Toward Better Interpretations of Vocabulary Size Test Scores for Estimating Reading Proficiency

Tatsu6 ISO
Reitaku University
Kazumi AIZAWA
Tokyo Denki University

Abstract

This study addresses the interpretation of the results of vocabulary size tests with reading proficiency estimation in mind. Despite the number of vocabulary size tests developed so far, how the raw scores should be interpreted, or put into a formula, to obtain estimated vocabulary size of L2 learners seems to leave room for improvement. In this study, three different interpretive approaches, namely the ‘level,’ ‘size,’ and ‘confidence’ interpretations were compared in terms of resulting estimated vocabulary sizes and their correlative strength with reading proficiency. According to test results for 162 Japanese university students, the study found that the size-approach yielded the largest estimated vocabulary size of 5442.77 words in average. It was also found that the vocabulary size calculated using a confidence-based approach showed the strongest correlation with L2 reading proficiency, followed by a size-based approach with a slightly lower correlation coefficient. The results indicated that although comparatively new, the confidence-based approach holds potential for estimating vocabulary size as well as reading proficiency, and the traditional ‘size’ interpretation was deemed as the most practical choice for the time being.

1. Introduction

Undoubtedly, knowledge of vocabulary is crucial to successful communication. Reading, of course, is a form of communication, and seems more susceptible to vocabulary knowledge size than other forms of communication due to its asynchronous characteristics. In fact, the strong relationship between the vocabulary knowledge and reading proficiency led one researcher to maintain that the size of vocabulary knowledge is “the single best predictor of text comprehension” (Alderson, 2000, p. 35).

When investigating its relationship with general English proficiency or reading proficiency, it has traditionally been the breadth of vocabulary knowledge that held the most importance (Ishii & Schmitt, 2009). Although some researchers currently regard vocabulary knowledge as having
multiple dimensions, such as size, organization, and/or accessibility (Henriksen, 1999; Meara, 1996), vocabulary size has enjoyed the most attention thus far. One of the biggest reasons for this popular focus is that, on the surface, measuring one’s vocabulary size has seemed more straightforward than other dimensions.

Unfortunately, this does not necessarily mean the estimation of vocabulary size is without problems. One aspect that needs to be addressed is the way vocabulary size test results are interpreted. Interpretation refers to how the scores of vocabulary size tests should be transformed into breadth of vocabulary. Although there are two traditional interpretive methods, it is not clear which interpretation should be employed in practice, for example, in estimating reading proficiency with learners’ vocabulary size. Further, with recent interest in learner confidence while answering each question in a vocabulary size test, it is possible that a new interpretation tool is most suitable in estimating vocabulary size. This study, therefore, aimed to discover the manner in which the results of a vocabulary size test should be interpreted so that the relationship between vocabulary size and reading proficiency would make better sense.

2. Literature Review

It is generally accepted that vocabulary knowledge is necessary for successful reading comprehension. As Alderson (2000) claims, vocabulary size best predicts a learner’s reading proficiency. Accordingly, when investigating the relationship between vocabulary knowledge and reading proficiency, it has traditionally been the size of vocabulary knowledge (Ishii & Schmitt, 2009). (See Aizawa & Iso, 2008; Hazenberg & Hulstijn, 1996, for example.)

One of the most widely-known, and widely used types of vocabulary size test is the multiple-choice format, represented by Vocabulary Levels Test (Nation, 1983, 1990). There are revised versions (Beglar & Hunt, 1999; Schmitt, Schmitt, & Clapham, 2001) and also localized versions of the test in a same or very similar format (Aizawa, 1998; Iso & Aizawa, 2008; Mochizuki, 1998). Although these tests of vocabulary size differ in several respects, such as the number of questions and whether or not they are computerized, the essential concept is the same. A certain number of words (usually ranging from 18 to 30 per 1000) are chosen from frequency-based wordlists, and test-takers demonstrate their knowledge of words by choosing the words’ closest descriptions or foreign language equivalents.

