmatory study of Milton

and Hopkins (2006). All of these studies validate the EVST as an accurate measure of vocabulary size. Additionally, Read (2000) praises the practical nature of this test from both a development and end-user standpoint; it is easy to generate, take, and score. Meara and Miralpeix (2015) state that the latest version of the test takes about ten minutes to complete the 200 items and the test is automatically scored and a score is given to the test-taker immediately upon completion.

For all of the positive points researchers have written about the EVST, there are some criticisms. Beeckmans et al (2001) states that this test may be problematic for test-takers suffering from dyslexia. They also mention that the EVST and other vocabulary size tests have been too short and that more items are necessary based on personal communication with Paul Meara. The article also criticizes test instructions and implications of having learners mark whether they know a word or not. “Knowing a word” may mean different things to different learners, so it is important that test instructions clarify what level of word knowledge is expected from them in order to mark a word as a real word or pseudoword. They also remark that all pseudowords should be checked to make sure they do not, in fact, exist. These issues will be addressed in the next chapter as considerations in development the instrument for this study.

Most of the criticism has been because of the scoring formula adopted from Zimmerman et al (1977). As this item type was still fairly new when Meara and his colleagues adopted it into the EVST, validation of yes/no tests was necessary. Beeckmans et al (2001), Huibregtse (2002), and Mochida and Harrington (2006) all examined yes/no format and concluded that the scoring method used in Zimmerman et al (1977) and Meara and Jones (1990) have some formulaic issues.

In scoring a yes/no test, there are two different ways of answering two different kinds of items resulting in four possible answers shown in the figure below. The labels come from Signal Detection Theory and are as follows:

- Hit: ticking a real word;
- False alarm: ticking a pseudoword;
- Miss: not ticking a real word;
- Correct rejection: not ticking a pseudoword.

		Response alternative (Do you know the word ?)	
		Yes	No
Item alternative	Word	Hit	Miss
	Pseudoword	False alarm	Correct rejection

Diagram illustrating the four possible responses in a 2x2 matrix. The top row represents 'Word' items and the bottom row represents 'Pseudoword' items. The left column represents 'Yes' responses and the right column represents 'No' responses. The four cells are: Hit (top-left, white), Miss (top-right, shaded), False alarm (bottom-left, shaded), and Correct rejection (bottom-right, white). To the right of the matrix, a legend shows two shaded squares labeled 'false responses' and two white squares labeled 'correct responses'.

Figure 14: Possible responses for yes/no test (Beeckmans et al, 2001, p. 237)

Different scoring formula account for these four possibilities in different ways. The original scoring formula used in the earliest versions of the EVST was a correction for guessing (*cfg*) procedure (Meara & Buxton, 1987) that takes into account the proportions of hits and false alarms by each individual. This and two other formulae that were used in Huibregtse et al (2002) and Mochida and Harrington (2006) are shown in the figure below.

<p>1) Correction for guessing (<i>cfg</i>):</p> $P(h) = \frac{(h)-(f)}{1-(f)}$ <p>where $P(h)$ = true hit rate, h = observed hit rate, f = observed false alarm rate.</p> <p>2) Meara's Δm</p> $\Delta m = \frac{(h-f)}{(1-f)} - \frac{f}{h}$ <p>3) Index of Signal Detection (I_{SDT})</p> $I_{SDT} = 1 - \frac{4h(1-f) - 2(h-f)(1+h-f)}{4h(1-f) - (h-f)(1+h-f)}$

Figure 15: Scoring formula for yes/no tests (Mochida & Harrington, 2006, p. 76)

The *cfg* formula is based on a 'blind guessing model'. This model assumes that either respondent knows the word or is guessing at random. Beeckmans et al (2001), Huibregtse et al (2002), and Mochida and Harrington (2006) all criticize this formula because it stresses the hit rate over the false alarm rate. For example, when the hit rate is 100%, the false alarm rate becomes irrelevant. It also does not account for different response styles. If a certain test-taker has a tendency for a high false alarm rate or a low false alarm rate, this formula does not account for this variability (Mochida & Harrington, 2006).

As a result, Meara (1992) offered a different formula in order to dispense with these issues, Δm . However, this formula turns out to be overly conservative in the opposite direction (Huibregtse et al, 2002; Mochida and Harrington, 2006). In this over-corrective formula, extremely low scores are produced when false alarm rates are high, even if the hit rate is also high. This formula, too, does not take into account individual response bias as a source of variability (Huibregtse et al, 2002; Mochida and Harrington, 2006).

Because neither of the previous formulae could account for this type of bias, Index of Signal Detection or I_{SDT} was suggested in Huibregtse et al (2002) and was tested in Mochida and

Harrington (2006). Mochida and Harrington (2006) compared test scores of 36 test-takers using these three formulae. A yes/no test was created, based on words from the VLT. Participants in the study took both the yes/no and VLT tests. The yes/no tests were scored using each of the three formula and found that the I_{SDT} formula correlated most closely with the VLT ($r = 0.88$).

Even though the other formulae also correlated nearly as well as I_{SDT} with the VLT, the hit rates were relatively high and the false alarm rates were relatively low. All of the participants were undergraduate or graduate L2 English students at the University of Queensland who had IELTS scores of at least 6.5. Meara (2010) maintains that more proficient ELLs will have lower false alarm rates, and therefore, all of the formulae would have scored them similarly. Therefore, if less proficient ELLs were to take a yes/no test, one would expect the I_{SDT} formula to be better at accounting for a wider range of response styles. For this reason, this is the scoring method that was chosen for use on the test used in this thesis.

Research Questions

Based on this review of literature, this thesis intends to address the following research questions.

1. To what degree is a vocabulary size test based on the COCA word list reliable and valid?
2. Is a vocabulary size test based on the COCA word list more reliable and valid than vocabulary size tests based on other word lists?
3. Do words across 1,000-word frequency bands vary in their item difficulty?

Chapter 3: Methodology

In order to answer these research questions, the study was designed according to the methodology described in this section. Participants from a variety of L1 backgrounds took a computer-based yes/no vocabulary size test based on the first 5,000 words of the COCA word list. Scores from that test underwent a Rasch-based analysis to determine reliability and validity. The scores were also compared against statistics reported by previous studies of vocabulary size tests. Finally, items across frequency bands of 1,000 words were analyzed to determine if and/or to what degree item difficulty and fit statistics varied.

Participants

Participants for this study were full-time missionaries from the Church of Jesus Christ of Latter-day Saints who were at either the Provo Missionary Training Center (MTC) in Utah or the Mexico Missionary Training Center in Mexico City. The participants were non-native speakers of English from a variety of L1 backgrounds. Four hundred and three missionaries completed the whole test. Two hundred and twenty of the participants were male, and 183 were female. The participants were between the ages of 18 and 30 with most of them being between the ages of 18 and 20. An exact breakdown of their ages is shown in Table 14. The participants varied widely in the amount of time they spent learning English. Many of the participants have been studying English for less than one year, which is actually a positive point because we are only testing the first 5,000 lemmas of English. Learners studying over an extended period would be expected to know all of these words. A breakdown of the number of years the participants had studied English is shown in Table 15.

Age of Participants	
Age	# of Participants
18	50
19	144
20	85
21	41
22	21
23	24
24	20
25	6
26	7
27	1
28	2
29	1
30	1

Table 13: Breakdown of age of participants

Number of Years Learning English by Participants	
# of Years Learning English	# of Participants
Less than 1	63
1	39
2	20
3	36
4	17
5	24
6	29
7	21
8	15
9	17
10	19
11	7
12	27
13	20
14	10
15	6
16	6
17	4
18	7
19	9
20	2
21	0
22	3
23	1
24	0
25	1

Table 14: Number of Years Learning English by Participants

Additionally, the learners came from a wide number of L1 backgrounds. Table 16 shows the 36 different L1s of the test-takers and the exact numbers of participants from each language. The languages with the most participants are Spanish ($n = 173$), Korean ($n = 31$), Mandarin Chinese ($n = 27$), Kiribati ($n = 24$), and Japanese ($n = 23$). Four other languages had more than ten participants: French, Portuguese, Samoan, and Tongan. Albanian, Cambodian, Cantonese, Cebuano, Chuukese, Czech, Fijian, Finnish, German, Haitian-Creole, Hungarian, Indonesian,

Italian, Kuanua, Lao, Malay, Mandarin, Mongolian, Norwegian, Pingelapese, Russian, Swahili, Tagalog, Thai, Turkish, and Waray-Waray all had fewer than ten participants.

L1s of Participants			
L1	# of Participants	L1	# of Participants
Albanian	1	Lao	1
Cambodian	4	Malay	1
Cantonese	8	Mandarin	27
Cebuano	4	Mongolian	4
Chuukese	2	Norwegian	2
Czech	1	Pingelapese	2
Fijian	2	Pohnpeian	4
Finnish	2	Portuguese	15
French	13	Russian	1
German	4	Samoan	11
Haitian-Creole	5	Spanish	173
Hungarian	1	Swahili	1
Indonesian	1	Swedish	3
Italian	3	Tagalog	8
Japanese	23	Thai	4
Kiribati	24	Tongan	14
Korean	31	Turkish	1
Kuanua	1	Waray-Waray	1

Table 15: L1s of Participants

Testing Instrument

The test designed for this thesis is a Vocabulary of American-English Size Test (VAST). Because COCA is the corpus/word list upon which this test is based, this test must, by necessity, be a test reflecting test-takers' knowledge of the 5,000 most frequent words of American-English as opposed to some other dialect or variety of English or English and as a global lingua franca.

VAST as a yes/no test.

The yes/no item type was selected for a number of reasons. First, the goal for this test was to test at least 500 vocabulary items. Every tenth word was selected so that variability of item difficulty across 1,000-word bands could be determined. Therefore, for the sake of

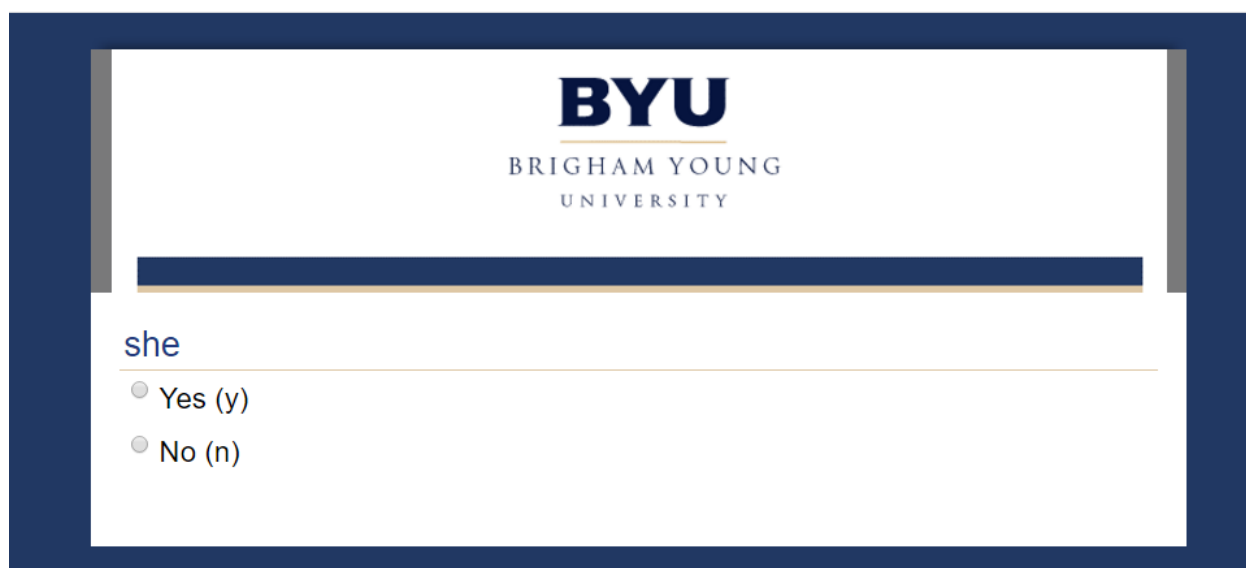
practicality, only the top 5,000 words were tested. Because the most frequent words have been determined by previous research to be the most important, 5,000 words was determined to be sufficient to answer the research questions for this thesis. In order to test this number of items, the most practical method of vocabulary size testing was selected that has been shown to be both highly valid and highly reliable. Creating distractors for multiple choice or matching or contexts for a productive test would have been difficult and time consuming to produce and pilot. For this number of test items, the yes/no item type seemed ideal because it would allow test-takers to see large numbers of items in a short period of time. At the same time, it could be automatically scored, and according to Shillaw (1996), it meets the assumptions of a Rasch-based model. Consequently, the yes/no test was selected as the item type for the VAST.

Test format.

This test was designed as a yes/no test to accommodate 510 real words and 340 pseudowords. Although the test has 850 items in total, each time participants take the test, they only see 250 test items. It is impractical to have a single test-taker take all 850 test items, so the test was divided into four forms. Each form of the test had 250 questions; 50 of these items (20%) were shared among all versions of the test to act as anchored items between the four forms of the test. Of the 200 non-shared items, 120 were real and 80 were pseudoword, and of the 50 shared items, 30 were real and 20 were pseudowords to maintain the real 60:40 word-pseudoword ratio suggested by previous researchers (Beeckmans et al, 2001; Meara, 1992; Meara, 2010; Mochida & Harrington, 2006). The 250 test items on each form of the test were administered randomly to eliminate any effect of a particular item sequencing. Originally, it was intended that words would be selected equally between five 1,000 word bands based on the first 5,000 words of the COCA word list. In addition to these words, nine words were accidentally

selected and included on the test from outside the first 5,000 words: rank. 5,021; 10,021; 15,021; 20,021; 25,021; 30,021; 35,021; 40,021; and 45,021. As a result, 101 words were taken from the first 1,000 words, and 100 words were taken from each band of 1,000 words from 2,000 to 4,000 words.

The test was designed in and administered through Qualtrics, which is an online surveying software. The software is easy to use from both the user's and test-designer's ends, which made distribution of the test and scoring of the data fairly convenient. The test was designed so that each person who took the test would get a random form of the test. The test was either answerable by clicking on *Yes* or *No* with the mouse or by pressing *y* or *n* on the keyboard respectively. An example of how a test item appears to a test-taker is shown in Figure 16. After completing the 250 questions of the test, the test-taker then would see a message thanking them for participation in the study and giving credit to Meara (1992) as the source of the pseudowords for this study.



BYU
BRIGHAM YOUNG
UNIVERSITY

she

☐ Yes (y)

☐ No (n)

Figure 16: Sample item from the Vocabulary of American Size Test

Real word selection.

