# COMPARING RECEPTIVE VOCABULARY KNOWLEDGE AND VOCABULARY PRODUCTION

**JESSICA FAST MICHEL & EMILY GAZDA PLUMB**

***University of Hawai'i at Manoa***

Vocabulary development in a second language is a complex process that has broad implications across all domains of language learning. In order for language learners to meaningfully engage with academic content in the target language, they must have a strong command of the kind of vocabulary used in an academic setting. The Vocabulary Levels Test (Nation, 1990; Beglar & Hunt, 1999), which assesses receptive vocabulary knowledge by asking learners to match lexical items to a short definition or description, is a common vocabulary assessment in academic settings. However, according to Coxhead and Nation (2001):

> For learners studying English for academic purposes, academic vocabulary is a kind of high frequency vocabulary and thus any time spent learning it is time well spent. The four major strands of a language course—meaning focused input, language focused learning, meaning focuses output, and fluency development—should all be seen as opportunities for the development of academic vocabulary knowledge, and it is important that the same words occur in each of these four strands. (p. 258)

Thus, in order to get a more balanced idea of learners' actual knowledge of academic vocabulary for both passive recognition and active output, tests for measuring it in both arenas are important.

Most studies of language learners' vocabulary knowledge have focused on only the measurement of their receptive knowledge (Beglar, 2010). Some have also considered learners' vocabulary production in a writing sample (Laufer & Nation, 1999; Zheng, 2012) and few have investigated vocabulary knowledge in the domains of listening and speaking (but see McLean, Kramer & Beglar, 2015, for a report on creating and validating a vocabulary levels listening test). For those studies that examine written vocabulary abilities, they generally focus on either passive or active measures of vocabulary. This study attempts to compare and contrast analyses of receptive and productive vocabulary size from the same group of students in order to explore how these two facets of vocabulary knowledge may manifest in different ways.

*Second Language Studies, 37*(2), Spring 2019, pp. 75-102.

***Measurement of Receptive Vocabulary Knowledge***

The Vocabulary Levels Test was originally created by Nation (1990). It consists of five subtests, each assessing a different 'level' of vocabulary knowledge, with 36 items at the 2,000 Word, 3,000 Word, 5,000 Word, 10,000 Word, and Academic Word Levels. Items are arranged in groups of three, with six possible definitions to choose from. Included is a sample item from the original version of Nation's (1990) 5,000 Word Level Test:

1. alcohol
2. apron _____ cloth worn in front to protect your clothes
3. lure
4. mess _____ stage of development
5. phase _____ musical instrument
6. plank (p. 268)

Designed to demonstrate mastery of various vocabulary levels, when taken together, these subtests work as a diagnostic instrument for students' vocabulary knowledge. However, over the years, alternative versions of the original test have been created and used for research and program placement. These alterations often include reducing or changing which levels appear on the test (to better fit the population being tested), and using classical test theory and/or other analyses to shorten test length.

The Vocabulary Levels Test has been used to assess learners' vocabulary knowledge based on the idea that their scores on any given subtest reflect their mastery of words at that level. For example, if a student gets all thirty items of the 2,000 Word Level Test correct, it can be assumed that that student has a very high comprehension of words at that level. Because the words within a cluster have very different meanings, even a small amount of knowledge about a target word's meaning should enable a student to choose the correct response. The Levels Test should, therefore, be seen as providing an indication of whether examinees have an initial knowledge of the most frequent meaning sense of each word in the test (Schmitt, Schmitt, & Clapham, 2001).

However, there are a few issues with this way of thinking. First is the assumption that knowledge of the representative items chosen for any given subtest automatically demonstrate comprehension of other/all words at that level. Correctly choosing definitions for 36 words at the 2,000 Word Level leaves 1,964 words at that level that the student is not tested on and which she may or may not recognize. Therefore, the test is not a very comprehensive measurement for

overall receptive vocabulary knowledge. In addition, this test only claims to measure receptive vocabulary knowledge and yet also claims to demonstrate a student's knowledge of vocabulary levels. This last claim is problematic, especially in academic settings, since students do not only encounter vocabulary words in written form, they also must produce those words themselves and be able to use them in context. The Vocabulary Size Test, which was developed by Nation and validated by Beglar (2010), attempts to address these issues by allowing for more detailed measurement, but the Vocabulary Levels Test remains popular because it is short and easy to administer.

### *Rasch Analysis of the Vocabulary Levels Test*

Several studies have analyzed results from the Vocabulary Levels Test using Rasch analysis. One of the earliest studies was done by Beglar and Hunt (1999), who investigated four forms of the Vocabulary Levels Test, two forms of the 2000 word frequency and university word frequency tests. Due to time constraints, they used classical test theory instead of Rasch analysis to equate the two forms of each test, and then they performed a Rasch analysis to confirm that the forms were equivalent. Additionally, the results of the Rasch analysis yielded only a handful of misfitting items, which an argument for unidimensionality and thus, they argued, construct validity.

In a later study, Schmitt, Schmitt, and Clapham (2001) used Rasch analysis, among other analytical tools, to conduct a systematic validation of the Vocabulary Levels Test. In deciding whether to use Rasch analysis, they raised the issue of local independence, an assumption made by the Rasch model that is not met by the Vocabulary Levels Test. Because test items on the Vocabulary Levels Test are clusters of six words and three definitions, the three items in each cluster are not strictly independent from each other. However, the authors decided that most test-takers treat individual items independently, so they proceeded with a Rasch analysis of the scores. Like Beglar and Hunt (1999), Schmitt, Schmitt, and Clapham (2001) equated two forms of several levels of the Vocabulary Levels Test; the later study employed Rasch analysis for this equation and the authors claim that it gave them a closer look at the forms.

Rasch analysis has been used to analyze other vocabulary tests in addition to the Vocabulary Levels Test. For example, Beglar (2010) used Rasch analysis to validate the Vocabulary Size Test, which is a multiple-choice test designed to get more detailed, nuanced information than the

Vocabulary Levels Test. Like Beglar and Hunt (1999) in their analysis of the Vocabulary Levels Test, Beglar (2010) found few misfitting items in the Vocabulary Size Test and thus argued that the test was unidimensional and made an argument for construct validity.

Rather than using Rasch analysis to validate a test, as in previous examples, Laufer, Elder, Hill, and Congdon (2004) investigated four different modes of vocabulary learning using the Computer Adapted Test of Size and Strength (CATSS): active-recognition, active-recall, passive-recognition, and passive-recall. They then performed a Rasch analysis and used the logit measures to perform statistical tests, finding significant differences between the four vocabulary modes. Because the Rasch model converts ordinal raw score data to an interval logit scale, using logits to perform statistical analysis provides more meaningful results than using raw scores.

### *Measurement of Vocabulary Production*

In contrast to receptive vocabulary knowledge measurements, instruments for productive measurement have been less explored in the literature. One common method of measuring productive vocabulary is with the Lexical Frequency Profile (LFP). This was designed by Laufer and Nation (1999) to examine "the proportions of high-frequency words, academic words and low-frequency words in learners' writing samples" (Zheng, 2012, p. 105). The profile is created by pasting a writing sample into a computer program, the most popular of which is the Lextutor Vocabulary Profiler (http://www.lextutor.ca/vp/eng/). Data from the profile includes the percentages and numbers, or "tokens," of words in the sample from four different categories: the first most frequent 1000 words (K1), the second most frequent 1000 words (K2), the Academic Word List compiled by Coxhead (1998), and everything else. These percentages and tokens make up a writing sample's LFP.