Although the test format seems stable for the time being, there is still an issue of test score interpretation. In this study, the word interpretation is used to mean how the raw scores of a test are converted into learners’ estimated vocabulary size. Traditionally, there are two major types of interpretations: the ‘level’ and the ‘size’ interpretations.

The ‘level’ interpretation is common in studies conducted overseas. The basic idea of the ‘level’ interpretation is that when a learner can show a sufficient amount of knowledge at a given frequency level within a vocabulary test, he/she is considered to have mastered that level. The
cut-off point is usually set at 80 percent for each frequency band. Further, a learner’s vocabulary level is estimated to be at the most difficult frequency level up to which he/she can maintain the score above the cut-off point. Table 1 shows this imaginary learner’s vocabulary level to be at the 3000-level. Notice that this learner’s vocabulary level is not at 5000-level where he/she scored above the cut-off point of 80%. This is because this learner did not maintain the score above the cut-off point at the 4000-level.

One of the problems with this type of interpretation is that it could be considered rather strict. According to the ‘level’ interpretation, once a learner fails to consecutively score above the cut-off point at any given frequency level, the vocabulary knowledge after such level is not included in the estimated vocabulary size. Due to this issue, the estimated vocabulary size may be smaller than the actual size.

Another problem is that the cut-off point is arbitrary. The authors hold the view that there is no empirical evidence showing that a certain cut-off point is high enough to show knowledge of all words in a given frequency band.

The second of the traditional interpretations is the ‘size’ interpretation. Many of the vocabulary acquisition studies conducted in Japan estimate vocabulary size based on the ‘size’ interpretation. With this interpretation, the knowledge of all the words in a word list is estimated by how many test items are correctly answered. As an example, let us consider the following data where 240 words are chosen from a list of 8000 words. When a learner’s score is 100%, his/her vocabulary size is estimated as 8000 words. If the score is 80%, the estimated vocabulary size would be 6400 words. While the ‘level’ interpretation views vocabulary size as how many levels a learner has mastered, the ‘size’ interpretation sees it as how many words has been learned.

Unfortunately, the ‘size’ interpretation is also problematic. When a vocabulary size test is in a multiple-choice format, the ‘size’ interpretation may estimate learners’ vocabulary size as larger than it actually is. This problem is inescapable when learners successfully employ a guessing strategy to answer questions, as suggested in Iso and Aizawa (2008). Since the ‘size’ interpretation does not distinguish the degree of knowing a word, i.e., how well a learner knows the word, even the correctly guessed answer contributes to the estimated size of vocabulary and is on par with the correct answers not guessed.

Although the use of guessing strategies have been recognized in vocabulary testing, only a few researchers have attempted to eliminate and correct test scores. Yes/No type vocabulary tests by Meara and his colleagues (Meara, 1992; Meara & Buxton, 1987; Meara & Jones, 1990) are the examples of this.

Quite recently, some researchers began incorporating learner confidence in vocabulary size
tests (Iso & Aizawa, 2008, 2009; Kamimoto, 2006). The idea is to fine-tune the vocabulary size estimation using learner confidence so that confident answers and guessed answers hold different weights in the estimation of vocabulary size.

By developing a vocabulary size test which unobtrusively collects learner confidence data for each test taken, Iso and Aizawa (2009) showed that the confidence-based corrected scores for each frequency band of JACET 8000 was significantly lower than the raw scores. Their result indicated elimination of guessing to some extent. However, it remains unknown whether the correction of the raw scores by the confidence value reflects the amount of vocabulary knowledge, as this issue has never been addressed.

As reviewed, three different types of vocabulary test score interpretations were identified; it is not clear which of them yields the best estimation of vocabulary size. If we are to support Alderson’s (2000) claim, investigation in this direction needs to be carried out. This study, therefore, investigated how a vocabulary size test score should be interpreted so that the estimation of vocabulary size as well as that of the reading proficiency by vocabulary size could be considered more accurate.