Words for this test were taken from the COCA word list because COCA is a large, representative and balanced corpus of modern English. In order to get a sampling of words across bands of frequency, 10% of all of words in the list were selected as test items. Every 10th word from the COCA word list was used starting with the most frequent word in English (*the*) then the 11th most frequent word (*I*), then the 21st (*they*), and so forth up to 5,000.

In order to maintain the real 60:40 word-pseudoword ratio used and recommended by previous researchers (Beeckmans et al, 2001; Meara, 1992; Meara, 2010; Mochida & Harrington, 2006), ten additional words were selected: one every 500 words, i.e., 2nd, 502nd, 1002nd, 1502nd, etc. to distribute the selection of those words across the 5,000 words. This brought the total number of real words to 510. The details of these exact distributions and proportions will be explained in greater detail later in this chapter.

Pseudoword selection.

Part of this study is to compare the COCA list against real word lists from previous vocabulary size assessment studies. Therefore, pseudowords from the EVST were adopted into this study to help control for the variable of pseudowords in comparisons with that test. Meara (1990) discusses how some pseudowords are inherently more difficult than others. Shillaw (1996) confirms this with his Rasch-based analysis.

In order to make sure no pseudoword actually existed as a real word, each one was searched in both the BNC and COCA. This procedure has never been performed for a yes/no test, but Beeckmans et al (2001) cites it as being a necessary precaution. The goal was to find 340 pseudowords from the EVST for use in the VAST. An equal number of pseudowords was taken from each of the levels from Meara's (1992) EFL Vocabulary Tests so that results by level

correlated with results from the VAST. In order to find 340 pseudowords, a total of 623 pseudowords from the EVST were searched in COCA and BNC. In other words, 283 pseudowords were found to have actual occurrences in at least one of these two corpora. Any word to have any occurrence in either corpus was excluded from the VAST.

The vast majority of pseudowords found with occurrences in the corpora were being used as proper nouns. However, as proper nouns, some of them such as *wray*, *channing*, *lauder*, *kiley*, *hammond*, *garrick*, *portman*, *deere*, *dring*, *jarvis*, and many, many others had several hundred or even thousands of tokens. Others were not proper nouns but had more lexical usage, such as *contemporize*, *maturate*, *integrality*, and *twining*. For these reasons, it was determined that these words could potentially be confusing to test-takers and would be excluded. Examples of concordance lines of each of these words from COCA can be seen in Table 17.

<u>Year</u>	<u>Genre</u>	<u>Source</u>	<u>Key Word in Context</u>
1998	NEWS	Atlanta	from "Aida,"Verdi's grand 1871 opera?" I want to <u>contemporize</u> the story and popularize it, "Woolverton says. "Now, is
1991	ACAD	InstrPsych	to strengths of students. As students adjust to the college or university environment and <u>maturate</u> beyond the learning style restrictions
2011	FIC	Bk:WinHerHeart	robbed the latticework of its color, but the promise of spring lingering in the <u>twining</u> stems. Levi rapped a knuckle against the door
2010	ACAD	Education	Turkish-teaching in our country as follows: The inability to realize the understanding of <u>integrality</u> in linguistic skill developping, The

Table 16: Pseudowords from EFL Vocabulary Test in COCA

Test instructions.

Mochida and Harrington (2006) point out the effect that test instructions can have on the performance of test-takers. Because yes/no tests are unfamiliar to most test-takers, many will not understand what is expected of them when taking the test and may be confused without clear

instructions. Yes/no tests with comprehensible and directive instructions tend to have lower false alarm rates and less sporadic responses (Eyckmans, 2004; Mochida & Harrington, 2006).

In order to make sure instructions were in their most understandable form to the participants, they were translated into the respective languages of the test-takers. Each translation was checked by at least one other native speaker before it was approved for implementation in the test. One of the demographics questions asks the test-takers about their native language, and from then on, the rest of the demographics questions and test instructions appear in the native language of the test-taker.

The instructions state that test-takers will be tested on vocabulary knowledge and that they are free to take as much time as they need to complete the test. They are told that they will be shown both words that exist and do not exist in English. Their task is to determine if they know that it is a real word of English. If they are sure it is not a word or if they do not know, they are instructed to say that it does not exist. They are told that random guessing will diminish their overall score and are discouraged from guessing. They are instructed on how to respond to the questions using the keyboard.

These instructions differ from those used by other yes/no vocabulary size tests, such as the EVST (Meara & Jones, 1987). Previous studies have used phraseology asking examinees to determine whether or not they know a word. Because the task is not measuring whether a test-taker knows a word, but rather, whether he or she thinks a word is a real or a fake word, the instructions were modified to reflect that difference. In other words, yes/no tests do not test word *knowledge*. They simply test word *recognition*, which is what the change in instructions reflects. This change in instructions does make it difficult to compare the results of this test

against the EVST. However, it more accurately reflects the task examinees are asked to perform than do the instructions for other yes/no tests.

After reading the instructions, test-takers see three practice examples of the questions which give immediate feedback on the correctness of their responses so that they know whether or not they are doing the test correctly. Both the exact wording of the instructions and examples are included in Appendix A.

Procedure

Piloting.

In order to work out potential issues this test might have had, five adult native American-English speakers and five adult non-native speakers took the test and gave feedback on the instructions and test format. Spelling and grammar errors in the instructions were corrected and wording was clarified prior to translation of the instructions. Originally, part of speech was also included in the test item next to the target word in parentheses. However, both non-native and native English speakers were often confused by the part of speech and, even after reading the instructions, some of the non-native speakers thought that the purpose of the test was to determine whether the part of speech matched the word. Because of the confounding effect it had on some test-takers, part of speech was eventually left out of the final form of the test.

Primary testing.

Testing took place in computer labs either at the Provo or Mexico Missionary Train Centers (MTC) when participants had available time. A trained test proctor was present to answer any questions the missionaries might have about the nature of the test. Participants in the study took this test on either a computer or mobile device. Upon arrival at the computer lab, the missionary was orally asked by the test proctor about his or her language background, and only

non-native English speakers were asked to take the test. The test-takers were also orally told to read the instructions carefully because the VAST is different than other tests. Most test-takers took between 20-30 minutes, although a few took more time and a large number took less time. If any test-taker had any questions, the test proctor could only answer using information that could be found in the instructions for the test. The data for any participant who did not finish the entire test was not included in this study.

Scoring.

The response data was exported from Qualtrics into a Microsoft Excel spreadsheet and was scored using simple algebraic formulae (simple raw hit rate (h) and hit rate minus false alarm rate ($h-f$)) in addition to the formulae described in the previous chapter. Thus, the test was scored by using h , $h-f$, c_{fg} , Δm , and I_{SDT} .

Data Analysis

The responses were scored automatically. Then, they were put through a Rasch model analysis. The Rasch-based analysis for this study was conducted in *WinSteps Version 3.91.1*. Person and item separation and fit statistics were calculated. Reliability was also determined through the Rasch model and in Cronbach's α , which is a commonly used statistical coefficient for test reliability.

Pearson correlations were then used to correlate item logit values with ranking and 1000-word frequency band. An ANOVA was also calculated between word frequency band and item logit values to determine the separability of levels. Descriptive statistics were also calculated for each 1,000-word frequency band.

Chapter 4: Results

Rasch-Based Analysis

With all 403 persons and 850 items included, the following results were obtained through Rasch analysis displayed in Tabled 18.

Summary of Rasch-Based Analysis									
INPUT: 403 Person 850 Item REPORTED: 403 Person 850 Item 2 CATS WINSTEPS									
3.92.1									

SUMMARY OF 403 MEASURED Person									

	TOTAL			MODEL		INFIT		OUTFIT	
	SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD	

MEAN	197.8	250.0	1.86	.20	1.00	.2	.88	-.2	
P.SD	29.0	.5	.98	.06	.13	1.8	.30	1.8	
S.SD	29.1	.5	.98	.06	.13	1.8	.30	1.8	
MAX.	246.0	251.0	4.70	.51	1.65	9.7	1.96	8.3	
MIN.	87.0	240.0	-.79	.14	.77	-2.9	.30	-2.6	

REAL RMSE	.21	TRUE SD	.96	SEPARATION	4.62	Person RELIABILITY	.96		
MODEL RMSE	.21	TRUE SD	.96	SEPARATION	4.68	Person RELIABILITY	.96		
S.E. OF Person MEAN = .05									

-									
Person RAW SCORE-TO-MEASURE CORRELATION = .96									

Table 17: Summary of Rasch-based analysis

Reliability.

An overall person Rasch reliability of 0.96 was obtained with a separation of 4.62. In other words, 96% of the time, the test is accurately placing people into 4.62 groups. An overall item reliability of 0.90 was obtained with a separation of 2.98. In other words, 90% of the time, the test is accurately separating items in 2.98 groups. Cronbach's α (KR-20) was found to be 0.86.

Rasch reliability was also calculated from each band of 1,000 words. In addition, reliability was determined for only pseudowords. The results are shown in Table 19.

Rasch-Based Analysis by Level		
Frequency Band	Item Separation	Item Reliability
1,000	1.51	0.69
2,000	2.18	0.83
3,000	2.74	0.88
4,000	3.3	0.92
5,000	3.14	0.91
Pseudowords	2.75	0.88

Table 18: Rasch-Based Analysis by Level

Fit.

Person Fit Statistics

Overall, person fit statistics showed a good general fit. The analysis showed a mean person infit mean-square of 1.00 and standardized value of 0.2. It also showed a person outfit mean-square of 0.88 and standardized value of -0.2. According to Linacre (2002), any mean square value ≤ 2 is not degrading for construction of measurement, and a standardized value of ≤ 2 has reasonable predictability. Ideal values of mean-square are between 0.5 and 1.5, and ideal values for standardized value are between -2 and 2. For the mean values of the overall test, person infit and outfit values fell within the ideal values for both mean square and standardized value. All 403 subjects had person mean-square values within the ideal range. However, 50 subjects had either input or output standardized scores that were > 2 and thus were potentially degradative to measurement.

Item Fit Statistics.

Overall, item fit statistics showed a good general fit. The analysis showed an overall item infit mean-square value of 0.98 and standardized value of 0.2. It also showed an outfit mean-square of 0.87 and standardized value of -0.1. For the mean values of the overall test, item infit and outfit values fell within the ideal values for both mean square and standardized value.

Even though the overall test was well-fit to the Rasch model, there were individual items which fell outside the ideal values. Of the 510 real words, 23 of them fell outside non-detrimental values, and of the 340 pseudowords, 59 of them fell outside of the non-detrimental values. Non-detrimental values are defined by Linacre (2002) as mean-square values > 2 or standardized values of ≥ 2 . Of the misfitting real words, three were from the 1,000 level, two were from the 3,000 level, three were from the 4,000 level, nine were from the 5,000 level, and six of them were the accidentally included words ranked above 5,000: *locale*, *childbearing*, *stoically*, *nonreactive*, *high-strength*, and *fornicate*. Both of the two interjections included in the test (*oh* and *mm-hmm*) were also among the misfitting items. Upon examination of the misfitting real words, it is not readily apparent why some of them are misfitting. There are no obvious patterns as to why certain words do not fit the Rasch model. Further analysis is needed to determine what factors cause certain real words to misfit. The misfitting items are shown in Table 20.

Misfitting Items from Rasch-Based Analysis of the VAST		
	Misfitting Pseudowords	Misfitting Real Words
1	F1000-certical	R101-only
2	F1000-degate	R381-second
3	F1000-disportal	R411-oh
4	F1000-factile	R2051-height
5	F1000-finalism	R2961-mm-hmm
6	F1000-innoculism	R3741-rhetoric
7	F1000-multiplifv	R3781-behavioral
8	F1000-oxylate	R3831-altogether
9	F1000-reconcilant	R4001-civic
10	F1000-retrogradient	R4231-sheer
11	F1000-strategyv	R4291-verdict
12	F2000-beautitude	R4341-metropolitan
13	F2000-benevolute	R4431-comprise
14	F2000-buttle	R4441-unprecedented
15	F2000-defunctionary	R4681-comply
16	F2000-descript	R4751-soar
17	F2000-extravagate	R4981-dictate
18	F2000-flamboymment	R10021-locale
19	F2000-mascarate	R15021-childbearing
20	F2000-motivize	R25021-stoically
21	F2000-provisual	R30021-nonreactive
22	F2000-quorant	R35021-high-strength
23	F2000-rudge	R40021-fornicate
24	F3000-baptistal	
25	F3000-bastin	
26	F3000-cardination	
27	F3000-contrivial	
28	F3000-detailoring	
29	F3000-distantial	
30	F3000-eluctant	
31	F3000-gummer	
32	F3000-hoult	
33	F3000-joice	
34	F3000-neutration	
35	F3000-paralogue	
36	F3000-refurge	
37	F4000-algoric	
38	F4000-barnish	
39	F4000-carpin	
40	F4000-fluctual	
41	F4000-graduabile	
42	F4000-legitimal	
43	F4000-obsolation	
44	F4000-oestrogeny	
45	F4000-preluminary	
46	F4000-presuppository	
47	F4000-professive	
48	F4000-savery	
49	F4000-solitist	
50	F4000-warman	
51	F4000-wellstead	
52	F4000-xenostrophic	
53	F5000-charlett	
54	F5000-conceitful	
55	F5000-doubltv	
56	F5000-homoglvph	
57	F5000-ickard	
58	F5000-nebulate	
59	F5000-whitelock	

Table 19: Misfitting items from Rasch-based analysis of the VAST

Non-Rasch-Based Analysis

Descriptive Statistics.

Descriptive statistics for item logit values by frequency band and for pseudowords are shown in the table below. A histogram of the item logit values for real and pseudowords was also generated and are shown in Figure 17. A boxplot of the logit values by level is also included in Figure 18.