The LFP has been used extensively in studies of written vocabulary production among learners of English. For example, Cho (2007) investigated the lexical variety, as measured by LFP analysis, in 90 placement compositions written for an intensive English program. Findings from this analysis indicated no significant difference in lexical variety among students who were placed into different levels in the intensive English program. In a more longitudinal study, Zheng (2012) used LFP to measure changes in Chinese EFL students' vocabulary production over a period of ten months. Findings indicated that participants' vocabulary production stabilized over

time, and rather than using academic or difficult words, the students in this study recycled the same simple words in their compositions.

Although the LFP was initially created to analyze English language texts, it can be adapted for use in other languages. For example, East (2007) used a German version of the LFP to measure the impact of bilingual dictionaries on vocabulary production; findings indicate that bilingual dictionary use is associated with an increase in both lexical variety and inaccurate word choice. To adapt the English LFP into German, three categories were analyzed: words from a list of common words, words from a list of less common words, and words that did not belong to either list. This analysis, which mimicked the structure of the original English language LFP, is one of the only examples of LFP analysis conducted in a language other than English.

The LFP is a useful tool, but it is not without its drawbacks. Meara (2005) criticized the LFP approach to estimating productive vocabulary size based on the results of probabilistic simulations. He found that that the model was insufficiently sensitive to small changes in lexical size, and therefore argued that more sensitive instruments should be developed for the measurement of productive vocabulary knowledge. Responding to Meara's (2005) article, Laufer (2005) questioned the validity of the computer-generated data used to criticize the LFP. Laufer directly addressed the criticism that the tool is insufficiently sensitive, arguing that small changes in learners' receptive vocabularies may not register in those learners' vocabulary production; the LFP's alleged lack of sensitivity, then, is not a flaw, but a characteristic of the differences between receptive and productive vocabulary knowledge.

### *Research Questions*

There remains a gap in the existing literature about receptive and productive vocabulary measurement that this study aims to fill. First, no study has yet compared learners' vocabulary production, as measured with the LFP method of analysis, with their receptive vocabulary knowledge, as measured by a test in the style of the Vocabulary Levels Test. Second, although it has been assumed that receptive and productive vocabulary knowledge contribute in different ways to performance, this has yet to be looked at using principal components analysis. Thus, this study sets out to investigate the following research questions:

1. What is the relationship between university ESL students' productive and receptive vocabulary levels?

2. To what extent do different vocabulary levels affect performance on both receptive and productive vocabulary test forms?

These questions will be addressed by first analyzing the receptive data using Rasch and the productive data using the LFP approach to see if correlations exist between them. Then, receptive and productive test score data will be analyzed using principal components analysis.

## METHODS

### *Participants*

The participants in this study were 95 students enrolled in reading and writing classes at the English Language Institute at the University of Hawai‘i at Manoa (UHM). The participants were mixed in gender and language background, with most students coming from Japan, China, and South Korea and the rest from Spain, Iran, Costa Rica, Sweden, Germany, Vietnam, Russia, Finland, Bulgaria, Chile, Slovenia, the Philippines, the Solomon Islands, and Bangladesh. The participants were a mix of undergraduate and graduate students, and most were in their first year of study at UHM, although some were in their second year. The English Language Institute is a requirement for students at UHM who have the equivalent of paper-based TOEFL scores between 500 and 600; as a result, the participants had a relatively narrow range of English language proficiency.

### *Materials*

The receptive test materials in this study consisted of two forms (Form A and Form B) of a receptive vocabulary test. The two forms of the receptive tests were created from a version of the Vocabulary Levels Test (Nation, 1990) used in the English Language Institute for diagnostic purposes in reading courses. The test was shortened to contain only the 5,000 word, Academic, and 10,000 word levels because the 1,000 and 2,000 word levels were too easy for the population. The shortened test was administered to an intermediate writing class and analyzed with classical test theory. The best-performing items from this test administration, based on item discrimination and item facility, were selected for inclusion on Form A and Form B of the receptive test for the current study. Form A and Form B have three sets of items each consisting of three nouns, three verbs, and three adjectives for the 5,000 word, Academic, and 10,000 word

levels, for a total of nine items per subtest. In Form A, the set of adjectives in the 5,000 word level had only two items because of a printing error, resulting in a total of 26 items. Form B had 27 items.

The productive materials consisted of three essay prompts, two of which were used for analysis (Form A and Form B) and one of which (Form C) was administered to students who were to take the essay test twice because of their course schedules. Form A of the productive writing test asked students to write about whether critical thinking is important for college students, and Form B asked students to write about the differences between high school and university and how they might adjust to life as a university student. Students were given approximately 50 minutes to hand-write their essays, and they were allowed to use a dictionary. The essays were transcribed electronically and analyzed using the Lextutor Frequency Profiler (http://www.lextutor.ca/vp/eng/). Following Laufer and Nation (1999), misspelled words were corrected for analysis, but word choice errors were not.

The reading and writing classes used in this study were randomly assigned to one of the receptive test forms and one of the writing prompts. Students who were enrolled in both reading and writing courses took the receptive tests assigned to their classes, resulting in some students taking Form A twice, some taking Form B twice, and some taking both Form A and Form B. For students who took the same form twice, the scores from the first test administration were used for analysis. Students who took both Form A and Form B of the receptive tests were used as anchor persons to equate the two forms of the test using Rasch analysis. To avoid students writing on the same essay prompt twice, these students took the essay prompt assigned to their writing class. In their reading class, they wrote based on a third writing prompt, Form C, which was not used for analysis.

## RESULTS

### *Rasch Analysis of Receptive Tests*

The first analysis performed on the receptive test data was Rasch analysis, which was chosen because the Rasch model transforms raw score data into an interval logit scale. An interval scale allows for student ability levels to be compared in a meaningful way across subtests and forms. Rasch analysis also provides fit statistics that indicate how well the various items are fitting the

model. Form A and Form B of the receptive test had no anchor items but instead were put on the same scale using the 14 anchor persons who took both tests. The Rasch logit scores produced by these analyses were then used for subsequent analyses.

First, the raw score data from Form A of the receptive test were entered into the Bond and Fox Steps program (Bond & Fox, 2007). The linear ruler for Form A, shown in Figure 1, displays *persons* on the left side of the dashed line and *items* on the right side. Both person ability and item difficulty are represented in logits, which are indicated on the far left of the figure.