3. The Study

3.1 Purpose

The purpose of this study was to investigate a better approach to interpreting vocabulary size test scores. The current study sought answers to the following questions:

1. How different would the estimated vocabulary size be when the interpretation emphasizes level, size, or confidence?
2. Which interpretation of the estimated vocabulary sizes shows the highest correlation coefficient with reading proficiency?

For the first research question, three types of interpretive methods were compared to demonstrate their general characteristics. Then, the study moved on to the second question where the results of the each interpretive type were compared again in terms of its correlation with reading proficiency.

3.2 Subjects

One hundred sixty-two Japanese university students participated in this study. The subjects consisted of 109 English majors and 53 science majors. The subjects’ TOEIC scores varied from 185 to 750, with the average and standard deviation of 404.94 and 129.71, respectively.

3.3 Materials

To gather data on the subjects’ vocabulary size, two forms of Flash VLT Online (Iso & Aizawa, 2009) were used. This receptive vocabulary size test was developed (1) to suit the needs
of Japanese learners of English and (2) to unobtrusively collect learner confidence data when answering each question. The format of the test closely follows that of Schmitt, et al. (2001). Three target words, along with six possible answers, are displayed on a computer screen as one question unit. Within a unit, three target words are drawn from the same frequency band of a vocabulary list, which in the case of this study was JACET List of 8000 Basic Words (JACET 8000). From each frequency band, 30 words were chosen as test items, forming 10 units. As JACET 8000 consists of eight frequency bands, one form of Flash VLT Online includes 80 units with 240 words as test items. Since two forms of the test were used in this study, the total number of words was 480.

In this test, the subjects were asked to match an English word and its Japanese equivalent. As can be seen in Figure 1, when matching an English word and its Japanese meaning, the subjects simply drag the ball in front of a question item to one of the holes below an equivalent Japanese word. When subjects were 100% sure of their answers, they dropped the ball in the hole with a double circle, while the hole with a single circle was used when the subjects thought they knew the answer but were not 100%. If subjects used a guessing strategy, they were told to use the hole with a triangle.

In order to prevent the subjects from using dictionaries, each unit was given a time limit of 30 seconds. Also, the ordering of units was randomized each time the test started. This is to counterbalance the possible negative effects of fatigue towards the end of the test.

3.4 Analysis

Upon applying the three types of interpretive methods to the current data, it was evident that a few changes had to be made, especially with the ‘level’ interpretation. First of all, knowledge of vocabulary in the frequency band on which the subjects did not score above 80% seemed too severe. Further, as mentioned earlier, there is no theoretical underpinning to support the traditional cut-off point at 80%. Therefore, this study prepared four versions of the ‘level’ interpretation. They included:

(1) Learners’ vocabulary level is at the last level up to where they can consecutively score over 80%. This is referred to as 80-C.

(2) Learners’ vocabulary level is at the last level where they can score over 80%. This is “80-H.”
(3) Learners’ vocabulary level is calculated by granting 1,000 words for each frequency level for scoring above 80% and adding them together. This is simply called 80.

(4) Learners’ vocabulary level is calculated by granting 1,000 words for each frequency level for scoring above 80% and 500 for frequency levels where they did not reach 80% but 50%. Due to the combination of 80% and 50% cut-off point, this is called 80/50.

The 80% cut-off point was maintained in this study, despite the lack of theoretical support. This was done so that the current study functions as an anchor to the previous studies and provides a rough estimate of how these previous results might be viewed utilizing interpretive methods employed in this study.

As for the ‘size’ interpretation, since one of the ‘confidence’ interpretations is very similar to the ‘size’ interpretation, the current authors did not recognize the need to make changes to the ‘size’ interpretation. Hence, each subject’s estimated vocabulary size was calculated by dividing the correct number of answers with the total number of question items (480 in this case) and then multiplying by the number of words in the wordlist where the test items were drawn (8000 in this study).

Regarding the ‘confidence’ interpretation, there were two methods of vocabulary size estimation employed. As previously mentioned, one such method, is to make a slight change to the ‘size’ interpretation by counting the highly confident and correct answers as correct. By doing so, it was expected that the guessing effect, which the ‘size’ interpretation does not address, could be eliminated.