Descriptive Statistics from VAST Results						
Frequency Band	1,000	2,000	3,000	4,000	5,000	Pseudo-
Mean	-1.27307	-0.7018	-0.339	0.0939	0.3481	0.526088
Standard Error	0.089934	0.108306	0.111257	0.123896	0.095122	0.041528
Median	-1.32	-0.825	-0.17	0.21	0.43	0.48
Mode	-1.81	-0.89	-1.14	-0.67	1.09	-0.09
Standard Deviation	0.903824	1.083061	1.112573	1.238956	0.95122	0.765744
Sample Variance	0.816897	1.173021	1.237819	1.535012	0.90482	0.586363
Kurtosis	0.311621	2.627789	0.875384	2.139822	-0.08717	-0.36875
Skewness	-0.17238	-0.84018	-0.37936	-0.94608	-0.46244	0.277451
Range	5.5	6.35	7.09	6.91	4.36	4.44
Minimum	-4.33	-4.43	-4.43	-4.43	-2.25	-1.54
Maximum	1.17	1.92	2.66	2.48	2.11	2.9
Count	101	100	100	100	100	340

Table 20: Descriptive statistics from VAST results

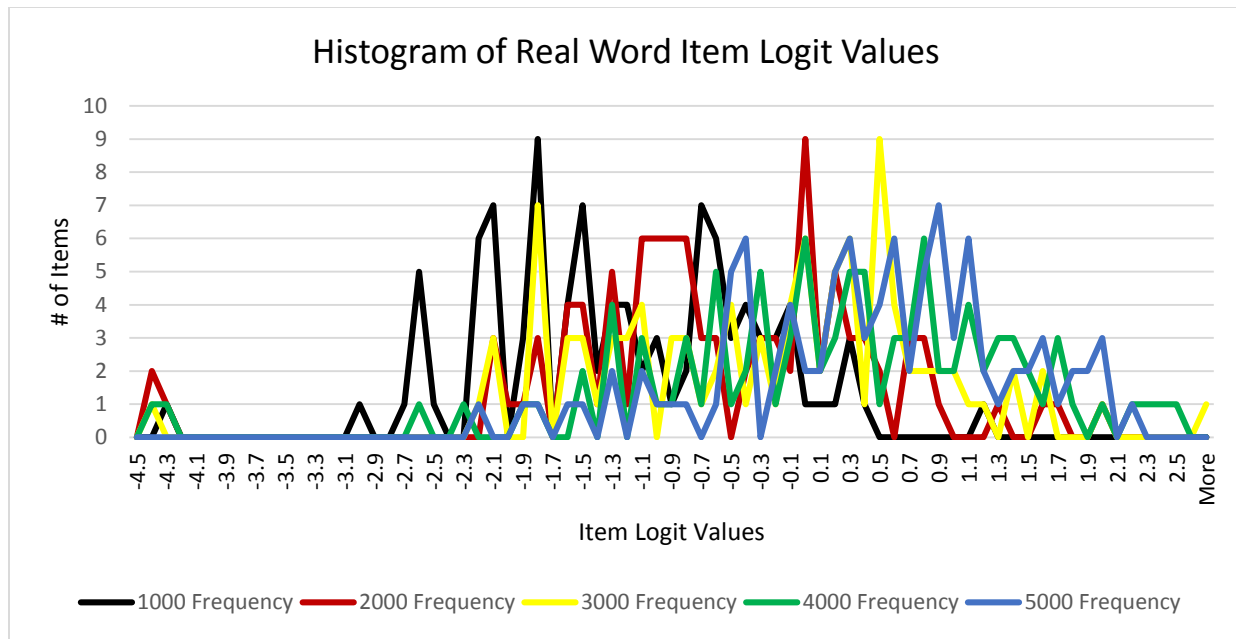


Figure 17: Histogram of real world item logit values

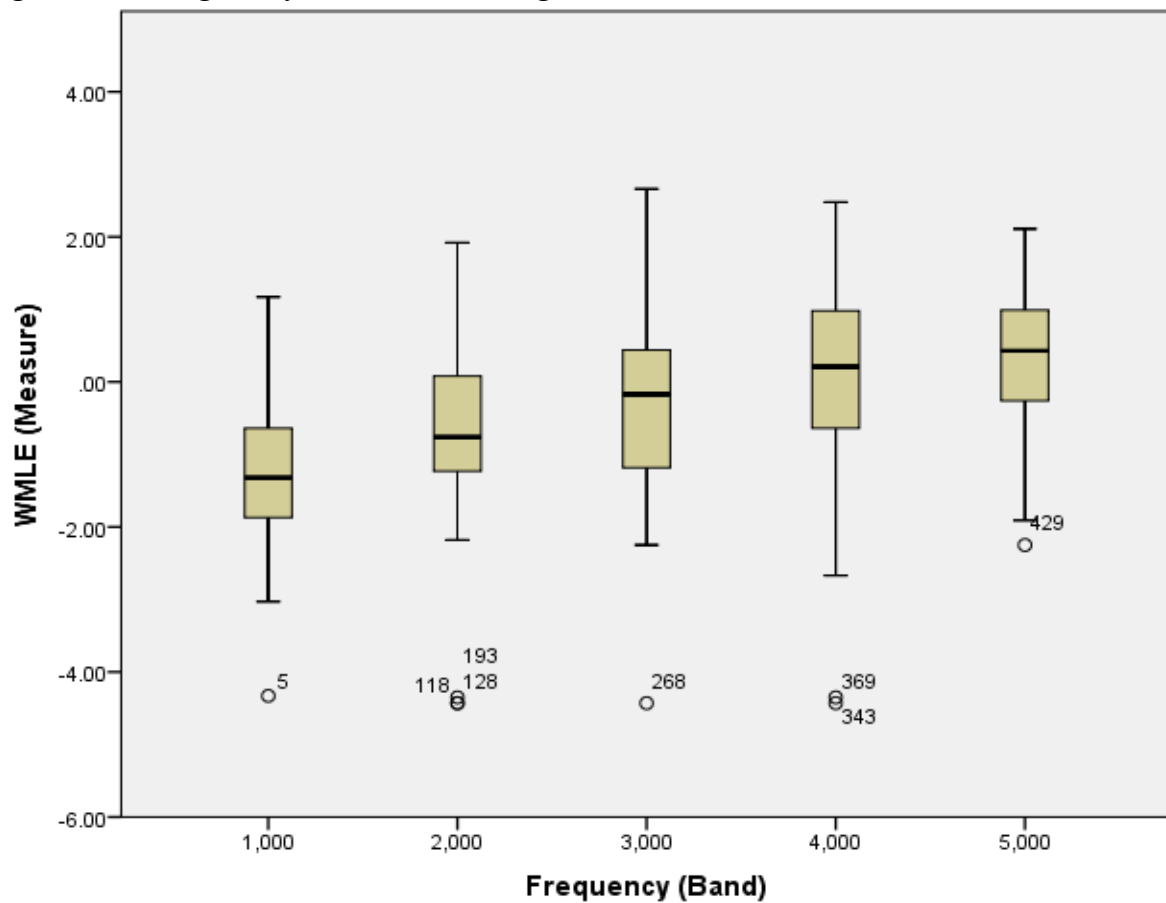


Figure 18: Boxplot of logit values—WMLE (Measure) or weighted maximum likelihood estimator is also called item logit value

Correlation.

Pearson correlations were conducted to determine the relationship between item logit values and ranking in the COCA word list. The correlation of item logit values and ranking in COCA of all 510 real items was $r = 0.329$ ($p < 0.001$, two-tailed). When the nine items that were accidentally included with ranks in COCA greater than 5,000 were excluded from the analysis, the results improved to $r = 0.474$ ($p < .001$, two-tailed). The r^2 of the two correlations are 0.108 and 0.225, respectively.

Pearson correlations were also conducted between logit values and frequency bands of 1,000 words from 1,000-5000, yielding a correlation of $r = 0.306$ ($p < 0.001$, two tailed) with an r^2 of 0.094.

Analysis of Variance.

An analysis of variance was conducted in order to determine if logit values of adjacent levels of 1,000 words were significantly different. The ANOVA used the five 1,000-word frequency bands and item logit values, yielding a result of $F(1, 4) = 36.752$, $p < 0.001$ with an adjusted r^2 of 0.222. A Tukey's post hoc test was also conducted. The results from this test can be seen in Table 22. The results showed significant differences between two of the four possible sequential frequency bands: the 1,000 and 2,000 levels at $p = 0.001$ and the 3,000 and 4,000 levels at $p = 0.034$, which indicated a total of three groups. There are also statistically significant differences between all the non-adjacent 1,000-word bands at $p < 0.00$.

ANOVA of 1,000-Word Levels of VAST by Logits

Dependent Variable: LOGIT

Tukey HSD

(I) FREQBAND	(J) FREQBAND	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1000.00	2000.00	-.5713*	.15015	.001	-.9824	-.1602
	3000.00	-.9341*	.15015	.000	-1.3452	-.5230
	4000.00	-1.3670*	.15015	.000	-1.7781	-.9559
	5000.00	-1.6212*	.15015	.000	-2.0323	-1.2101
2000.00	1000.00	.5713*	.15015	.001	.1602	.9824
	3000.00	-.3628	.15052	.114	-.7749	.0493
	4000.00	-.7957*	.15052	.000	-1.2078	-.3836
	5000.00	-1.0499*	.15052	.000	-1.4620	-.6378
3000.00	1000.00	.9341*	.15015	.000	.5230	1.3452
	2000.00	.3628	.15052	.114	-.0493	.7749
	4000.00	-.4329*	.15052	.034	-.8450	-.0208
	5000.00	-.6871*	.15052	.000	-1.0992	-.2750
4000.00	1000.00	1.3670*	.15015	.000	.9559	1.7781
	2000.00	.7957*	.15052	.000	.3836	1.2078
	3000.00	.4329*	.15052	.034	.0208	.8450
	5000.00	-.2542	.15052	.442	-.6663	.1579
5000.00	1000.00	1.6212*	.15015	.000	1.2101	2.0323
	2000.00	1.0499*	.15052	.000	.6378	1.4620
	3000.00	.6871*	.15052	.000	.2750	1.0992
	4000.00	.2542	.15052	.442	-.1579	.6663

Based on observed means.

The error term is Mean Square(Error) = 1.133.

*Highlighted values are adjacent frequency levels to the level examined

*. The mean difference is significant at the 0.05 level.

Table 21: ANOVA of 1,000-word levels of VAST by Logits

Chapter 5: Discussion

This chapter will be organized according to the three research questions for this thesis.

1. To what degree is a vocabulary size test based on the COCA word list reliable and valid?
2. Is a vocabulary size test based on the COCA word list more reliable and valid than vocabulary size tests based on other word lists?
3. Do words across 1,000-word frequency bands vary in their item difficulty in a vocabulary size test?

Research Question 1: To what degree is a vocabulary size test based on the COCA word list reliable and valid?

Construct Validity.

Construct validity is defined as the “degree to which it is appropriate to interpret a test score as an indicator of the construct of interest” (Carr, 2011, p. 315). Vocabulary size tests are designed to be measurements of the number of words learners know at certain word levels from word lists. These word lists should be reflective of the general language learners are likely to encounter and learn.

The COCA word list is a very large contemporary corpus that draws on a variety of genres and incorporates dispersion statistics into its word list. Additionally, COCA bases its word list on lemmas and not word families. Logically, since L2 learners are typically seeking to communicate in English with others (contemporaries), they are more likely to encounter words from a contemporary corpus than a dated one. Also, a corpus that draws on a wide variety of genres is a better representation of general language than is a list that draws on only one (Davies,

2010; Davies & Gardner, 2010). Language learners also more likely to learn words as lemma than as word families (Schmitt & Meara, 1997; Schmitt & Zimmerman, 2002).

Therefore, COCA produces a word list that is highly valid for its intended purpose—to measure the vocabulary size of contemporary test-takers. In short, because the COCA word list is designed to reflect the construct of interest, it contributes to the high overall construct validity of the VAST.

Reliability.

Carr (2011) states that determining test reliability “plays a vital role in demonstrating construct validity” (p. 122). Reliability is a measure of the consistency of a test—how well does this test maintain consistency internally or if a person takes this test multiple times, how consistently will they score the same? High reliability is an essential characteristic for any psychometric measure because it indicates the test is stable and, therefore, trustworthy.

Reliability statistics were calculated both in terms of Rasch person reliability and Cronbach’s α . For Rasch person reliability, a coefficient of 0.96 was obtained, which is a very high level of reliability. This means that in the Rasch model, the instrument is able to accurately discriminate persons into separate levels very well. The Cronbach’s α (KR-20) reliability for the VAST was found to be $\alpha = .86$, which is notably lower than the Rasch reliability but still very reliable. According to Linacre (1997), the true reliability for a test generally lies between the Rasch person reliability and Cronbach’s α . Either way, Carr (2011) recommends that a high-stakes test have a reliability of at least .80. The VAST showed higher reliability than Carr (2011)’s recommendation in both Rasch reliability and Cronbach’s α .

Separation.

Person separation is important for this test because it is a determination of the degree to which the test can separate examinees into different groups. This is important for the validity of a vocabulary size test because the goal is to separate test-takers into different levels of vocabulary knowledge. The data collected produced separation of 4.62, which is very high. This also shows good construct validity because it fulfills the purpose of the test—to separate test-takers into different levels of vocabulary knowledge. The test was composed of five 1,000-word levels, and the results showed test separating people into 4.62 levels. This shows a high degree of construct validity and test validity for the VAST.

Fit Statistics.

The overall mean-squared and standardized values for both infit and outfit showed that the items generally fit the Rasch model well. However, there were a large number of items that did not fit the model. Fifty-nine of the misfitting items were pseudowords. It is difficult to determine why these pseudowords did not fit the Rasch model when others did, without further in-depth analysis. However, if these misfitting pseudowords were replaced by better fitting items, the already-high reliability and separation values obtained through Rasch analysis would improve, which would make the test an even better vocabulary size measurement instrument. These misfitting words can be found in Table 20 on page 110.

For the real words, many of the mistakenly inserted words with ranks above 5,000 appeared to be misfitting, which was expected. The interjections *oh* and *mm-hmm* also seemed to pose a particular problem for test-takers of all proficiency levels. Test-takers often asked test proctors about these two words and were unsure whether to call them real words or not. So, it

seems, at least from this limited data, that interjections are confusing for test-takers in a yes/no vocabulary test and might need to be excluded from future vocabulary size tests.

Other than words ranked above 5,000 and interjections, there were 15 other misfitting real words. It is not clear why these words were confusing for test-takers and future investigations should be undertaken to determine why certain words fit the Rasch model better than others in a vocabulary size test. However, these misfitting real words were only a small percentage (4.5%) of the total real words. This evidence also indicates a strong validity for the VAST because 95.5% of the real words fit the Rasch model well.

ANOVA.

Levels were grouped in 1,000 words in order to follow the methodology of previous studies. The ANOVA showed that 2,000/3,000 and 4,000/5,000 word levels were not significantly different. Thus, there were only three significantly different groups among the five levels of items. These results match the Rasch results of item separation of 2.98. Because the test is separating items into multiple groups, it can be determined that the test has moderate validity, but the levels of the test as they are formatted currently do not display a high degree of validity. The level of the test would have to be reworked to be more closely aligned with item difficulty in order to improve the construct validity of the test.