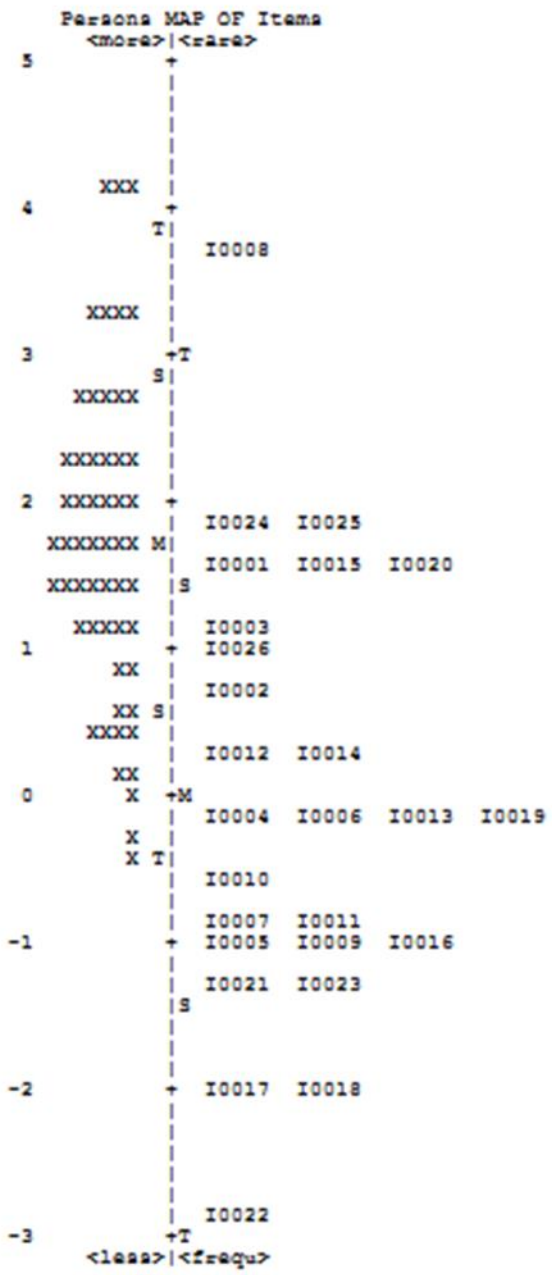



*Figure 1*. Receptive Form A Linear Ruler.

This linear ruler indicates that, in general, the test items are too easy for this group of participants. There are several items that are easier than the ability of the lowest ability person, as measured in logits, and there are persons whose abilities are higher than the items on the test can account for. Scores are normally distributed across a range of abilities, which is important because normal distribution is an underlying assumption that must be met for Rasch analysis.

Item statistics for Form A are shown in Figure 2. Item 8 is the most difficult, and item 22 is the easiest item. This is surprising, since item 8 is from the 5,000-Level subtest, while item 22 is from the 10,000-Level subtest. Intuitively, items from higher level subtests should be more difficult than those from lower level subtests, but the item logit scores for Form A do not match the expected pattern. Most items are a good fit to the Rasch model, with the notable exception of item 25, which is underfitting with a *z*-standard infit statistic of -3.2. However, the rest of the items fall within the acceptable fit range of -2 to +2.

| ENTRY NUMBER | TOTAL SCORE | COUNT | MEASURE | MODEL S.E. | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD | PTMEA CORR. | EXACT MATCH OBS% | EXACT MATCH EXP% | Item |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8 | 9 | 56 | 3.73 | .40 | 1.07 | .4 | 1.56 | 1.2 | .28 | 85.7 | 85.2 | I0008 |
| 24 | 26 | 56 | 1.87 | .30 | 1.11 | .9 | 1.19 | 1.1 | .34 | 67.9 | 69.0 | I0024 |
| 25 | 26 | 56 | 1.87 | .30 | .67 | -3.2 | .61 | -2.7 | .71 | 89.3 | 69.0 | I0025 |
| 1 | 29 | 56 | 1.60 | .30 | 1.03 | .3 | 1.05 | .3 | .41 | 71.4 | 68.3 | I0001 |
| 15 | 30 | 56 | 1.51 | .30 | 1.22 | 1.9 | 1.27 | 1.5 | .25 | 58.9 | 68.5 | I0015 |
| 20 | 30 | 56 | 1.51 | .30 | .90 | -.9 | .86 | -.8 | .52 | 69.6 | 68.5 | I0020 |
| 3 | 34 | 56 | 1.15 | .30 | 1.19 | 1.5 | 1.21 | 1.0 | .27 | 64.3 | 69.9 | I0003 |
| 26 | 35 | 56 | 1.06 | .31 | 1.10 | .8 | 1.04 | .2 | .35 | 64.3 | 70.5 | I0026 |
| 2 | 39 | 56 | .67 | .32 | 1.16 | 1.2 | 1.59 | 2.0 | .23 | 71.4 | 73.9 | I0002 |
| 12 | 43 | 56 | .24 | .34 | .96 | -.1 | .77 | -.6 | .42 | 75.0 | 78.2 | I0012 |
| 14 | 43 | 56 | .24 | .34 | .92 | -.4 | .81 | -.5 | .44 | 82.1 | 78.2 | I0014 |
| 4 | 46 | 56 | -.15 | .37 | 1.01 | .1 | 1.13 | .4 | .30 | 85.7 | 82.5 | I0004 |
| 6 | 46 | 56 | -.15 | .37 | .81 | -.9 | .61 | -.9 | .50 | 85.7 | 82.5 | I0006 |
| 13 | 46 | 56 | -.15 | .37 | 1.20 | 1.0 | 1.10 | .4 | .19 | 78.6 | 82.5 | I0013 |
| 19 | 46 | 56 | -.15 | .37 | .93 | -.3 | .68 | -.7 | .42 | 82.1 | 82.5 | I0019 |
| 10 | 49 | 56 | -.62 | .43 | .98 | .0 | .75 | -.3 | .33 | 87.5 | 87.5 | I0010 |
| 7 | 50 | 56 | -.81 | .45 | .82 | -.5 | .47 | -.9 | .45 | 89.3 | 89.3 | I0007 |
| 11 | 50 | 56 | -.81 | .45 | .99 | .1 | .89 | .0 | .28 | 89.3 | 89.3 | I0011 |
| 5 | 51 | 56 | -1.03 | .49 | .96 | .0 | .80 | -.1 | .28 | 91.1 | 91.1 | I0005 |
| 9 | 51 | 56 | -1.03 | .49 | .89 | -.2 | .48 | -.7 | .39 | 91.1 | 91.1 | I0009 |
| 16 | 51 | 56 | -1.03 | .49 | 1.12 | .5 | 1.08 | .3 | .15 | 91.1 | 91.1 | I0016 |
| 21 | 52 | 56 | -1.30 | .54 | .96 | .0 | 2.24 | 1.4 | .17 | 92.9 | 92.9 | I0021 |
| 23 | 52 | 56 | -1.30 | .54 | .87 | -.2 | .52 | -.4 | .36 | 92.9 | 92.9 | I0023 |
| 17 | 54 | 56 | -2.07 | .73 | .99 | .2 | .53 | -.2 | .21 | 96.4 | 96.4 | I0017 |
| 18 | 54 | 56 | -2.07 | .73 | 1.04 | .3 | 1.64 | .9 | .08 | 96.4 | 96.4 | I0018 |
| 22 | 55 | 56 | -2.80 | 1.02 | .95 | .3 | .30 | -.3 | .22 | 98.2 | 98.2 | I0022 |
| MEAN | 42.2 | 56.0 | .00 | .44 | .99 | .1 | .97 | .1 | | 82.6 | 82.5 | |
| S.D. | 11.3 | .0 | 1.46 | .17 | .13 | .9 | .43 | 1.0 | | 11.1 | 9.9 | |

*Figure 2*. Receptive Form A Item Statistics, Measure Order.