Another confidence-based interpretation involved the use of Clustered Objective Probability Scoring (COPS) by Shizuka (2004). This is a method to weigh each answer based on confidence. To simply describe the mechanisms of COPS, each test-taker’s answers are grouped based on the confidence value (high, mid, low, for example) and the weight of the answers in a given confidence group is decided based on the ratio of correct and incorrect answers in the group. Let us consider a hypothetical case where a learner took a 10-item test and answered four questions with high confidence, and three of them were correct. In this case, the weight assigned to each of the four highly confident answers is .75 (3 divided by 4). Yet, since only three of the four answers were correct, the score for the highly confident answers is 2.25 (.75 times 3). Without COPS, the score would be 3.00 as the number of the correct answers is three.

Altogether, this study compared seven interpretive methods to calculate each subject’s estimated vocabulary size. The traditional ‘level’ interpretation had four versions, while the ‘size’ interpretation remained the same. This study also added two ‘confidence’ interpretations to the vocabulary size test.

All figures of estimated vocabulary size were compared in terms of the correlation coefficients with reading proficiency. Due to practical reasons, the subjects’ reading proficiency was defined as the score of the reading section of TOEIC test.
Table 2 Descriptive statistics of the TOEIC reading section and seven different estimations of vocabulary size

<table>
<thead>
<tr>
<th>TOEIC</th>
<th>Mean</th>
<th>SD</th>
<th>Max</th>
<th>Min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>167.72</td>
<td>64.41</td>
<td>375.00</td>
<td>35.00</td>
</tr>
<tr>
<td>80-C</td>
<td>2376.54</td>
<td>1285.44</td>
<td>7000.00</td>
<td>0.00</td>
</tr>
<tr>
<td>80-H</td>
<td>2598.77</td>
<td>1542.20</td>
<td>8000.00</td>
<td>0.00</td>
</tr>
<tr>
<td>80</td>
<td>2481.48</td>
<td>1370.60</td>
<td>7000.00</td>
<td>0.00</td>
</tr>
<tr>
<td>80 / 50</td>
<td>4552.47</td>
<td>1479.26</td>
<td>7500.00</td>
<td>500.00</td>
</tr>
<tr>
<td>Size</td>
<td>5442.77</td>
<td>911.13</td>
<td>7350.00</td>
<td>1967.00</td>
</tr>
<tr>
<td>Confidence</td>
<td>Confident</td>
<td>4070.27</td>
<td>1149.31</td>
<td>6833.33</td>
</tr>
<tr>
<td></td>
<td>COPS</td>
<td>4539.21</td>
<td>1077.40</td>
<td>7073.40</td>
</tr>
</tbody>
</table>

3.5 Results

Prior to the data analysis for the research questions, the reliability of Flash VLT Online was calculated. Cronbach’s alpha for the raw scores was .977, which indicated the test was highly reliable. Please note that although two versions of the test were administered, the reliability coefficient was calculated as if they were one test. This is because the only difference between the versions was the test items.

The descriptive statistics of the TOEIC reading section and each of the seven interpretive conversions are shown in Table 2. Even though the figures for the seven conversions were based on the same dataset, the values were surprisingly different. As is apparent from Figure 2, the vocabulary size estimated with ‘size’ interpretation showed the largest figure of all the seven interpretations. Furthermore, as expected, the ‘level’ interpretation produced the lowest estimation with an exception of 80/50 interpretation where there were two cut-off points. The results of the ‘confident’ interpretations fell somewhere in the middle. ANOVA revealed that the effect of the interpretations was significant (F (6, 966) = 835.55, p < .01). Further, LSD multiple comparison showed that the ‘size’ interpretation yielded the highest mean of estimated vocabulary size with a significant difference from 80/50 and COPS (MSe = 292458.72, p < .05). The confident conversion of the ‘confidence’ interpretation was significantly lower than the 80/50 and COPS, but significantly higher than 80-C, 80-H, and 80 conversion of the ‘level’ interpretation. Therefore, regarding the first research question, it was clear that the estimated vocabulary size significantly varied depending on the interpretations of the vocabulary size test scores. The ‘size’ interpretation was the most generous, and the ‘level’ interpretation in general was the most severe. The ‘confidence’ interpretation fell between the two.