In brief, the sum of evidences above shows that the VAST is a highly valid and reliable measure of L2 vocabulary size. However, the fact that some 1,000-word levels appear to be statistically indistinct is evidence that dividing levels by frequency is a less valid construct. The idea that frequency equates to vocabulary item difficulty and vocabulary items size is the whole theoretical framework upon which modern vocabulary size tests are built. Nevertheless, based

on this evidence, perhaps a reworking of how vocabulary size is determined ought to be considered.

Research Question 2: Is a vocabulary size test based on the COCA word list more reliable and valid than vocabulary size tests based on other word lists?

In order to answer this question, the results obtained in other studies about the VLT, PVL, VST, and EVST will be compared against result obtained in this study. All existing relevant literature will be compared against result obtained from the VAST.

Vocabulary Levels Test.

Beglar and Hunt (1999) measured the VLT to be reliable at 0.97 with Rasch reliability and 0.95 with Cronbach's α . Both of these are higher than the VAST which only obtained results of 0.96 and 0.86 respectively. Schmitt, Schmitt, and Clapham (2001) confirms that the VLT has higher reliability than the VAST with a Cronbach's α of 0.9325 taking into account all levels of the VLT.

These results are surprising because the VLT only has 27 items per level for each form of the test whereas the VAST has 50 items per level for each form of the test. Thus, the VLT maintains a higher reliability with fewer items. However, when considering the way in which the VLT was designed, these results become less surprising. Like many other vocabulary size tests, words were not chosen at random from the different word levels. Rather, vocabulary items were handpicked by the test designers to be "representative of all the words at that level" (Nation, 1983, p. 14). When taken in this light, it then becomes more logical that internal reliability would be high because the researchers would have picked items they believed to be of similar difficulty from that level and left out items that seemed too difficult or too easy.

Beglar and Hunt (1999) report a range of Rasch item difficulty for the 2,000 level to be 4.46 (-2.49 to 1.97) and a variance of 1.14. This is very comparable to the results obtained from the VAST for its 1,000-word levels. The mean range for the VAST by level was 6.042 and the mean variance for each level was 1.134. One would logically expect a larger number of items from across the 1,000-word bands would yield a wider range of logit values. This explains the larger range by level in the VAST. However, the fact that the variance by level between the two tests is almost identical indicates that the two tests are comparable in consistency of item difficulty within levels.

Productive Vocabulary Levels Test.

Laufer (1999) used the KR-21 formula to calculate the reliability of four forms of the PVLT to be 0.585, 0.5175, and 0.5075. The KR-21 is a simplified version of the KR-20 formula, which was used to determine a level of 0.86 in the VAST. In all four forms of the PVLT, the reliability is remarkably less than in the VAST. Thus, in terms of reliability alone, the PVLT falls far short of the VAST.

Vocabulary Size Test.

Beglar (2009) determined the internal reliability of the VST to be 0.98 with a Rasch reliability coefficient of > 0.96 . Both of these are also higher than the results obtained from the VAST, although the Rasch reliability seems to be only slightly higher. However, there is practically no difference here and both tests maintain an extremely high degree of validity. Thus, according to the studies cited in this thesis, the most reliable of these vocabulary size tests is the VST, then the VLT, followed by the VAST, with the PVLT being the least reliable.

Beglar (2009) also determined that the person strata statistics from his Rasch analysis showed separation into seven distinct groups, which is three greater than the VAST. However,

the VST sampled words up to the 14,000th word family, 2.2 times the number of the VAST, and so, one might logically expect the separation to be perhaps even greater than what the VST actually achieved.

Eurocentres Vocabulary Size Test.

For the EVST, Shillaw (1996) reports a Rasch reliability coefficient of 0.7148, which is much lower than the 0.96 value obtained in the VAST. This study also reports 15% of the items and 7% of the subjects misfit compared to the 9.64% of items and 12.4% of misfitting subjects found in this study. It makes sense for this thesis study to have more misfitting subjects because of the diversity of L2 learning experience, L2 learning environments, and L1 backgrounds of the participants. However, the VAST did have only two-thirds of the number of misfitting real words compared to the EVST. Perhaps the modern and better designed word list used in the VAST allowed the selection of better fitting real words than did an outdated and unbalanced corpora used in the EVST.

However, these findings have their limitations. Reliability can be easily influenced by any number of factors, especially the number and types of participants in the study. Perhaps, a better research design to compare the results of the VAST with previously designed vocabulary size tests would be to have the same group of test-takers take both tests and then to compare the results of the two tests.

Construct Validity

Because vocabulary size tests are built around frequency lists, having a suitable list is a key component towards having construct validity for these types of test. COCA produces a better list from which to test words for a vocabulary size test of English than do less contemporary and less representative lists used for other vocabulary size tests. The COCA word

list has many advantages over the lists used for the other vocabulary size tests. Table 9 on page 72 contains a summary of these advantages. The table shows that COCA is many times larger and newer than the other corpora. Additionally, it contains many more genres than do any of the other corpora. Finally, the COCA word list is based off of lemma rather than word family, which is a much more accurate representation of the way language learners acquire words (Nagy et al, 1993; Schmitt & Meara, 1997; Schmitt & Zimmerman, 2002).

Because the list produced from COCA is more suitable for generating general word lists than other dated or less-extensive corpora, it has a higher degree of construct validity than other vocabulary size tests.

Another reason why the VAST has a higher degree of construct validity than other tests is because of the way words were selected for the test. In the other tests, the words were hand-picked by the researcher so as to be “representative of all the words at that level” (Nation, 1983, p. 14). This methodology assumes that all words at a given word level can be represented as a cohesive group by only a few words. However, no empirical investigation into what makes words “representative of all the words” at a given level has ever been undertaken.

In summation, regarding whether the VAST is more reliable and valid than other vocabulary size tests, the results are not entirely conclusive. However, the evidence gathered in this thesis give evidence that the VAST might be a more accurate measure of the construct of interest than other vocabulary size tests. At the very least, VAST is comparable in many respects to other vocabulary size tests which have been created before it in terms of reliability and various aspects of validity. According to the studies explained above, the VAST is statistically less reliable than the VLT and VST but more reliable than the PVL or EVST. In terms of validity, it maintains a higher degree of construct validity because it is based on a more up-to-date and

lemmatized word list. More research needs to be done into the VAST to clearly determine a satisfactory answer to this question.

Research Question 3: Do words across 1,000-word frequency bands vary in their item difficulty?

Variance and range.

The variance for each level is shown in Table 21 on page 111. These figures show that for all five of the 1,000-word levels, the variance was > 0.8 logits with the 4,000 level having a variance of 1.535. The range for all of the levels was > 4.3 logits with the 3,000 level having a range of 7.09 logits. These results are shown graphically in Figures 17 and 18 on pages 103 and 107, respectively. Although this shows a consistent upward trend in item logit values by level, there is a great deal of overlap between the difficulty of words in each level. According to Figures 17 and 18, the relationship between item difficulty, and frequency appears to be moderate or weak, which was confirmed when Pearson correlations were taken between the two factors.

Correlations.

When a Pearson correlation was calculated to determine the relationship between item difficulty of all 510 real words and their frequency rank in COCA, a coefficient of $r = 0.329$ was obtained. The resultant r^2 was equal to .108. That means that frequency rank only accounted for 10.8% of the variability of item difficulty, which then leaves 89.4% to other factors.

When the nine words over rank 5000 were excluded, the correlation improved to 0.474 with an r^2 of 0.225, meaning that for the most frequent 5,000 lemmas in COCA, frequency rank only accounted for 22.5% of the variability of item difficulty.

The correlation decreased greatly when item difficulty was correlated against bands of 1,000 words for just the first 5,000 words to a weak $r = 0.306$. This gives an r^2 of 0.094 meaning that band of frequency only accounts for 9.4% of the variance.

These results seriously call into question the long-held assumptions that frequency is a good predictor of item difficulty and that learners who know words of certain frequencies will know other words of that same frequency level. According to these results, tests which place people into levels of 1,000 words may not be accounting for as much as 91.6% of the variance of item difficulty. Even by the best correlation obtained in this study between frequency and item difficulty, 87.5% of the variability is unaccounted. So, at least for this test, frequency was shown to be a poor predictor of item difficulty.

Finally, there is the matter of the 29 misfitting real words. It is not entirely clear why some of these words found in Table 20 are misfitting. Some of the misfitting real words are even highly frequent words on the COCA list such as rank 101 *only* rank and 381 *second*. Future investigation must go into determining the factors that make certain words misfit the Rasch modes.

In summation, words do vary in item difficulty across 1,000 word levels and can do so to a very great degree. For decades, scholars designed vocabulary size tests under the assumption that frequency predicts which words learners do and do not know. However, the evidences above indicate that this clearly is not always the case. For this tests, for bands of 1,000 words, 87.5% of the variance cannot be accounted for by frequency alone. It is likely that a variety of other factors affect which words learners acquire, and it is certain that these factors should be taken into consideration when designing future vocabulary size tests.

Chapter 6: Conclusion

Implications

Future vocabulary size test researchers have many factors to reconsider based on the results of this thesis. However, perhaps the most meaningful finding of this study has been the nature of the relationship between frequency of words in a corpus and the likelihood that learners know them. There are a variety of factors that have been discussed in this thesis outside of raw frequency that affect whether or not learners acquire certain words. These include necessity, cover, semantic neutrality, regularity of morphology, regularity of orthography, regularity of pronunciation, frequency in classroom, teaching, or environmental settings, language needs, linguistic distance, cultural distance, cognates, and possibly many others (Bauer & Nation, 1993; Meara, 2010; Nagy et al, 1993). In order to create more suitable lists for vocabulary size tests, it is important to consider how and to what degree these other factors affect what words learners acquire. As these factors are studied in greater depth, insight will be gained about how many words those learners know.

Figuring out how these factors behave and interact in different types of learners is an extensive process. As an alternative to depending on a frequency list alone, there is another solution. Because frequency may not be as good of a predictor of overall item difficulty as has been previously assumed, each word needs to be tested in order to see how particular types of learners handle them. After a sufficient range of learners is tested on a word, item difficulty for that word can be determined. Item difficulty might then show general tendencies of the order of acquisition of vocabulary.

Future computer-adaptive vocabulary tests should then be designed so that person scores can inform approximately how many words a person is likely to know. If logit values are determined for every word, then words can be ordered and ranked according to their item

difficulty. After completing a test, a test-taker receives a person score, which would then correspond to a logit value. That logit value would have a ranking which would show the learner their approximate vocabulary size.

For example, in a hypothetical situation, a test-taker scores a 1.75 person measure on a Rasch-based computer adaptive vocabulary size test. That value corresponds to the item *dilemma* which has an item measure of 1.75. *Dilemma* is ranked as the 3,500th easiest word in English. Therefore, the test-taker can be determined to know approximately 3,500 words in English. This might be a considerably more accurate representation of the lexicon of language learners.

Limitations

This study has a number of limitations that potentially had unknown effects on the results of the data. Perhaps the biggest limitation for this study was that because it does not parallel any of the previously designed tests in format, it is not entirely comparable. The VAST combined a number of confounding variables, and it is difficult to determine how these variables are interacting to influence the results that were obtained. The VAST varies in both test format and item type from the VLT, VST, and PVL. For the EVST, the VAST varies in its instructions and test format. In order to more directly compare previously designed tests with the VAST, these multiple confounding variables should have to be controlled.

For one, the subjects came from vastly diverse language learning backgrounds and have been learning English for varying amounts of time. Because of this, both the amount and type of vocabulary the subjects encountered and studied is likely to be extremely different. This could very well have had an impact on the way in which our particular respondents answered the test

questions. One might expect very different test results from a less diverse group of learners who all come from a similar language learning background.

Another limitation of this study was that only adult English L2 learners were observed. One might expect that the results would be different if the subjects were L1 learners or child L2 learners. Different results might also be obtained for different languages.

Yet another limitation is that this thesis only examines the first 5,000 lemma of English. It is important to determine whether the results of this study remain consistent beyond 5,000 lemmas in order to completely determine whether item difficulty and frequency maintain weak correlations at higher levels.

Finally, the test items themselves should be put to further study. Some of the pseudoword items were reverse discriminating. Such items should probably be modified or replaced to be better fitting. Also, only 10% of the first 5,000 words were examined. Because item difficulty in a vocabulary size test is not necessarily based on word frequency, in order to accurately determine a learner's vocabulary size, the difficulty of all words of a language need to be tested. This might also be an area of future research.

Future Research

This thesis reveals many areas where future research is where further confirmatory research is necessary. First, further evidence is necessary to fully determine the relationship between word frequency and likelihood that learners know words of certain frequencies. Future studies might investigate this relationship in corpora other than COCA. Little is known about the relationship between how corpora with different genre-balances and compositions might more accurately reflect item difficulty for certain words in vocabulary size tests.

This study should also be replicated in other languages and with different types of learners. For example, child learners will likely acquire very different vocabulary than adults. It is likely that different types of language learners encounter and learn different types of words. Replicating this study with different types of learners might reveal the types of words that learners of different ages and language learning settings are acquiring.

In order to confirm that COCA is, in fact, more suitable for a frequency list than other corpora words, words of the same frequency ranking in respective lists should be given to the same learners as test items in order to see which list is a better predictor of item difficulty. Another way to test this same assumption is to compare the frequency of words tested in this study relative to other word lists to determine if other word lists are more accurate predictors of item difficulty.

Also, there are other ways to validate the VAST than those performed in this study. Other vocabulary size tests have used scalograms and cross-validations with other measures in order to show validity (Read, 2000; Milton, 2010; Milton, 2013). In the future, these types of validations can and should also be done in order to confirm construct validity for the VAST.

Future research also needs to determine what factors also affect the order of vocabulary that learners acquire other than raw frequency. West (1953) suggests that such factors might include as necessity, cover, stylistic level, and emotional words. Any number of other factors not yet considered by any known researcher might also conceivably affect the order in which words are learning including similarity/dissimilarity to known words, length in terms of number of phonemes or written characters, irregularity of form, semantic complexity, linguistic or cultural distance, learning context, and many others.

Pseudowords are another area that warrants further investigation. What factors make certain pseudowords more or less easy to reject as fake words in yes/no tests? What are these factors for less or more advanced learners? What are these factors for learners from different L1 backgrounds? Although scholars have speculated (Meara, 1992; Meara, 2010; Meara & Jones, 1990; Read, 2000), no formal studies have been conducted researching these topics.