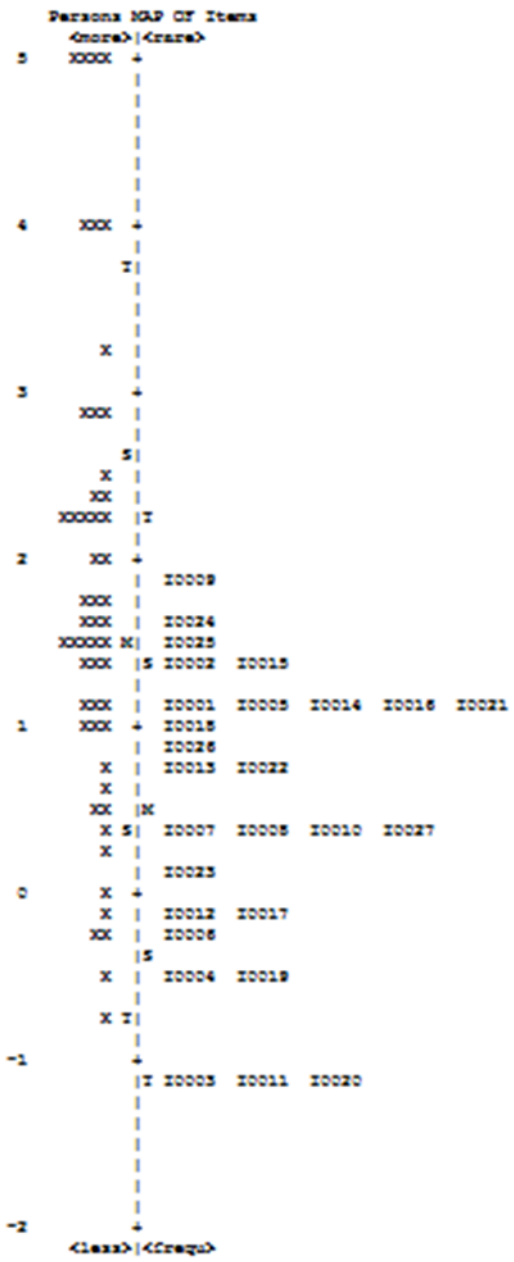



*Figure 3*. Receptive Form B Linear Ruler.

Form B was anchored to the logit scale from the Rasch analysis of Form A using the Form A logit scores from the 14 participants who took both forms of the receptive test. Anchor people function similarly to anchor items, which are used commonly in language test Rasch analysis (see Beglar, 2010). The purpose of an anchor in Rasch analysis is to compare multiple forms of a test and put all persons and items on the same logit scale (Bond & Fox, 2015). The anchor person

scores from Form A ranged from 2.32 to -0.26 logits, so the anchor scores covered a wide range of abilities. The linear ruler for Form B, shown in Figure 3, shows the distribution of the person and item logit scores.

Item statistics for Form B can be seen in Figure 4. Like in Form A, the most difficult item (9) came from the 5,000-Level, and the most difficult item (20) came from the 10,000-Level, which is unexpected based on the assumption that less common words are more difficult for learners. Most items from Form B are a good fit to the Rasch model, with the exception of item 21, which is overfitting with a *z*-standard infit statistic of 2.3. However, the rest of the items fall within the acceptable fit range of -2 to +2.

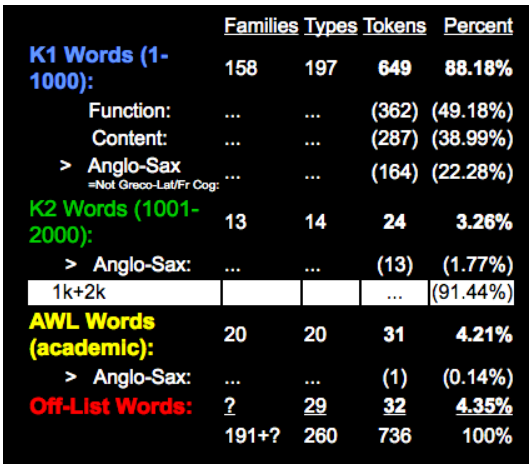
| ENTRY NUMBER | TOTAL SCORE | COUNT | MEASURE | MODEL S.E. | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD | PTMEA CORR. | EXACT MATCH OBS% | EXACT MATCH EXP% | Item |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 9 | 25 | 53 | 1.88 | .32 | 1.00 | .0 | .95 | -.2 | .52 | 67.3 | 69.2 | I0009 |
| 24 | 28 | 53 | 1.57 | .32 | 1.19 | 1.6 | 1.29 | 1.5 | .38 | 65.3 | 68.0 | I0024 |
| 25 | 29 | 53 | 1.47 | .32 | 1.17 | 1.4 | 1.19 | 1.0 | .39 | 61.2 | 68.1 | I0025 |
| 2 | 30 | 53 | 1.37 | .32 | 1.07 | .6 | 1.10 | .6 | .44 | 71.4 | 68.4 | I0002 |
| 15 | 30 | 53 | 1.37 | .32 | 1.08 | .7 | 1.25 | 1.3 | .42 | 63.3 | 68.4 | I0015 |
| 1 | 32 | 53 | 1.16 | .32 | 1.17 | 1.3 | 1.14 | .7 | .38 | 61.2 | 69.4 | I0001 |
| 5 | 32 | 53 | 1.16 | .32 | 1.25 | 1.9 | 1.20 | 1.0 | .34 | 53.1 | 69.4 | I0005 |
| 14 | 32 | 53 | 1.16 | .32 | 1.04 | .3 | 1.00 | .1 | .45 | 69.4 | 69.4 | I0014 |
| 16 | 32 | 53 | 1.16 | .32 | .91 | -.7 | .83 | -.8 | .53 | 73.5 | 69.4 | I0016 |
| 21 | 32 | 53 | 1.16 | .32 | 1.30 | 2.3 | 1.58 | 2.5 | .27 | 57.1 | 69.4 | I0021 |
| 18 | 33 | 53 | 1.06 | .32 | 1.17 | 1.3 | 1.08 | .5 | .38 | 61.2 | 70.0 | I0018 |
| 26 | 35 | 53 | .84 | .33 | 1.02 | .2 | .96 | -.1 | .44 | 71.4 | 71.7 | I0026 |
| 13 | 36 | 53 | .74 | .33 | 1.06 | .5 | 1.04 | .2 | .40 | 71.4 | 72.6 | I0013 |
| 22 | 36 | 53 | .74 | .33 | .93 | -.4 | .85 | -.5 | .48 | 75.5 | 72.6 | I0022 |
| 7 | 39 | 53 | .39 | .35 | 1.03 | .2 | .87 | -.3 | .41 | 77.6 | 75.6 | I0007 |
| 8 | 39 | 53 | .39 | .35 | .89 | -.6 | .90 | -.2 | .46 | 85.7 | 75.6 | I0008 |
| 10 | 39 | 53 | .39 | .35 | .95 | -.2 | .96 | .0 | .43 | 77.6 | 75.6 | I0010 |
| 27 | 39 | 53 | .39 | .35 | 1.18 | 1.1 | 1.04 | .2 | .32 | 69.4 | 75.6 | I0027 |
| 23 | 41 | 53 | .14 | .36 | .78 | -1.2 | .65 | -1.0 | .51 | 83.7 | 78.2 | I0023 |
| 12 | 43 | 53 | -.14 | .39 | 1.07 | .4 | .83 | -.3 | .35 | 79.6 | 80.7 | I0012 |
| 17 | 43 | 53 | -.14 | .39 | .94 | -.2 | .74 | -.5 | .41 | 83.7 | 80.7 | I0017 |
| 6 | 44 | 53 | -.30 | .40 | 1.03 | .2 | .96 | .1 | .33 | 85.7 | 82.2 | I0006 |
| 4 | 45 | 53 | -.46 | .42 | .95 | -.1 | 2.55 | 2.3 | .28 | 87.8 | 83.9 | I0004 |
| 19 | 45 | 53 | -.46 | .42 | 1.10 | .5 | 1.76 | 1.4 | .22 | 83.7 | 83.9 | I0019 |
| 3 | 48 | 53 | -1.08 | .50 | 1.06 | .3 | .92 | .1 | .24 | 89.8 | 89.8 | I0003 |
| 11 | 48 | 53 | -1.08 | .50 | .80 | -.5 | .54 | -.6 | .39 | 89.8 | 89.8 | I0011 |
| 20 | 48 | 53 | -1.08 | .50 | .79 | -.5 | .79 | -.1 | .37 | 89.8 | 89.8 | I0020 |
| MEAN | 33.1 | 49.0 | .51 | .36 | 1.03 | .4 | 1.07 | .3 | | 74.3 | 75.5 | |
| S.D. | 6.6 | .0 | .84 | .06 | .13 | .8 | .39 | .9 | | 10.6 | 7.1 | |

*Figure 4*. Receptive Form B Item Statistics, Measure Order.

**Productive Tests**

To analyze the productive test data, the Lexical Frequency Profile (LFP) was used to analyze each writing sample. The LFP provides the number of tokens, which is defined as the total number of words, and percentages for four different groups of words. K1 refers to words from the 1,000-Level, and K2 to words from the 2,000-Level. AWL refers to words from the Academic Word List, while Off-List (OL) words are any words that do not fit into any of these

lists, which are usually proper nouns or field-specific vocabulary in academic writing. A sample LFP for a participant's essay is shown in Figure 5.



| | Families | Types | Tokens | Percent |
|---|---|---|---|---|
| K1 Words (1-1000): | 158 | 197 | 649 | 88.18% |
| Function: | ... | ... | (362) | (49.18%) |
| Content: | ... | ... | (287) | (38.99%) |
| > Anglo-Sax =Not Greco-Lat/Fr Cog: | ... | ... | (164) | (22.28%) |
| K2 Words (1001-2000): | 13 | 14 | 24 | 3.26% |
| > Anglo-Sax: | ... | ... | (13) | (1.77%) |
| 1k+2k | | | ... | (91.44%) |
| AWL Words (academic): | 20 | 20 | 31 | 4.21% |
| > Anglo-Sax: | ... | ... | (1) | (0.14%) |
| Off-List Words: | ? | 29 | 32 | 4.35% |
| | 191+? | 260 | 736 | 100% |

*Figure 5*. Sample LFP Output from Lextutor for One Participant's Written Essay.

The LFP provides data about the extent to which participants produced words from the various lists. The LFP, like the Vocabulary Levels Test, uses sets of vocabulary words as its foundation, but the lists used in the instruments are different except for the Academic Word List.

### *Correlation Between Receptive and Productive Tests*

Normality is an assumption of Pearson's product-moment correlation (Pearson's *r*), so prior to calculating a correlation, all measures from the receptive and productive vocabulary tests were analyzed for skewness and kurtosis. Descriptive statistics for the scores are shown in Table 1.

Table 1

*Descriptive Statistics for Logit (Receptive) and LFP (Productive) Scores*

| **Measure** | ***N*** | **Min** | **Max** | **Mean** | **SD** | **Skewness** | **Kurtosis** |
|---|---|---|---|---|---|---|---|
| Logits | 95 | -.81 | 5.21 | 1.86 | 1.30 | 0.680 | .400 |
| K1 Token | 96 | 100 | 662.00 | 284.94 | 118.29 | 1.197 | 2.197 |
| K1 Percent | 96 | 77.69 | 95.98 | 87.12 | 3.60 | -.421 | -.177 |
| K2 Token | 96 | 2.00 | 28.00 | 14.91 | 6.51 | .154 | -.806 |
| K2 Percent | 96 | .73 | 9.49 | 4.78 | 1.88 | .368 | -.486 |
| AWL Token | 96 | 1.00 | 38.00 | 14.28 | 7.68 | .783 | .378 |
| AWL Percent | 96 | .92 | 10.00 | 4.50 | 2.10 | .611 | -.128 |
| OL Token | 96 | 1.00 | 41.00 | 12.02 | 8.06 | 1.261 | 1.1614 |
| OL Percent | 96 | .56 | 10.31 | 3.61 | 1.83 | .716 | .725 |

*Note*: Min = minimum score; Max = maximum score; SD = standard deviation.

The skewness and kurtosis for K1 tokens and OL tokens are particularly high with values greater than 1.0. These high values indicate a deviation from the normal distribution, so log transformations were applied to all productive measures in order to achieve normal distributions for the correlational analysis. The logit scores for the receptive measures include negative values so they could not be log transformed, but since the skewness and kurtosis values were less than 1.0, the values did not need to be transformed.

Table 2 shows the skewness and kurtosis for the productive measures before and after transformations. The LFP measures that became more normal with a log transformation are K1 tokens, AWL percent, and OL tokens, so these measures were used in place of the raw scores in order to achieve the assumption of normality for the Pearson's product-moment correlation.

Table 2

*Descriptive Statistics for LFP (Productive) Scores Before and After Log Transformation*

| Measure | Before Log Transformation | | After Log Transformation | |
|---|---|---|---|---|
| | Skewness | Kurtosis | Skewness | Kurtosis |
| K1 Token | 1.197 | 2.197 | **-.201** | **.206** |
| K1 Percent | **-.421** | **-.177** | -.523 | -.056 |
| K2 Token | **0.154** | **-.806** | -.994 | 1.082 |
| K2 Percent | **.368** | **-.486** | -.895 | 1.946 |
| AWL Token | **0.783** | **.378** | -1.062 | 2.141 |
| AWL Percent | .611 | -.128 | **-.578** | **.216** |
| OL Token | 1.261 | 1.1614 | **-.758** | **.665** |
| OL Percent | **.716** | **.725** | -.861 | .721 |

*Note*: The skewness and kurtosis values for each measure that are closest to the normal distribution are in **bold**.

Correlations between LFP measures and logit scores for the receptive test, shown in Table 3, were quite low. The highest correlation, that between receptive and the AWL list, is still well below the 0.70 required to be considered a medium correlation.

Table 3

*Correlations Between Logit (Receptive) and LFP (Productive) Measures.*

| **LFP Measure** | **Correlation with Logit Score (Pearson's *r*)** | **Significance Value (*p*)** |
|---|---|---|
| K1 Tokens* | .006 | .47 |
| K1 Percent | -.136 | .09 |
| K2 Tokens | .100 | .17 |
| K2 Percent | .125 | .11 |
| AWL Tokens | .147 | .08 |
| AWL Percent* | .234 | .01 |
| OL Tokens* | -.109 | .15 |
| OL Percent | -.114 | .14 |

*Note*: * = log transformed scores were used in correlations.