Perhaps another area to explore is whether results for a vocabulary size test would change if the word list from which the test is derived were based on lexemes instead of word families or lemma. Having a list that is more meaning-based and less form-based would allow researchers to determine if learners acquiring new words learn them more as word families or as lemma. It would also allow them to examine how they learn polysemous words.

Finally, one huge task that needs to be undertaken is developing vocabulary size tests for all of the major languages of the world and determining how different factors discussed in the review of literature of this thesis affect languages other than English. More issues surface in other languages. For example, some languages have opaque orthography such as Chinese, Japanese, and French. Others lack large and reliable corpora. Some languages have a good deal more homophony or polysemy than English. Others have more complex morphology.

In short, linguists have barely scratched the surface of a wide variety of issues relating to vocabulary size tests. Nobody in the 30 years since Nation (1983) first created the modern vocabulary size test has stopped to question its construct validity. Vocabulary size testing is an area of language testing rife with utilitarian potential and possibility for improvement. However, researchers are only just now beginning to question the underlying assumptions behind modern vocabulary size tests and examining more closely.

References

- Anderson, R. C. & Freebody, P. (1983). Reading comprehension and the assessment and acquisition of word knowledge. In B. Hutson (Ed.), *Advances in reading/language research: A research annual* (p.231-256). Greenwich, CT: JAI Press.
- Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing*, 22(3), 301-320.
- Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography*, 6(4), 253-279.
- Beeckmans, R., Eyckmans, J., Janssens, V., Dufranne, M., & Van de Velde, H. (2001). Examining the Yes/No vocabulary test: some methodological issues in theory and practice. *Language Testing*, 18(3), 235-274.
- Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, 27(1).
- Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 word level and university word level vocabulary tests. *Language Testing*, 16(2), 131-162.
- Berko, J. (1958). *The child's learning of English morphology* (Doctoral dissertation, Radcliffe College). Retrieved from <http://www.tandfonline.com/action/journalinformation?journalCode=rwr20>. (February 20, 2016)
- Biber, D. (1990). A typology of English texts. *Linguistics*, 27, 3-43.
- Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). *Longman grammar of spoken and written English*. London, UK: Longman.

- Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of Educational Psychology*, 93(3), 498.
- Bogaards, P., & Laufer, B. (Eds.). (2004). *Vocabulary in a second language: Selection, acquisition, and testing* (Vol. 10). Philadelphia, PA: John Benjamins Publishing
- Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary?: introducing the New General Service List. *Applied Linguistics*, 36(1), 1-22.
- Bruton, A. (2009). The vocabulary knowledge scale: A critical analysis. *Language Assessment Quarterly*, 6(4), 288-297.
- Burnard, L. (2007). BNC User Reference Guide. Retrieved March 04, 2016, from <http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#spodes>
- Cameron, L. (2002). Measuring vocabulary size in English as an additional language. *Language Teaching Research*, 6(2), 145-173.
- Carlisle, J. F. (2000). Awareness of the structure and meaning of morphologically complex words: Impact on reading. *Reading and writing*, 12(3), 169-190.
- Carlisle, J. F., & Katz, L. A. (2006). Effects of word and morpheme familiarity on reading of derived words. *Reading and Writing*, 19(7), 669-693.
- Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford University Press.
- Campion, M. E., & Elley, W. B. (1971). *An academic vocabulary list*. New Zealand Council for Educational Research.
- Chui, A. S. Y. (2006). A study of the English vocabulary knowledge of university students in Hong Kong. *Asian Journal of English Language Teaching*, 16, 1-23.
- Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 213-238.

- Davies, Mark. (2008-) *The Corpus of Contemporary American English: 520 million words, 1990-present*. Available online at <http://corpus.byu.edu/coca/>. Retrieved May 27, 2015.
- Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2), 159-190.
- Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary & Linguistic Computing*, 25(4).
- Davies, M. (2011-). Word frequency data. Retrieved February 13, 2016, from <http://www.wordfrequency.info/>
- Davies, M., & Gardner, D. (2010). *A frequency dictionary of contemporary American English: Word sketches, collocates and thematic lists*. New York, NY, USA: Routledge.
- Derwing, B. L., & Baker, W. J. (1979). Recent research on the acquisition of English morphology. *Language Acquisition*, 209-223.
- East, M. (2007). Bilingual dictionaries in tests of L2 writing proficiency: do they make a difference?. *Language Testing*, 24(3), 331-353.
- Edwards, R., & Collins, L. (2011). Lexical frequency profiles and Zipf's law. *Language Learning*, 61(1), 1-30.
- Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text-Interdisciplinary Journal for the Study of Discourse*, 20(1), 29-62.
- Eyckmans, J. (2004). *Measuring receptive vocabulary size*. Utrecht, The Netherlands: LOT (Landelijke Onderzoekschool Taalwetenschap).
- Fan, M. (2000). How big is the gap and how to narrow it? An investigation into the active and passive vocabulary knowledge of L2 learners. *RELC Journal*, 31(2), 105-119.

- Freyd, P. & Baron, J. (1982). Individual differences in acquisition of derivational morphology. *Journal of Verbal Learning and Verbal Behavior*, 21, 282-295.
- Gardner, D. (2004). Vocabulary input through extensive reading: A comparison of words found in children's narrative and expository reading materials. *Applied Linguistics*, 25(1), 1-37.
- Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics*, 28(2), 241-265.
- Gardner, D. (2013). *Exploring vocabulary: Language in action*. New York, NY: Routledge.
- Gardner, D., & Davies, M. (2007). Pointing out frequent phrasal verbs: A corpus-based analysis. *TESOL Quarterly*, 41(2), 339-359.
- Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics*, 35(3).
- Ghadessy, M. (1979). Frequency counts, word lists, and materials preparation: a new approach. *English Teaching Forum*, 17(1), 24-27.
- Goulden, R. P., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be?. *Applied Linguistics*, 11, 341-363.
- Grabe, W. (1991). Current developments in second language reading research. *TESOL Quarterly*, 25(3), 375-405.
- Grabe, W. (2002). Narrative and expository macro-genres. In A. Johns (Ed.). *Genre in the classroom: Multiple perspectives*, 249-267. Mahwah, NJ: LEA.
- Gyllstad, H., Vilkaite, L., & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL International Journal of Applied Linguistics*.

- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, 15(3), 309-333.
- Harsch, C., & Hartig, J. (2015). Comparing C-tests and Yes/No vocabulary size tests as predictors of receptive language skills. *Language Testing*.
- Hatch, E., & Brown, C. (1995). *Vocabulary, semantics, and language education*. 40 West 20th Street, New York, NY: Cambridge University Press.
- Hatch, E., & Farhady, H. (1982). *Research design and statistics for applied linguistics*. Rowley, MA: Newbury House.
- Hazenberg, S., & Hulstun, J. H. (1996). Defining a minimal receptive second-language vocabulary for non-native university students: An empirical investigation. *Applied Linguistics*, 17(2), 145-163.
- Hindmarsh, R. (1980). *Cambridge English lexicon*. Cambridge: Cambridge University Press.
- Hu, M. H., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1), 403-30.
- Huibregtse, I., Admiraal, W., & Meara, P. (2002). Scores on a yes-no vocabulary test: Correction for guessing and response style. *Language Testing*, 19(3), 227-245.
- Huertas, C., Gómez-Ruelas, M., Juárez-Ramírez, R., & Plata, H. (2011). A formal approach for measuring the lexical ambiguity degree in natural language requirement specification: Polysemes and Homonyms focused. In *Uncertainty reasoning and knowledge engineering (URKE), 2011 International Conference on* (Vol. 1, pp. 115-118). IEEE (Institute of Electrical and Electronics Engineers).

- Ishii, T., & Schmitt, N. (2009). Developing an integrated diagnostic test of vocabulary size and depth. *RELC Journal*, 40(1), 5-22.
- Jang, E. E., & Roussos, L. (2009). Integrative analytic approach to detecting and interpreting L2 vocabulary DIF. *International Journal of Testing*, 9(3), 238-259.
- Juilland, A. G., Brodin, D. R., & Davidovitch, C. (1971). *Frequency dictionary of French words*. Mouton.
- Juilland, A., & Chang-Rodriguez, E. (1964). *Frequency dictionary of Spanish words*. Mouton.
- Klepousniotou, E. (2002). The processing of lexical ambiguity: Homonymy and polysemy in the mental lexicon. *Brain and Language*, 81(1), 205-223.
- Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.
- Laufer, B. (1992). Reading in a foreign language: how does L2 lexical knowledge interact with the reader's general academic ability. *Journal of Research in Reading*, 15(2), 95-103.
- 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. (2005). Focus on form in second language vocabulary learning. *EUROSLA Yearbook*, 5(1), 223-250.
- Laufer, B. (2013). Lexical frequency profiles. *The encyclopedia of applied linguistics*.
- Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: do we need both to measure vocabulary knowledge?. *Language Testing*, 21(2), 202-226.
- Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, 16(1), 33-51.

- Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, 54(3), 399-436.
- Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: based on the British National Corpus*. London: Longman.
- Linacre, J. M. (1997). KR-20 or Rasch reliability: Which tells the “truth”. *Rasch Measurement Transactions*, 11(3), 580-581.
- Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean. *Rasch Measurement Transactions*, 16(2), 878.
- Lyne, A. A. (1986). In praise of Juillard’s “D”: A contribution to the empirical evaluations of various measure of dispersion applied to word frequencies. In *Proceedings, Méthodes quantitatives et informatiques dans l’étude des textes*. (pp. 587-597). Geneva, Switzerland: Slatkine.
- Lynn, R. W. (1973). Preparing word lists: a suggested method. *RELC Journal*, 11(2), 25-32.
- Malabonga, V., Kenyon, D. M., Carlo, M., August, D., & Louguit, M. (2008). Development of a cognate awareness measure for Spanish-speaking English language learners. *Language Testing*, 25(4), 495-519.
- McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research* 19(6), 741-760.
- Meara, P. (n.d.). Free software from _lognostics. Retrieved March 18, 2016, from <http://www.lognostics.co.uk/tools/index.htm>
- Meara, P. (1990). Some notes on the Eurocentres vocabulary tests. *Foreign Language Comprehension and Production*, 103-113.
- Meara, P. (1992). *EFL vocabulary tests*. Swansea, UK: ERIC Clearinghouse.

- Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer, & J. Williams (Eds.), *Competence and performance in second language acquisition* (pp. 35-53). Cambridge, UK: Cambridge University Press.
- Meara, P. (2005). Lexical frequency profiles: A Monte Carlo analysis. *Applied Linguistics*, 26(1), 32-47.
- Meara, P. (2010). *EFL vocabulary tests: second edition*. Swansea University. Retrieved from <http://www.lognostics.co.uk/vlibrary/meara1992z.pdf>
- Meara, P. (2013). Imaginary words. *The encyclopedia of applied linguistics*.
- Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, 4(2), 142-154.
- Meara, P., & Jones, G. (1988). Vocabulary Size as a Placement Indicator.
- Meara, P., & Jones, G. (1990). *Eurocentres 10K Vocabulary Size Test 10KA*. Zurich: Eurocentres.
- Meara, P. M., & Milton, J. L. (2003). *X_Lex: The Swansea Vocabulary Levels Test*. Newbury: Express.
- Meara, P. M., & Miralpiex, I. (2015). *Tools for vocabulary research*. Bristol: Multilingual Matters.
- Melka, F. (1997). Receptive vs. productive aspects of vocabulary. In N. Schmitt & M. McCarthy (Eds.) *Vocabulary: Description, acquisition and pedagogy*, (pp. 84-102). Cambridge, UK; Cambridge University Press.
- Millett, J., Atwill, K., Blanchard, J., & Gorin, J. (2008). The validity of receptive and expressive vocabulary measures with Spanish-speaking kindergarteners learning English. *Reading Psychology*, 29(6), 534-551.

- Milton, J. (2009). *Measuring second language vocabulary acquisition* (Vol. 45). Multilingual Matters.
- Milton, J. (2010). The development of vocabulary breadth across the CEFR levels. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: intersections between SLA and language testing research*, (p. 211-232). Eurosla Monographs Series, Rome, Italy: Eurosla.
- Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C. Bardel, C. Lindqvist, & B. Laufer. (Eds.), *L2 vocabulary acquisition, knowledge, and use: New perspectives on assessment and corpus analysis*, (pp. 57-78). Eurosla Monographs Series, Rome, Italy: Eurosla.
- Milton, J., & Alexiou, T. (2009). Vocabulary size and the common European framework of reference for languages. In *Vocabulary studies in first and second language acquisition* (p. 194-211). Palgrave Macmillan UK.
- Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners?. *Canadian Modern Language Review*, 63(1), 127-147.
- Milton, J., & Treffers-Daller, J. (2013). Vocabulary size revisited: the link between vocabulary size and academic achievement. *Applied Linguistics Review*, 4(1), 151-172.
- Milton, J., Wade, J., & Hopkins, N. (2010). 6. Aural word recognition and oral competence in English as a foreign language. In R. C. Beltran, C. Abello-Contesse, & M. del mar Torreblanca-Lopez, (Eds.). *Insights into non-native vocabulary teaching and learning* (pp. 83-98). Multilingual Matters.

- Mochida, A. & Harrington, M. (2006). The Yes/No test as a measure of receptive vocabulary knowledge. *Language Testing*, 23(1), 73-98.
- Molina, M. T. L. M. (2009). A Computer-Adaptive Vocabulary Test. *Indian Journal of Applied Linguistics*, 35(1), 121-138.
- Moon, R. (1997). Vocabulary connections: Multi-word items in English. In N. Schmitt, & M. McCarthy, *Vocabulary: Description, acquisition and pedagogy* (pp. 40-63). Cambridge, UK: Cambridge University Press.
- Nagy, W. E. (1997). On the role of context in first- and second language vocabulary learning. In N. Schmitt, & M. McCarthy, *Vocabulary: Description, acquisition and pedagogy*, (pp. 64-83). Cambridge, UK: Cambridge University Press.
- Nagy, W. E. & Anderson, R. C. (1984). How many words are there in printed school English?. *Reading Research Quarterly*, 19, 304-330.
- Nagy, W. E., Diakidoy, I. A. N., & Anderson, R. C. (1993). The acquisition of morphology: Learning the contribution of suffixes to the meanings of derivatives. *Journal of Literacy Research*, 25(2), 155-170.
- Nation, I.S.P. (1983) *Testing and teaching vocabulary. Guidelines* 5(1), 12-25.
- Nation, I.S.P. (1986). *Vocabulary lists: words, affixes, and stems*. Wellington, NZ: English Language Institute, Victoria University of Wellington.
- Nation, I.S.P. (1990) *Teaching and learning vocabulary*. New York: Heinle and Heinle.
- Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language*, (pp. 3-13). Amsterdam: John Benjamins.

- Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening?. *Canadian Modern Language Review*, 63(1), 59-82.
- Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, 31(7), 9-13.
- Nation, P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt, & M. McCarthy, *Vocabulary: Description, acquisition and pedagogy*, (pp. 6-19). Cambridge, UK: Cambridge University Press.
- Noro, T. (2002). The roles of depth and breadth of vocabulary knowledge in reading comprehension in EFL. *ARELE*, 13, 71-80.
- Olinghouse, N. G., & Leaird, J. T. (2009). The relationship between measures of vocabulary and narrative writing quality in second-and fourth-grade students. *Reading and Writing*, 22(5), 545-565.
- Pearson, P. D., Hiebert, E. H., & Kamil, M. L. (2007). Vocabulary assessment: What we know and what we need to learn. *Reading Research Quarterly*, 42(2), 282-296.
- Praninskas, J. (1972). *American University Word List*. London: Longman.
- Qian, D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review*, 56(2), 282-308.
- Qian, D. D. (2008). From single words to passages: Contextual effects on predictive power of vocabulary measures for assessing reading performance. *Language Assessment Quarterly*, 5(1), 1-19.
- Qian, D. D., & Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing*, 21(1), 28-52.

- Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC nonword database. *The Quarterly Journal of Experimental Psychology: Section A*, 55(4), 1339-1362.
- Read, J. (1988). Measuring the Vocabulary Knowledge of Second Language Learners. *RELC Journal*, 19(2), 12-25.
- Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, 10(3), 355-371.
- Read, J. (2000). *Assessing Vocabulary*. Cambridge: Cambridge University Press.
- Read, J. (2007). Second language vocabulary assessment: Current practices and new directions. *IJES, International Journal of English Studies*, 7(2), 105-126.
- Read, J. (2013). Second language vocabulary assessment. *Language Teaching*, 46(01), 41-52.
- Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing*, 18(1), 1-32.
- Schmitt, N. (1994). Vocabulary testing: Questions for test development with six examples of tests of vocabulary size and depth. *Thai TESOL Bulletin*, 6(2), 9-16.
- Schmitt, N. (2014). Size and Depth of Vocabulary Knowledge: What the Research Shows. *Language Learning*, 64(4), 913-951.
- Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework. *Studies in Second Language Acquisition*, 19(01), 17-36.
- Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, 47(04), 484-503.
- 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.

- Schmitt, N., & Zimmerman, C. B. (2002). Derivative word forms: What do learners know?. *TESOL Quarterly*, 36(2), 145-171.
- Schonell, F. J., Meddleton, I. G., Shaw, B. A., Routh, M., Popham, D., Gill, G., Mackrell, G., & Stephens, C. (1956). *A Study of Oral Vocabulary of Adults*. Brisbane and London: University of Queensland Press/University of London Press.
- Schoonen, R., & Verhallen, M. (2008). The assessment of deep word knowledge in young first and second language learners. *Language Testing*, 25(2), 211-236.
- Shillaw, J. (1996). The application of Rasch modelling to yes/no vocabulary tests. *Vocabulary Acquisition Research Group*. Discussion document number js96a, available at www.swan.ac.uk/cals/vlibrary/js96a.htm
- Shimamoto, T. (2000). An analysis of receptive vocabulary knowledge: Depth versus breadth. *JABAET Journal*, 4, 69-80.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. London: Routledge.
- Singleton, D. (1999) *Exploring the second language mental lexicon*. Cambridge: Cambridge University Press.
- Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, 36(2), 139-152.
- Stalnaker, J. M., & Kurath, W. (1935). A comparison of two types of foreign language vocabulary test. *Journal of Educational Psychology*, 26(6), 435.
- Stein, G. (1913). Sacred Emily. Retrieved February 22, 2016, from <http://www.lettersofnote.com/p/sacred-emily-by-gertrude-stein.html>

- Sutarsyah, C., Nation, P., & Kennedy, G. (1994). How useful is EAP vocabulary for ESP? A corpus based case study. *RELC Journal*, 25(2), 34-50.
- Takala, S. (1985). *Evaluation of students' knowledge of English vocabulary in the Finnish comprehensive school*. University Microfilms. Finland: Institute of Educational Research.
- Thorndike, E. L., & Lorge, I. (1944). *The Teacher's Word Book of 30,000 Words*. New York: Teachers College, Columbia University.
- Tseng, W. T., & Schmitt, N. (2008). Toward a model of motivated vocabulary learning: A structural equation modeling approach. *Language Learning*, 58(2), 357-400.
- Tyler, A., & Nagy, W. (1989). The acquisition of English derivational morphology. *Journal of memory and language*, 28(6), 649-667.
- Van Zeeland, H. (2013). *Second language vocabulary knowledge in and from listening*. Unpublished doctoral dissertation. University of Nottingham, UK.
- Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics*, 22(02), 217-234.
- Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, 30(01), 79-95.
- Webb, S. A., & Sasao, Y. (2013). New directions in vocabulary testing. *RELC Journal*, 44(3), 263-277.
- Webster's Ninth New Collegiate Dictionary*. (1988). Springfield, MA; Merriam-Webster Inc.
- Wesche, M., & Paribakht, T. S. (1996). Assessing Second Language Vocabulary Knowledge: Depth Versus Breadth. *Canadian Modern Language Review*, 53(1), 13-40.

- West, M. (1953). *A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. London: Addison-Wesley Longman Limited.
- Wilks, C., & Meara, P. (2002). Untangling word webs: graph theory and the notion of density in second language word association networks. *Second Language Research*, 18(4), 303-324.
- WordNet. (2003). Princeton University electronic database (Version 2.0). Retrieved February 11, 2004 and February 26, 2016 from <http://wordnet.princeton.edu/obtain>.
- Wright, B. D. (1996). Reliability and separation. *Rasch Measurement Transactions*, 9(4), 472.
- Wright, C. W. (1965). *An English Word Count*. Government Printer, South Africa.
- Zimmerman, J., Broder, P. K., Shaughnessy, J. J., & Underwood, B. J. (1977). A recognition test of vocabulary using signal-detection measures, and some correlates of word and nonword recognition. *Intelligence*, 1(1), 5-31.
- Zimmerman, K. J. (2004). *The Role of Vocabulary Size in Assessing Second Language Proficiency* (Master's Thesis, Brigham Young University, Provo, Utah). Retrieved January 3, 2016 from <http://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=1577&context=etd>

Appendix A: Vocabulary of American English Size Test Format and Sample Items

What is your full name (Last, First)? ex. Smith, Joseph

What is your native or first language?

Albanian/Shqip

Cambodian/Khmer/□□□□□□□□

Cantonese/廣東話/广东话

Cebuano/Binisaya

Chuukese/Trukese

Czech/čeština

Fijian/Nā Vosa Vakaviti

Filipino/Tagalog

Finnish/Suomen kieli

French/Français

German/Deutsch

Haitian-Creole/Kreyòl

Hungarian/Magyar

Indonesian/Bahasa Indonesia

Italian/Italiano

Japanese/日本語

Kiribati/Taetae ni Kiribati

Korean/한국어

Kuanua/Ye'kuana

Lao/Phasa lao/□□□□□□□

Malay/Bahasa Melayu/الليوب هاس

Mandarin/ 國語/国语

Mongolian/Монгол хэл/ᠮᠣᠩᠭᠣᠯ ᠬᠡᠯ

Norwegian/Norsk

Pingelapese/Pingelap

Pohnpeian/Ponapean

Portuguese/Português

Romanian/Română

Russian/русский язык

Samoan/Gagana Sāmoa

Spanish/Español

Swahili/Kiswahili

Swedish/Svenska

Thai/ภาษาไทย

Tongan/Lea Fakatonga

Turkish/Türkçe

Waray-Waray/Samar-Leyte

What is your gender?

- ☐ Male
- ☐ Female

How old are you?

How many years have you studied English?

Please read all the instructions. Test takers usually take 30 minutes to complete the 250 questions of this test. If you wish, you may stop taking the test at any time.

In this test, you will be shown real and fake words. Each page will have one word. If that word is a real word, press 'y' for 'Yes' on the keyboard or 'n' for 'No' on the keyboard if it is not a real word. It is important to note that guessing will hurt your score. If you are not sure if the word is a real word or a fake word, mark it as a fake word. You must answer every question on the test.

You are not being timed so take as much time as you want as you answer the questions. You will now see a set of practice items to help you before you begin the real test. If you have any questions, please ask them to the test administrator now.

Did you read and understand all of the instructions? If not, please go back and read them now.

- ☐ Yes
- ☐ No

Practice Items

go

- ☐ Yes (y)
- ☐ No (n)

laniff

- ☐ Yes (y)
- ☐ No (n)

cordim

- ☐ Yes (y)
- ☐ No (n)

You are now finished with the practice items. You may now begin the test.

The list of words and pseudowords used in the real test are found below.

Real Words (510 Words in Total)

Rank Lemma (Part of Speech)

1	the (Article)
2	be (Verb)
11	I (Pronoun)
21	they (Pronoun)
31	she (Pronoun)
41	would (Verb)
51	one (Number)
61	me (Pronoun)
71	could (Verb)
81	more (Adverb)
91	more (Determiner)
101	only (Adverb)
111	woman (Noun)
121	should (Verb)
131	ask (Verb)
141	high (Adjective)
151	put (Verb)
161	same (Determiner)
171	problem (Noun)
181	place (Noun)
191	system (Noun)
201	government (Noun)
211	point (Noun)
221	all (Adverb)
231	national (Adjective)
241	book (Noun)
251	head (Noun)
261	long (Adverb)
271	power (Noun)
281	stand (Verb)
291	almost (Adverb)
301	white (Adjective)
311	idea (Noun)
321	whether (Conjunction)
331	anything (Pronoun)
341	office (Noun)

351	party (Noun)
361	win (Verb)
371	teacher (Noun)
381	second (Number)
391	process (Noun)
401	serve (Verb)
411	oh (Interjection)
421	behind (Preposition)
431	class (Noun)
441	pass (Verb)
451	role (Noun)
461	drug (Noun)
471	pull (Verb)
481	son (Noun)
491	arm (Noun)
501	building (Noun)
502	action (Noun)
511	early (Adverb)
521	space (Noun)
531	couple (Noun)
541	court (Noun)
551	industry (Noun)
561	quite (Adverb)
571	wall (Noun)
581	open (Adjective)
591	attention (Noun)
601	cause (Verb)
611	culture (Noun)
621	hundred (Number)
631	place (Verb)
641	material (Noun)
651	thousand (Number)
661	security (Noun)
671	officer (Noun)
681	goal (Noun)
691	plan (Verb)
701	reduce (Verb)
711	share (Verb)
721	hot (Adjective)
731	article (Noun)
741	career (Noun)
751	lie (Verb)
761	list (Noun)

771	left (Adjective)
781	particularly (Adverb)
791	attack (Noun)
801	election (Noun)
811	arrive (Verb)
821	glass (Noun)
831	ok (Adverb)
841	gun (Noun)
851	truth (Noun)
861	rather (Adverb)
871	design (Verb)
881	sound (Verb)
891	green (Adjective)
901	that (Adverb)
911	tonight (Adverb)
921	respond (Verb)
931	employee (Noun)
941	wide (Adjective)
951	structure (Noun)
961	treat (Verb)
971	worry (Verb)
981	writer (Noun)
991	dream (Noun)
1001	somebody (Pronoun)
1002	magazine (Noun)
1011	fall (Noun)
1021	agent (Noun)
1031	test (Verb)
1041	investment (Noun)
1051	civil (Adjective)
1061	mouth (Noun)
1071	score (Noun)
1081	relate (Verb)
1091	senior (Adjective)
1101	speech (Noun)
1111	global (Adjective)
1121	release (Verb)
1131	version (Noun)
1141	hurt (Verb)
1151	plane (Noun)
1161	perfect (Adjective)
1171	vote (Verb)
1181	spirit (Noun)

1191	brain (Noun)
1201	battle (Noun)
1211	stick (Verb)
1221	ship (Noun)
1231	park (Noun)
1241	truck (Noun)
1251	refuse (Verb)
1261	club (Noun)
1271	shape (Noun)
1281	band (Noun)
1291	demand (Noun)
1301	facility (Noun)
1311	basis (Noun)
1321	feed (Verb)
1331	river (Noun)
1341	ear (Noun)
1351	gather (Verb)
1361	aspect (Noun)
1371	mean (Noun)
1381	measure (Verb)
1391	engage (Verb)
1401	youth (Noun)
1411	apparently (Adverb)
1421	intelligence (Noun)
1431	context (Noun)
1441	dress (Verb)
1451	average (Noun)
1461	dangerous (Adjective)
1471	Internet (Noun)
1481	finding (Noun)
1491	famous (Adjective)
1501	cut (Noun)
1502	actor (Noun)
1511	circle (Noun)
1521	train (Verb)
1531	hate (Verb)
1541	intend (Verb)
1551	danger (Noun)
1561	northern (Adjective)
1571	climb (Verb)
1581	ticket (Noun)
1591	lunch (Noun)
1601	enemy (Noun)

1611	impossible (Adjective)
1621	commitment (Noun)
1631	consequence (Noun)
1641	connect (Verb)
1651	regional (Adjective)
1661	theme (Noun)
1671	yellow (Adjective)
1681	regulation (Noun)
1691	appearance (Noun)
1701	anymore (Adverb)
1711	but (Preposition)
1721	muscle (Noun)
1731	cash (Noun)
1741	content (Noun)
1751	setting (Noun)
1761	duty (Noun)
1771	slow (Adjective)
1781	shirt (Noun)
1791	snow (Noun)
1801	soil (Noun)
1811	golf (Noun)
1821	governor (Noun)
1831	golden (Adjective)
1841	long (Conjunction)
1851	trust (Verb)
1861	confirm (Verb)
1871	issue (Verb)
1881	debt (Noun)
1891	file (Verb)
1901	now (Conjunction)
1911	clean (Verb)
1921	totally (Adverb)
1931	rest (Verb)
1941	aim (Verb)
1951	overall (Adjective)
1961	league (Noun)
1971	tie (Noun)
1981	apart (Adverb)
1991	beside (Preposition)
2001	definitely (Adverb)
2002	bomb (Noun)
2011	invest (Verb)
2021	solid (Adjective)