*The p*-values were adjusted using a Bonferroni correction, with the alpha level set at 0.05 for eight comparisons (0.05/8=0.006). The correlation that is the closest to that value is the one between logit scores and the percentage of words from the AWL in a composition, which is .234 and has $p = 0.01$. However, even if the correlation were significant, it would still only explain 5.48% of the variance between the two sets of scores. Therefore, we cannot conclude that these receptive and productive vocabulary measures are correlated due to anything but chance.

### *Principal Components Analysis for Productive Tests*

Principal components analysis was used to investigate the low correlations between receptive and productive test scores described above. First, in order to decide whether to use percent or token scores from the LFP data for factor analysis, a principal components analysis was run on the percent and token scores from the K1, K2, Academic, and Off-List categories. Similar to the correlations above, log transformations were used for K1 token, Academic percent, and Off-List token scores. In the principal components analysis, four eigenvalues over 1.0 were found, and those components accounted for 97.675% of the variance, as shown in Table 4.

Table 4

*Components Analysis for Productive LFP Data with Varimax Rotation*

| **Measure** | **Component** | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | $h^2$ |
| K1 Percent | **-.660** | .190 | **-.614** | -.372 | .987 |
| K1 Tokens* | -.013 | **.964** | -.161 | .141 | .975 |
| K2 Percent | .153 | -.150 | **.965** | -.118 | .991 |
| K2 Tokens | .134 | **.678** | **.714** | .005 | .987 |
| AWL Percent* | **.964** | -.078 | .163 | -.039 | .963 |
| AWL Tokens | **.800** | **.561** | .055 | -.011 | .957 |
| OL Percent | .015 | -.077 | .009 | **.989** | .963 |
| OL Tokens* | .017 | **.506** | -.092 | **.839** | .970 |

Note: * = log transformation used for scores, AWL = Academic Word List, OL = Off-List. Loadings higher than 0.5 are presented in **bold** to clearly display component loadings and complexity.

All the token scores load on component 2 higher than .50, which possibly indicates a method effect, perhaps related to the number of words written. Only the off-list tokens and percent scores load highly on component 4, while academic word list tokens and percent scores load on component 1 and K2 tokens and percent scores load on component 3. K1 percent scores load on both components 1 and 3, indicating complexity. The results of this principal components analysis indicate that, when analyzed together, the token LFP scores load on the same component and therefore have a shared source of variance. Therefore, the percent scores should be used in a factor analysis of receptive and productive data together because they have more varied sources of variance.

### *Principal Components Analysis for Receptive Tests*

**Form A.** Factor analysis at the item level requires separate analyses for Form A and Form B of the receptive test, which have no items in common. Table 5 shows descriptive statistics for LFP percent scores, alongside log transformations of those scores, for participants who took Form A of the receptive test.

Table 5

*Descriptive Statistics for LFP Percent Scores for Receptive Form A Participants*

| Score | *N* | Min | Max | Mean | SD | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| K1 | 56 | 77.69 | 95.58 | 87.48 | 3.73 | -.710 | .476 |
| K1 Log | 56 | 1.89 | 1.98 | 1.94 | .02 | -.832 | .655 |
| K2 | 56 | 1.19 | 9.49 | 4.59 | 1.70 | .550 | .230 |
| K2 Log | 56 | .08 | .98 | .63 | .17 | -.547 | .663 |
| AWL | 56 | 1.10 | 10.00 | 4.25 | 2.12 | 1.001 | .542 |
| AWL Log | 56 | .04 | 1.00 | .58 | .220 | -.247 | .084 |
| OL | 56 | .56 | 10.31 | 3.68 | 2.03 | .743 | .643 |
| OL Log | 56 | -.25 | 1.01 | .49 | .28 | -.718 | .103 |

*Note:* Min = minimum score; Max = maximum score; SD = standard deviation; AWL = academic word list; OL = off-list.

In order to satisfy the requirement for normal distributions, factor analysis for Form A utilized raw percent scores for K1 and K2 lists and log transformations of percent scores for Academic and Off-List word lists.

Principal components analysis was run for all 26 items on Form A and the LFP percent data (with log transformations of Academic and Off-List word lists). A total of 11 components had eigenvalues greater than 1.0, accounting for 74.381% of the variance. Three components were selected for further analysis based on the scree plot shown in Figure 6, which levels off after three components. The three components used for this principal components analysis account for 31.257% of the variance in the data.

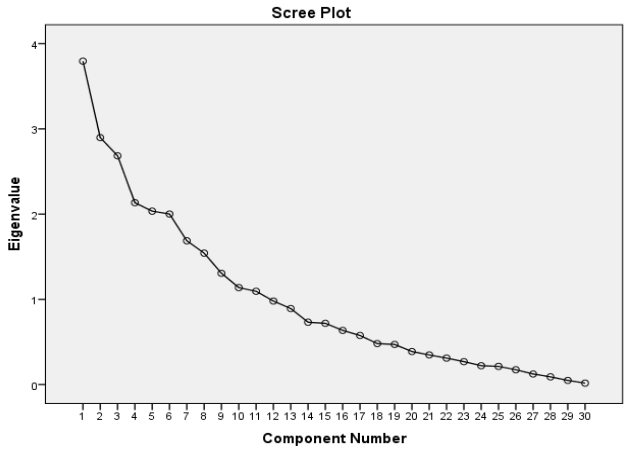



*Figure 6.* Scree Plot for Form A.

Table 6 shows the component matrix for Form A and LFP percent data with a Varimax rotation. The left-hand column in Table 6 shows the item number and the word being assessed for the items on Form A of the receptive test. Items 1-8 are from the 5,000 word level, items 9-17 are from the Academic Word List, and items 18-26 are from the 10,000 word level. Component loadings greater than 0.3 are shown in bold.

Table 6

*Components for Form A Receptive Test Data & LFP Percent Scores with Varimax Rotation*

| **Measure** | **Component** | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | $h^2$ |
| 1 plot | **.382** | .056 | -.225 | .199 |
| 2 sample | -.185 | **.439** | .231 | .281 |
| 3 fiber | -.048 | **.356** | .096 | .138 |
| 4 harsh | **.321** | .126 | .081 | .125 |
| 5 solitary | .086 | **.405** | -.118 | .185 |
| 6 prescribe | **.663** | .160 | .091 | .474 |
| 7 relax | **.686** | .132 | -.054 | .491 |
| 8 resent | .299 | .049 | -.236 | .148 |
| 9 modify | **.364** | **.331** | -.034 | .243 |
| 10 exhibit | **.533** | -.195 | .090 | .330 |
| 11 reinforce | .070 | **.309** | .046 | .102 |
| 12 granted | .273 | .114 | **.506** | .343 |
| 13 complex | **.374** | **-.449** | .153 | .365 |
| 14 precise | **.563** | -.041 | .280 | .397 |
| 15 label | **.382** | -.185 | -.012 | .181 |
| 16 element | **.409** | **-.410** | -.106 | .346 |
| 17 ministry | .069 | .252 | -.015 | .069 |
| 18 triple | -.151 | .166 | .002 | .051 |
| 19 specific | **.313** | **.380** | .048 | .245 |
| 20 bizarre | **.547** | .113 | .191 | .348 |
| 21 poison | .160 | .147 | **.429** | .231 |
| 22 psychiatrist | .032 | .325 | **.579** | .442 |
| 23 flu | .115 | **.509** | **.445** | .471 |
| 24 contemplate | .113 | **.594** | -.188 | .401 |
| 25 contaminate | **.490** | **.595** | -.054 | .597 |
| 26 dissipate | -.087 | **.640** | .109 | .429 |

| | | | | |
|---|---|---|---|---|
| K1 Percent | .079 | .093 | **-.801** | .656 |
| K2 Percent | .066 | **-.333** | **.599** | .474 |
| AWL Percent Log | .052 | -.113 | **.331** | .125 |
| OL Percent Log | -.177 | .163 | **.658** | .490 |

*Note:* Highly loading variables are presented in **bold** to clearly display component loadings and complexity.