2031	either (Determiner)
2041	talent (Noun)
2051	height (Noun)
2061	creative (Adjective)
2071	live (Adjective)
2081	weak (Adjective)
2091	passenger (Noun)
2101	lab (Noun)
2111	journalist (Noun)
2121	permit (Verb)
2131	dramatic (Adjective)
2141	airline (Noun)
2151	initiative (Noun)
2161	post (Noun)
2171	violent (Adjective)
2181	layer (Noun)
2191	portion (Noun)
2201	display (Noun)
2211	shall (Verb)
2221	print (Noun)
2231	atmosphere (Noun)
2241	discovery (Noun)
2251	grand (Adjective)
2261	coat (Noun)
2271	online (Adjective)
2281	jacket (Noun)
2291	substance (Noun)
2301	gene (Noun)
2311	employer (Noun)
2321	competitive (Adjective)
2331	another (Pronoun)
2341	coach (Verb)
2351	spending (Noun)
2361	emphasis (Noun)
2371	digital (Adjective)
2381	increasing (Adjective)
2391	twin (Noun)
2401	so-called (Adjective)
2411	light (Verb)
2421	block (Verb)
2431	confront (Verb)
2441	personnel (Noun)
2451	perfectly (Adverb)

2461	watch (Noun)
2471	salary (Noun)
2481	plant (Verb)
2491	assist (Verb)
2501	occasionally (Adverb)
2502	mayor (Noun)
2511	grandmother (Noun)
2521	install (Verb)
2531	concert (Noun)
2541	roll (Noun)
2551	speaker (Noun)
2561	pop (Verb)
2571	depth (Noun)
2581	pack (Noun)
2591	dealer (Noun)
2601	routine (Noun)
2611	activist (Noun)
2621	valuable (Adjective)
2631	developing (Adjective)
2641	extraordinary (Adjective)
2651	clock (Noun)
2661	button (Noun)
2671	portrait (Noun)
2681	burden (Noun)
2691	lost (Adjective)
2701	destruction (Noun)
2711	apple (Noun)
2721	dispute (Noun)
2731	initially (Adverb)
2741	retain (Verb)
2751	expansion (Noun)
2761	solar (Adjective)
2771	strip (Noun)
2781	balance (Verb)
2791	guarantee (Verb)
2801	awareness (Noun)
2811	dialogue (Noun)
2821	delivery (Noun)
2831	relevant (Adjective)
2841	partly (Adverb)
2851	justify (Verb)
2861	lie (Noun)
2871	originally (Adverb)

2881	external (Adjective)
2891	shelter (Noun)
2901	net (Noun)
2911	target (Verb)
2921	introduction (Noun)
2931	steady (Adjective)
2941	nowhere (Adverb)
2951	correspondent (Noun)
2961	mm-hmm (Interjection)
2971	buyer (Noun)
2981	stability (Noun)
2991	psychology (Noun)
3001	slight (Adjective)
3002	math (Noun)
3011	store (Verb)
3021	briefly (Adverb)
3031	besides (Adverb)
3041	preference (Noun)
3051	ski (Noun)
3061	porch (Noun)
3071	scandal (Noun)
3081	contest (Noun)
3091	publisher (Noun)
3101	tennis (Noun)
3111	rate (Verb)
3121	Catholic (Noun)
3131	curious (Adjective)
3141	taxpayer (Noun)
3151	laboratory (Noun)
3161	demonstration (Noun)
3171	cabin (Noun)
3181	manufacturing (Noun)
3191	boom (Noun)
3201	sense (Verb)
3211	extension (Noun)
3221	cluster (Noun)
3231	operator (Noun)
3241	weekly (Adjective)
3251	seize (Verb)
3261	frustration (Noun)
3271	correct (Verb)
3281	powder (Noun)
3291	cooking (Noun)

3301	rhythm (Noun)
3311	transformation (Noun)
3321	intensity (Noun)
3331	concrete (Adjective)
3341	recording (Noun)
3351	reinforce (Verb)
3361	criminal (Noun)
3371	sigh (Verb)
3381	lap (Noun)
3391	surprisingly (Adverb)
3401	boyfriend (Noun)
3411	upset (Verb)
3421	invent (Verb)
3431	trading (Noun)
3441	counsel (Noun)
3451	compound (Noun)
3461	serving (Noun)
3471	pleased (Adjective)
3481	slam (Verb)
3491	essence (Noun)
3501	pitcher (Noun)
3502	retail (Adjective)
3511	pig (Noun)
3521	reverse (Verb)
3531	Roman (Adjective)
3541	tip (Verb)
3551	van (Noun)
3561	swallow (Verb)
3571	enforce (Verb)
3581	frankly (Adverb)
3591	monster (Noun)
3601	integration (Noun)
3611	ownership (Noun)
3621	forgive (Verb)
3631	prosecution (Noun)
3641	medium (Adjective)
3651	wrist (Noun)
3661	walking (Noun)
3671	ideology (Noun)
3681	chronic (Adjective)
3691	pad (Noun)
3701	colony (Noun)
3711	particle (Noun)

3721 alarm (Noun)
3731 research (Verb)
3741 rhetoric (Noun)
3751 pause (Noun)
3761 matter (Determiner)
3771 corruption (Noun)
3781 behavioral (Adjective)
3791 suspicion (Noun)
3801 pleasant (Adjective)
3811 theoretical (Adjective)
3821 hook (Noun)
3831 altogether (Adverb)
3841 invisible (Adjective)
3851 exhibit (Noun)
3861 carbohydrate (Noun)
3871 magic (Noun)
3881 opera (Noun)
3891 giant (Noun)
3901 elevator (Noun)
3911 fist (Noun)
3921 thereby (Adverb)
3931 practically (Adverb)
3941 realm (Noun)
3951 accounting (Noun)
3961 worry (Noun)
3971 diplomatic (Adjective)
3981 confess (Verb)
3991 prevention (Noun)
4001 civic (Adjective)
4002 magnitude (Noun)
4011 angel (Noun)
4021 prohibit (Verb)
4031 outstanding (Adjective)
4041 tide (Noun)
4051 cook (Noun)
4061 trap (Noun)
4071 coastal (Adjective)
4081 way (Adverb)
4091 Dutch (Adjective)
4101 bid (Noun)
4111 shock (Verb)
4121 diabetes (Noun)
4131 ours (Pronoun)

4141	buddy (Noun)
4151	dilemma (Noun)
4161	stadium (Noun)
4171	condemn (Verb)
4181	courtroom (Noun)
4191	productivity (Noun)
4201	combined (Adjective)
4211	orbit (Noun)
4221	rent (Noun)
4231	sheer (Adjective)
4241	clip (Noun)
4251	empire (Noun)
4261	web (Noun)
4271	draft (Verb)
4281	verdict (Noun)
4291	puzzle (Noun)
4301	utilize (Verb)
4311	near (Adverb)
4321	ambition (Noun)
4331	metropolitan (Adjective)
4341	helmet (Noun)
4351	minimal (Adjective)
4361	flexibility (Noun)
4371	experienced (Adjective)
4381	upset (Adjective)
4391	supplier (Noun)
4401	associate (Noun)
4411	fever (Noun)
4421	dried (Adjective)
4431	comprise (Verb)
4441	unprecedented (Adjective)
4451	counter (Verb)
4461	banker (Noun)
4471	speculation (Noun)
4481	swimming (Noun)
4491	someday (Adverb)
4501	ideal (Noun)
4502	colorful (Adjective)
4511	cease (Verb)
4521	dot (Noun)
4531	marketplace (Noun)
4541	planner (Noun)
4551	invade (Verb)

4561 ambassador (Noun)
4571 likewise (Adverb)
4581 publicity (Noun)
4591 builder (Noun)
4601 artifact (Noun)
4611 rib (Noun)
4621 ash (Noun)
4631 halfway (Adverb)
4641 carrot (Noun)
4651 blink (Verb)
4661 rain (Verb)
4671 peanut (Noun)
4681 comply (Verb)
4691 awake (Adjective)
4701 butt (Noun)
4711 liver (Noun)
4721 banana (Noun)
4731 plain (Noun)
4741 brutal (Adjective)
4751 soar (Verb)
4761 unhappy (Adjective)
4771 routinely (Adverb)
4781 objection (Noun)
4791 rental (Noun)
4801 suitable (Adjective)
4811 regard (Noun)
4821 fare (Noun)
4831 leave (Noun)
4841 broadcast (Verb)
4851 spark (Verb)
4861 substantially (Adverb)
4871 surveillance (Noun)
4881 soak (Verb)
4891 within (Adverb)
4901 brave (Adjective)
4911 dense (Adjective)
4921 sudden (Adverb)
4931 economically (Adverb)
4941 weave (Verb)
4951 skilled (Adjective)
4961 fog (Noun)
4971 butterfly (Noun)
4981 dictate (Verb)

4991 pistol (Noun)

Pseudowords (340 in total)

1000 Word Level

cantileen

ralling

contortal

lapidoscope

gandle

dowrickfic

dogmatile

aistrope

justal

youde

cotargent

ballotage

renigrade

oligation

bundock

lore

investebrate

certical

ventrice

reconcilant

lunorous

pocock

loving

redivate

misabrogate

skene

callisthemia

maidment

pardoe

wallage

proctalise

climaximal

nonagrate

lannery

retrogradient

equalic

cordle

elode

ordinisation
spedding
roscrow
mabey
factile
twose
canarify
martlew
multiplify
curify
arain
demaine
prelatoriat
kitely
acquince
colliver
werrell
devoidance
finalism
willment
innoculism
disportal
batstone
frutal
minestory
gasson
strategy
proscratify
oxylate
degate (Verb
2000 Word Level
condimented
loveridge
rudge
descript
reservory
horozone
almanical
amagran
abrogative
swithin
cheatle
nichee

restificate
antile
logalation
kellett
worrall
beautitude
keable
linocat
amelicant
presential
centipath
limbrick
dumbrill
majury
hignall
spratling
defunctionary
bargery
libidnize
extravagate
galpin
benevolate
hudd
burse
hermantic
ashill
bowring
mynott
sedgebeer
flamboyment
fumicant
skelding
mascarate
mollet
webbert
dyslaxative
primality
challinor
matsell
quorant
lampard
motivize
agrinomy

batteric
leaity
auner
provisual
preconagulative
louvragé
chlorosate
watchorn
practicate
yallop
lamble
buttle
horobin

3000 Word Level

berrow
limidate
pernicate
humberoid
eluctant
detailoring
stimulcrate
bastionate
asslam
seclunar
churchlow
neutratriation
refurge
carotic
kearle
paralogue
andow
crucialate
floralate
dagless
kerkin
barmion
recentile
remonic
moule
jemmett
hegedoxy
attard
deliction

troake
fancett
ionopose
gummer
contrivial
distantial
contrammand
surman
leopradate
rhind
candish
bowring
farinize
lediard
laudalize
ebullible
savourite
bastin
absolvention
tearle
coath
garrisotte
escrotal
eckett
damnifest
sacrumate
tindle
cardination
ackery
dring
baptistal
atribus
wintle
captivise
interisation
rainish
joice
hoult
whitrow
4000 Word Level
fluctual
cambule
ridout

charactal
hapgood
menstruable
batcock
hemiaphrodite
peritonic
savery
ashment
boobier
viggers
doole
amphlett
carpin
bickle
samphirate
obsolation
annobile
dyment
cockram
expostulant
loaring
decorite
causticate
graduable
transcendiary
shattock
warman
perceptacle
prowt
suddery
acklon
mastaphitis
hawther
vardy
genderation
coppard
schismal
biforcal
waygood
gotargent
negalogue
disaddle
pimlott

hislop
hermantic
mabbitt
manomize
localitude
franternism
kellett
xenostrophic
eade
barnish
professive
wellstead
algoric
hospite
tandulous
legitimal
beament
cymballic
presuppository
microphant
duffin
oestrogeny
5000 Word Level
draconite
combustulate
scudamore
homoglyph
abrogative
nickling
charlett
woolnough
haque
investebrate
arkless
logam
mourant
whitelock
incarminate
saratogal
expostulant
sacrumate
dictalate
jotham

rundle
correctivate
nebulate
frequent
briochemistry
amroth
proctalise
condick
haswell
appertonal
inertible
ickard
litholect
scurrilise
baldock
porlock
cicatration
powling
aspection
conceitful
cundy
pungid
enigmanic
obsolation
rendle
brind
dunster
apricoterie
perceptacle
filterite
choreostat
gamage
ackrill
cartledge
bendall
pitten
innoculism
propend
documentate
pegler
gravology

paladine
manolect
cunnion
mabille
rudall
bodelate
bance

Appendix B: Item-Person Map

INPUT: 403 Person 850 Item REPORTED: 403 Person 850 Item 2 CATS WINSTEPS 3.92.1

```

MEASURE                                     MEASURE
<more> ----- Person +- Item ----- <rare>
5                                     5

# |
.# |
. |
.# |
4 .## +                                     4
   ### T|
   .### |
   ### |
   ##### |
   .#### |
3 .### +                                     3
   ##### S|
   .##### |
   .##### |
   .##### |
   .##### |
2 .##### T| #.                                     2
   .##### M| #####.
   .##### | #####.
   .##### | #####.
   .##### | #####.
1 .##### | S #####.                                     1
   .##### + #####.
   .##### S| #####.
   .##### | #####.
   .##### | #####.
   .##### | #####.
0 . +M #####.                                     0
   . T| #####.
   . | #####.
   . | #####.
   # | #####.
-1 + #####.                                     -1
   | S #####
   | #####.
   | #####.
   | #####.
   | .
-2 + #####.                                     -2
   | T .
   | #.
   | #####
   |
-3 + #####.                                     -3
   | ##.
   |
   |
   |
-4 + ##.                                     -4
<less> ----- Person +- Item ----- <freq>
EACH "#" IN THE Person COLUMN IS 2 Person: EACH "." IS 1
EACH "#" IN THE Item COLUMN IS 3 Item: EACH "." IS 1 TO 2
M IS THE MEAN; S IS 1 STANDARD DEVIATION FROM THE MEAN; T IS 2 STANDARD DEVIAT
MEAN