This rotated component matrix shows that communalities are quite low for most of the variables in this factor analysis, indicating that most of the variance for these variables is not explained by the factors shown. In spite of the low communalities, it seems to be the case that the first two components explain more of the variance in the receptive test items, while the third component explains more of the variance in the productive LFP data. Four items from the receptive test (granted, poison, psychiatrist, and flu) loaded exclusively on component 3; one (flu) was complex and also loaded on component 2. The rest of the receptive test items loaded either component 1 or component 2, both components, or neither component. No patterns emerged about which receptive test items loaded on components 1 or 2 based on linguistic origin of the item, whether the word is a loan word from English in Chinese, Japanese, or Korean, the word level for the item, or how the item difficulty according to Rasch analysis.

***Form B.*** Table 7 shows descriptive statistics for LFP percent scores, alongside log transformations of those scores, for participants who took Form B of the receptive test. In order to satisfy the requirement for normal distributions, factor analysis for Form B utilized raw percent scores for K1, K2, and academic lists and log transformations of percent scores for the Off-Word list.

Table 7

*Descriptive Statistics for LFP Percent Scores for Receptive Form B Participants*

| **Score** | ***N*** | **Min** | **Max** | **Mean** | **SD** | **Skewness** | **Kurtosis** |
|---|---|---|---|---|---|---|---|
| K1 | 53 | 79.20 | 93.14 | 86.75 | 3.67 | -.186 | -.874 |
| K1 Log | 53 | 1.90 | 1.97 | 1.94 | .018 | -.256 | -.815 |
| K2 | 53 | .73 | 9.35 | 5.16 | 1.92 | .066 | -.714 |
| K2 Log | 53 | -.14 | .97 | .68 | .20 | -1.439 | 4.209 |
| AWL | 53 | .92 | 9.72 | 4.70 | 2.08 | .360 | -.386 |
| AWL Log | 53 | -.04 | .99 | .62 | .22 | -.779 | .526 |
| OL | 53 | .56 | 10.31 | 3.38 | 1.85 | 1.188 | 2.526 |
| OL Log | 53 | -.25 | 1.01 | .46 | .27 | -.748 | .787 |

*Note*: Min = minimum score; Max = maximum score; SD = standard deviation; AWL = academic word list; OL = off-list.

Principal components analysis was run for all 27 items on Form B and the LFP percent data (with log transformations the Off-List word list). A total of 11 components had eigenvalues greater than 1.0, accounting for 77.098% of the variance. Five components were selected for further analysis based on the scree plot shown in Figure 7, which drops off sharply after one component levels off after five components. The five components used for this principal components analysis account for 50.877% of the variance in the data.

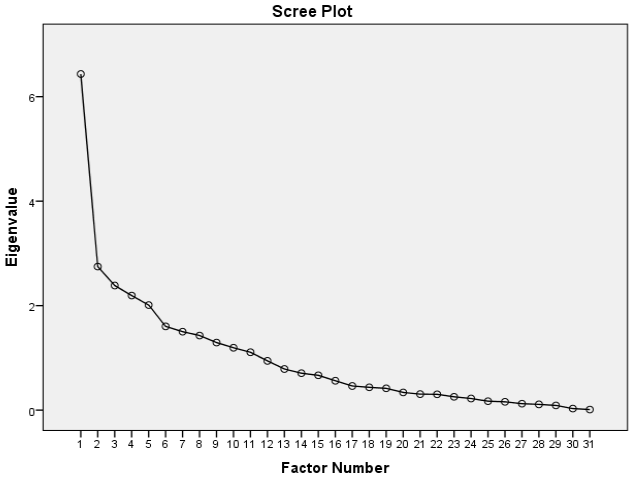



*Figure 7*. Scree Plot for Form B.

Table 8 shows the component matrix for Form B and LFP percent data with a Varimax rotation. The left-hand column in Table 8 shows the item number and the word being assessed for the items on Form B of the receptive test. Items 1-9 are from the 5,000 word level, items 10-18 are from the Academic Word List, and items 19-27 are from the 10,000 word level. Component loadings greater than .40 are shown in bold.

Table 8

*Components for Form B Receptive Test Data & LFP Percent Scores with Varimax Rotation*

| **Measure** | **Component** | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | $h^2$ |
| 1 contemplate | **.607** | -.118 | .380 | .081 | -.002 | .533 |
| 2 entitle | **.551** | .003 | .299 | .129 | .022 | .410 |
| 3 wake | **.417** | .179 | -.066 | .237 | -.127 | .283 |
| 4 hydrogen | -.016 | .227 | .153 | **.537** | .000 | .364 |
| 5 sermon | -.054 | -.021 | **.454** | **.485** | .097 | .454 |
| 6 trumpet | .254 | -.112 | .283 | **.759** | -.004 | .733 |
| 7 municipal | .176 | .032 | **.601** | .321 | -.087 | .504 |
| 8 tragic | .026 | .035 | **.654** | .231 | .098 | .493 |
| 9 profound | **.413** | **.428** | .037 | .277 | .122 | .447 |
| 10 emerge | .242 | .148 | **.456** | .010 | .242 | .347 |
| 11 contrast | -.075 | .302 | **.659** | .005 | .080 | .538 |
| 12 ensure | .366 | .272 | .302 | -.285 | -.243 | .440 |
| 13 incentive | **.752** | .255 | .018 | -.099 | -.029 | .642 |
| 14 norm | **.710** | .035 | .155 | .094 | .290 | .622 |
| 15 implication | **.565** | .353 | -.069 | .071 | .185 | .489 |
| 16 rigid | .112 | **.510** | .281 | .326 | .113 | .471 |
| 17 neutral | -.034 | **.691** | .181 | .215 | -.085 | .565 |
| 18 marginal | -.032 | **.568** | .010 | -.113 | .194 | .374 |
| 19 respectable | .139 | -.055 | **.479** | .051 | .152 | .277 |
| 20 voluntary | -.236 | **.409** | **.466** | .339 | .199 | .594 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 21 incredible | .080 | -.028 | .060 | **.501** | .056 | .265 |
| 22 entitle | **.452** | .377 | .230 | .347 | -.244 | .580 |
| 23 glance | .176 | **.510** | **.410** | .193 | -.094 | .505 |
| 24 prefer | .202 | **.591** | .061 | -.226 | -.125 | .460 |
| 25 fringe | .194 | .332 | -.308 | **.634** | .318 | .746 |
| 26 canopy | .153 | **.692** | -.271 | .387 | .090 | .733 |
| 27 botany | -.025 | .163 | .251 | **.563** | -.103 | .418 |
| K1 Percent | .096 | -.100 | -.215 | -.058 | **-.882** | .846 |
| K2 Percent | .043 | .127 | .295 | -.166 | **.742** | .683 |
| AWL Percent | .253 | -.104 | -.020 | .238 | **.747** | .690 |
| OL Percent Log | **-.500** | .106 | .070 | .020 | -.010 | .267 |

*Note:* Highly loading variables are presented in **bold** to clearly display component loadings and complexity.

Compared to the communalities for the principal components analysis for Form A, the analysis of Form B indicates that a higher amount of the variance is accounted for by the selected components. The LFP percent scores loaded highly on component 5 except for the Off-List score, which loaded on component 1. No receptive test items loaded highly on component 5. Most of the receptive items loaded on components 1, 2, 3, and 4, but some did not load on any and some were complex on multiple components. No patterns emerged about which receptive test items loaded on components 1, 2, 3, or 4 based on linguistic origin of the item, whether the word is a loan word from English in Chinese, Japanese, or Korean, the word level for the item, or how the item difficulty according to Rasch analysis.

## DISCUSSION

### *1. What is the Relationship Between University ESL Students' Productive and Receptive Vocabulary Levels?*

After the two forms of the receptive test were anchored to each other with Rasch analysis and some measures of the productive LFP data were logarithmically transformed to achieve a normal distribution, correlations were run between the receptive logit scores and LFP measures. These correlations were expected to be high because of the common-sense notion that receptive and

productive vocabulary abilities are related to each other. However, they were so low that not a single correlation was significant at $p < 0.05$, given the eight correlations run. According to the correlations between the scores on the test instruments used in this study, no claim can be made that there is any kind of relationship between receptive vocabulary knowledge and vocabulary production.

### *2. To What Extent Do Different Vocabulary Levels Affect Performance on Both Receptive and Productive Vocabulary Test Forms?*

In order to investigate the very low correlations between productive and receptive vocabulary knowledge found in the first part of this study, principal components analysis was performed. First, principal components analysis for just the productive data indicated that all the token scores tended to load on the same component, thus indicating some kind of practice or measurement effect. This component might be a fluency effect, which could be related to how many words a student produces in a composition overall. In general, the percentages of words from different word levels loaded on different components, indicating that the percent measures explained more of the variance in the productive scores.

LFP measures, some of which were transformed logarithmically in order to achieve the assumption of normality for factor analysis, were included in the principal components analysis alongside the item-level receptive data. Two separate principal components analyses were performed. The principal components analysis for Form A had low communalities for many items, indicating that not much of the variance was explained by the components analyzed. However, Form B had higher communalities, perhaps because more components were used in analysis as determined by the eigenvalues and the scree plot. The principal components analysis for both forms indicated that LFP percent scores for vocabulary production tended to load on a separate components from receptive test items. However, no pattern was found indicating a relationship between the components found and the vocabulary level for the receptive items. In fact, no pattern was found regarding any linguistic feature of the words in the receptive test items.

### *Future Research*

This study has several limitations that should be addressed by future research. The first limitation is that Form A of the receptive test contained an error, and it is unknown how this error effected the subsequent data. Additionally, the two forms of the receptive test were created using classical test theory based on the administration of the test to a single intermediate-level writing class. The tests in the study were administered to students in both intermediate- and advanced-level classes, which perhaps explains why the items in Rasch analysis were too easy for the ability levels of the students. Future studies would benefit from creating equivalent forms using logit scores from Rasch analysis and ensuring the pilot population is equivalent to the population of study participants.

A question that remains unanswered in the current study is why certain receptive test items load on certain components. Future research should explore this further by investigating linguistic features of the vocabulary words tested such as the presence of English loan words in other languages and whether the word has Latinate or Germanic etymology. In the current study, no relationship was found between the word level (5,000, AWL, or 10,000) of an item and its difficult, and future research should investigate whether this result holds with different learner populations and with different vocabulary items. Finally, this study provides insight into the relationship between productive and receptive vocabulary skills through principal components analysis, and structural equation modeling (SEM) should be used in future research to further explore the extent to which test score data fits a proposed model in which receptive and productive vocabulary knowledge are fundamentally different constructs.

## CONCLUSION

The results of this study question the common-sense notion that receptive and productive vocabulary knowledge are highly related to each other. The measures of vocabulary production and receptive vocabulary knowledge were found to have very low correlations, and productive and receptive vocabulary measures tended to load highly on different components in a principal components analysis. These results raise the idea that perhaps these two aspects of overall vocabulary knowledge have less to do with each other than it is generally thought. Although common sense and many ESL reading and writing curricula assume that these two

manifestations of vocabulary knowledge would be related, the low correlations and complex loadings shown here indicate otherwise.

This lack of a clear relationship between two aspects of vocabulary knowledge has implications for both language pedagogy and vocabulary assessment. Pedagogically, writing instruction often presupposes that students are able to produce the words in a composition that they can identify receptively, especially at the academic level. The results of this study call this assumption into question and suggest that second language writing instruction should incorporate the direct teaching of productive vocabulary skills, rather than assuming that students will be able to use words in their writing if they can identify them receptively.

In terms of vocabulary assessment, the results of this study indicate that testing should approach vocabulary from the multiple perspectives of reading and writing, and perhaps even listening and speaking as well. Testers should keep in mind that receptive and productive knowledge of vocabulary are multi-faceted constructs (see Laufer et al., 2004) and one measure of vocabulary knowledge may not be adequate for measuring a student's knowledge at any one word level or in any one medium. It is of fundamental importance for both testers and teachers to recognize the complexity of vocabulary assessment and approach it from multiple perspectives and modalities.

## REFERENCES

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

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 Word Level and University Word Level Vocabulary Tests. *Language Testing, 16*(2), 131-162.

Bond, T., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). New York: Routledge.

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York: Routledge.

Cho, H. (2007). Fluency, lexical diversity, and selected syntactic features in L2 students' writing (master's scholarly paper). Retrieved from http://scholarspace.manoa.hawaii.edu/bitstream/10125/20185/1/Cho2007.pdf

Coxhead, A. (1998). *An academic word list*. ELI Occasional Publications #18, School of Linguistics and Applied Language Studies. Wellington, Australia: Victoria University of Wellington.

Coxhead, A., & Nation, P. (2001). The specialized vocabulary of English for academic purposes. In J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp. 252-267). Cambridge: Cambridge University.

Laufer, B. (2005). Lexical frequency profiles: From Monte Carlo to the real world: A response to Meara (2005). *Applied Linguistics, 26*(4), 582-588.

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing, 21*, 202-226.

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

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. (2005). Lexical frequency profiles: A Monte Carlo analysis. *Applied Linguistics, 26*(1), 32-47.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. New York: Newbury House.

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.

Zheng, Y. (2012). Exploring long-term productive vocabulary development in an EFL context: The role of motivation. *System*, *40*, 104-119.