```

Appendix C: Item: Measure Map

MEASURE	Person - MAP - Item	
	<more> <rare>	
5	+	
	.	
	#	
	.	
	#	
4	+	
	.#	
	## T	
	.##	
	##	R25021-stoically
	.###	
	###	
3	+	
	#### S	F2000-descript
	.#####	R2961-mm-hmm
	.#####	F5000-whitelock
	.#####	F2000-buttlet
		R3781-behavioral
	.####	F2000-presential
		F5000-brind
		R4871-surveillance
2	+	F1000-certical
		F2000-burse
		F2000-practicat
		F3000-cardination
		R1311-basis
		R4071-coastal
		R4291-verdict
	.##### M	F1000-disportal
		F1000-multiplify
		F2000-benevolate
		F2000-provisual
		F4000-fluctual
		F4000-professive
		F5000-aspection
		F5000-doublty
		R4151-dilemma
		R4771-routinely
	#####	F1000-factile
		F3000-dring
		F3000-neutratriation
		R1821-governor
		R3251-seize
		R3391-surprisingly
		R4561-ambassador
	#####	F1000-minestory
		F2000-abrogative
		F2000-primality
		F3000-gummer
		F4000-legitimal
		F5000-correctivate
		R2301-gene
		R3301-rhythm
		R3831-altogether
		R4101-bid
		R4621-ash
	.#####	F1000-curify
		F3000-baptistal
		F3000-paralogue
		F4000-algoric
		F5000-charlett
		R1881-debt
		F1000-reconcilant
		F2000-extravagat
		F2000-rudge
		F4000-obsolation
		R3811-theoretical
		R4171-condemn
		R3741-rhetoric
		R10021-locale
		R3941-realm
		F4000-graduabile
		R3791-suspicion
		F2000-condimented
		F3000-joice
		F3000-tindle
		R30021-nonreactive
		R3351-reinforce
		R3631-prosecution
		R4681-comply
		F1000-spedding
		F2000-mascarate
		F3000-bowring
		F4000-carpin
		F4000-wellstead
		R15021-childbearing
		R2801-awareness
		R3801-pleasant
		R4001-civic
		R4311-utilize
		R4911-dense
		F2000-lampard
		F3000-contrivial
		F3000-refurge
		F4000-loaring
		F5000-propend
		R2571-depth

		R2791-guarantee	R3051-ski
		R3141-taxpayer	R3171-cabin
		R3221-cluster	R3651-wrist
		R4571-likewise	R4861-substantially
.#####	S	F1000-finalism	F1000-maidment
		F1000-strategy	F1000-wallage
		F2000-cheatle	F2000-mollet
		F2000-motivize	F2000-restificate
		F3000-candish	F3000-coath
		F3000-deliction	F3000-interisation
		F4000-barnish	F4000-bickle
		F4000-menstruable	F4000-microphant
		F4000-peritonic	F4000-solitist
		F4000-transcendary	F4000-warman
		F5000-eventualise	F5000-paladine
		F5000-perceptacle	F5000-powling
		F5000-rendle	F5000-rundle
		R2401-so-called	R3061-porch
		R3561-swallow	R3581-frankly
		R4511-cease	R4881-soak
		R4991-pistol	R741-career
1	.#####	+ F1000-cordle	F1000-degate
		F1000-equalic	F1000-fruital
		F1000-glandle	F1000-redivate
		F1000-retrogradient	F1000-ventrice
		F2000-bargery	F2000-batteric
		F2000-reservory	F3000-absolvention
		F3000-berrow	F3000-carotic
		F3000-limdate	F3000-recentile
		F3000-surman	F4000-ashment
		F4000-hospite	F4000-presuppository
		F4000-ridout	F4000-waygood
		F5000-beament	F5000-conceitful
		F5000-enigmanic	F5000-nickling
		F5000-pitten	R2361-emphasis
		R2591-dealer	R2671-portrait
		R3281-powder	R3502-retail
		R3551-van	R3571-enforce
		R3921-thereby	R40021-fornicate
		R4121-diabetes	R4341-metropolitan
		R4361-minimal	R45021-hatched
		R4531-marketplace	R4551-invade
		R4601-artifact	R4611-rib
#####	S	F1000-arain	F1000-ballotage
		F1000-batstone	F1000-innoculism
		F1000-oligation	F2000-antile
		F2000-hermantic	F2000-skelding
		F3000-rainish	F4000-decorite
		F4000-duffin	F4000-genderation
		F4000-viggers	F5000-bance
		F5000-cicatration	F5000-homoglyph
		F5000-incarminate	R1941-aim
		R1951-overall	R2611-activist
		R2731-initially	R2841-partly
		R2971-buyer	R3451-compound
		R3611-ownership	R3671-ideology
		R3821-hook	R3851-exhibit
		R3861-carbohydrate	R4091-Dutch
		R4181-courtroom	R4231-sheer
		R4431-comprise	R4711-liver
		R4751-soar	R4791-rental
		R4801-suitable	R4811-regard
		R4821-fare	R4941-weave
		R4961-fog	
.#####		F1000-cotargent	F1000-demaine
		F1000-devoidance	F1000-oxylate
		F1000-proscratify	F2000-lamble
		F3000-hoult	F3000-wintle

	F4000-suddery	F5000-arkless
	F5000-cartledge	F5000-dunster
	F2000-laminastic	F2000-loveridge
	F2000-majury	F2000-quorant
	F2000-spratling	F3000-attard
	F3000-captivise	F3000-contrammand
	F3000-crucialate	F3000-detailoring
	F5000-frequid	F5000-haswell
	R1541-intend	R1561-northern
	R1691-appearance	R1801-soil
	R1871-issue	R20021-rebuilt
	R2502-mayor	R2941-nowhere
	R3031-besides	R3071-scandal
	R3331-concrete	R3371-sigh
	R3441-counsel	R3911-fist
	R4351-helmet	R4391-supplier
	R4701-butt	R4741-brutal
.###	F1000-acquince	F1000-bundock
	F1000-colliver	F1000-lapidoscope
	F1000-mabey	F1000-renigrade
	F1000-twose	F1000-willment
	F2000-agrinomy	F2000-logalation
	F3000-atribus	F3000-bastin
	F3000-moule	F3000-pernicate
	F3000-remonic	F3000-whitrow
	F4000-beament	F4000-disaddle
	F4000-franternism	F4000-hemiaphrodite
	F5000-filterite	F5000-gravology
	F5000-investebrate	F5000-logam
	F5000-mourant	F5000-nebulate
	R2121-permit	R2231-atmosphere
	R2601-routine	R2721-dispute
	R2741-retain	R2811-dialogue
	R2891-shelter	R2931-steady
	R2951-correspondent	R3001-slight
	R3481-slam	R3491-essence
	R3681-chronic	R3691-pad
	R4002-magnitude	R4211-orbit
	R4391-upset	R4421-dried
	R4471-speculation	R4581-publicity
	R4671-peanut	R4691-awake
	R4981-dictate	
##	F1000-gasson	F1000-kitely
	F1000-lannery	F1000-lorey
	F1000-lunarus	F1000-ordinisation
	F1000-werrell	F2000-almanical
	F2000-horozone	F2000-watchorn
	F3000-ackery	F3000-andow
	F3000-churchlow	F3000-dagless
	F3000-savourite	F3000-seclunar
	F4000-batcock	F4000-biforcal
	F4000-causticate	F4000-charactal
	F4000-gotargent	F4000-xenostrophic
	F5000-combustulate	F5000-cundy
	F5000-proctalise	R1051-civil
	R1351-gather	R1381-measure
	R1391-engage	R1961-league
	R2111-journalist	R2151-initiative
	R2311-employer	R2471-salary
	R2681-burden	R2851-justify
	R2881-external	R3091-publisher
	R3181-manufacturing	R3431-trading
	R3501-pitcher	R3601-integration
	R3881-opera	R3891-giant
	R4041-tide	R4371-flexibility
	R4381-experienced	R4631-halfway
	R4651-blink	R4921-sudden
	R4931-economically	R841-gun
.	F1000-callisthemia	F1000-climaximal

F1000-elode	F1000-justal
F1000-nonagrate	F1000-roscrow
F2000-centipath	F2000-flamboyment
F2000-kellett	F2000-limbrick
F2000-webbert	F3000-eluctant
F3000-escrotal	F3000-floralate
F3000-tearle	F4000-coppard
F4000-doole	F4000-dyment
F4000-hislop	F4000-localitude
F4000-shattock	F5000-baldock
F5000-bendall	F5000-briochery
F5000-dictalate	F5000-gamage
F5000-inertible	F5000-pegler
F5000-torpedal	R1041-investment
R1101-speech	R1121-release
R1251-refuse	R1361-aspect
R1411-apparently	R1451-average
R1661-theme	R2001-definitely
R201-government	R2101-lab
R861-rather	
F1000-aistroke	F1000-cantileen
F1000-martlew	F1000-misabrogate
F1000-skene	F1000-stephonitis
F1000-youde	F2000-amagran
F2000-ashill	F2000-challinor
F2000-hudd	F2000-matsell
F3000-asslam	F3000-barmion
F3000-damnifest	F3000-humberoid
F3000-laudalize	F3000-lediard
F3000-rhind	F4000-catalypso
F4000-negalogue	F4000-pimlott
F4000-schismal	F5000-ackrill
F5000-amroth	F5000-litholect
F5000-rudall	F5000-saratogal
R1131-version	R1341-ear
R1741-content	R1761-duty
R1971-tie	R2051-height
R2351-spending	R2391-twin
R2871-originally	R3121-Catholic
R3161-demonstration	R3191-boom
R3361-criminal	R3531-Roman
R3771-corruption	R401-serve
R4131-ours	R421-behind
R4401-associate	R4641-carrot
R541-court	R551-industry
R561-quite	R851-truth
. F2000-linocat	F2000-nichee
F3000-jemmett	F3000-leopradata
F3000-troake	F4000-mabbitt
F4000-tandulous	F5000-jotham
F5000-manolect	F5000-scudamore
R1111-global	R1861-confirm
R2131-dramatic	R2771-strip
R2821-delivery	R2911-target
R3261-frustration	R3721-alarm
R3731-research	R411-oh
R4301-puzzle	R4411-fever
R4451-counter	R4731-plain
R4781-objection	R4901-brave
R4951-skilled	R4971-butterfly
R5021-lift	R531-couple
R641-material	R801-election
R941-wide	R951-structure
. F1000-dowrickfic	F1000-prelatoriat
F2000-hignall	F2000-sedgebeer
F2000-yallop	F3000-eckett
F3000-farinize	F3000-kerkin
F4000-acklon	F4000-cambule
R2261-coat	R2341-coach

		F4000-eade	F4000-hapgood
		F5000-choreostat	F5000-draconite
		R1141-hurt	R1201-battle
		R1371-mean	R2211-shall
		R3151-laboratory	R3231-operator
		R3241-weekly	R3421-invent
		R3951-accounting	R4161-stadium
		R4221-rent	R4241-clip
		R461-drug	R591-attention
		R681-goal	R691-plan
		R731-article	R751-lie
		R871-design	
	.	F1000-pardoe	F1000-pocock
		F4000-cockram	F4000-mastaphitis
		F5000-ickard	R1211-stick
		R1401-youth	R151-put
		R1701-anymore	R1851-trust
		R1991-beside	R2002-bomb
		R2531-concert	R2561-pop
		R291-almost	R3081-contest
		R3621-forgive	R3761-matter
		R451-role	R471-pull
		R651-thousand	R811-arrive
		R921-respond	
-1	+	F3000-kearle	F5000-woolnough
		R1241-truck	R1271-shape
		R1281-band	R1431-context
		R1461-dangerous	R1511-circle
		R1531-hate	R1681-regulation
		R1811-golf	R1921-totally
		R2091-passenger	R2171-violent
		R2701-destruction	R2901-net
		R3541-tip	R361-win
		R3901-elevator	R4111-shock
		R4831-leave	R491-arm
		R611-culture	
	S	F3000-ionopose	R1002-magazine
		R1071-score	R1151-plane
		R1221-ship	R1321-feed
		R1502-actor	R1981-apart
		R2061-creative	R2081-weak
		R2541-roll	R2581-pack
		R3101-tennis	R3341-recording
		R3641-medium	R441-pass
		R4481-swimming	R4541-planner
		R671-officer	R831-ok
		R991-dream	
		F2000-libidnize	R1331-river
		R1491-famous	R1551-danger
		R1611-impossible	R1751-setting
		R1781-shirt	R1831-golden
		R1891-file	R191-system
		R2041-talent	R2141-airline
		R2221-print	R2241-discovery
		R2661-button	R331-anything
		R3311-transformation	R3591-monster
		R4011-angel	R41-would
		R4321-near	R511-early
		R601-cause	R661-security
		R971-worry	
		R1231-park	R1441-dress
		R161-same	R1791-snow
		R1841-long	R2201-display
		R2251-grand	R2781-balance
		R3002-math	R3291-cooking
		R3461-serving	R4491-someday
		R501-building	R711-share
		R911-tonight	R981-writer
		R2421-block	R2551-speaker

```

      | F2000-mynott      R1001-somebody
      | R101-only         R1061-mouth
      | R11-I             R1181-spirit
      | R1591-lunch       R1671-yellow
      | R1731-cash        R1771-slow
      | R1931-rest        R2-be
      | R211-point        R2161-post
      | R231-national     R2371-digital
      | R351-party        R3751-pause
      | R2691-lost        R301-white
      | R3271-correct     R381-second
      | R3961-worry       R4081-way
      | R4661-rain        R4721-banana
      | R631-place        R821-glass
      | R91-more
-2    + R3511-pig
      | R1011-fall        R111-woman
      | R1581-ticket      R1641-connect
      | R1711-but         R2071-live
      | R221-all         R2271-online
      | R2281-jacket      R2331-another
      | R241-book         R2411-light
      | R2451-perfectly   R281-stand
      | R2921-introduction R341-office

      R4051-cook          R51-one
      R571-wall           R61-me
      R761-list           R771-left
      R81-more
      |T R1501-cut         R2511-grandmother
      | R1031-test        R121-should
      | R581-open         R71-could
      | R1471-Internet    R171-problem
      | R181-place        R1911-clean
      | R21-they          R2461-watch
      | R2481-plant       R251-head
      | R271-power        R2711-apple
      | R3011-store       R4261-web
      | R431-class        R481-son
      | R502-action       R721-hot
      | R881-sound        R901-that

-3    |
      |
      +
      | R1-the            R131-ask
      | R141-high         R261-long
      | R371-teacher      R3871-magic
      | R391-process      R891-green

      |
      |
      |
-4    + R1161-perfect     R1261-club
      | R1901-now         R2651-clock
      | R31-she           R3401-boyfriend
      | R3661-walking

      <less>|<freq>
EACH "#" IS 3: EACH "." IS 1 TO 2
R#### IS A REAL WORD; F#### IS A FAKE WORD

```