To answer the second research
Table 3 Correlation coefficients between estimated vocabulary sizes and TOEIC reading score

<table>
<thead>
<tr>
<th>Reading</th>
<th>80-C</th>
<th>80-H</th>
<th>80</th>
<th>80/50</th>
<th>Confident</th>
<th>COPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>.613</td>
<td>.562</td>
<td>.607</td>
<td>.570</td>
<td>.617</td>
<td>.562</td>
</tr>
<tr>
<td>Confidence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.622</td>
</tr>
</tbody>
</table>

(p < .01)

question in which the seven interpretive methods were compared in terms of their correlation with reading performance, correlation coefficients were calculated. Table 3 shows the results. As is apparent from the table, the interpretation using COPS showed the strongest correlation with reading proficiency, followed by the ‘size’ interpretation and 80-C of the ‘level’ interpretation with slightly lower correlation coefficients of .617 and .613, respectively. Hence, to answer research question 2, the ‘confidence’ interpretation, especially COPS, revealed the strongest correlation with reading performance in this study.

4. Discussion

This study compared the traditional and alternative interpretive methods of vocabulary size test scores. When the seven estimated vocabulary sizes for each subject were directly compared, the traditional ‘size’ interpretation was found to be the most generous. On the other hand, the smallest estimation came from one of the ‘level’ interpretations, 80-C to be specific, which was the original ‘level’ interpretation used in other studies. The difference was more than 3,000 words. We should be well aware of the fact that the vocabulary size estimation can vary considerably depending on the interpretations employed in studies and educational settings.

When the estimated vocabulary sizes were compared from another perspective, the results looked different. Upon investigating which of the interpretive methods show the strongest correlation with reading performance, the ‘confidence’ interpretation using COPS was the most promising of all the seven interpretive methods.

Although the COPS conversion seemed useful, the traditional ‘size’ interpretation was not so far behind in terms of the correlation coefficients with reading proficiency. In fact, there are a few reasons to choose the ‘size’ over COPS. One is the complexity of the calculation. As explained above, COPS involves counting the number of correct and incorrect answers for all the confidence-index, such as high, medium, and low confidence. With a vocabulary size test where 240 to 480 questions are asked, the calculation is very cumbersome.

Another drawback is that if a subject, irrespective of confidence, does not answer any of the questions incorrectly, COPS’ weighing system does not function as the weight is determined by the proportion of correct and incorrect answers. Until such issues in COPS are resolved, it might be more realistic to employ the ‘size’ conversion.

Due to the issues inherent in COPS, it is important to consider the possibility of another
‘confidence’ based interpretation. In this study, in order to eliminate guessing effect from the data, only the correct answers with high confidence were counted as correct. As a result, the estimated vocabulary size this method produced was not as strongly correlated as other methods. One of the possible reasons, a limitation of this method, was that this method did not count all the correct answers with medium confidence or by guessing. If there is a better way to weight answers depending on learners’ confidence, correct answers with other degree of confidence would contribute to better vocabulary size estimation.

As for the traditional ‘level’ interpretation, the correlation with reading proficiency tended to be weaker than ‘size’ and COPS. It can be argued based on this tendency that the ‘size’ interpretation and COPS were slightly better methods of vocabulary size estimation. Placing a strong focus on judging which frequency levels have been mastered by setting cut-off points, the ‘level’ interpretation may not be as sensitive to the amount of vocabulary knowledge as other interpretations. However, one may argue that the difference in correlation coefficients of the ‘size,’ COPS, and the ‘level’ interpretation was minor. What can be inferred from the study is that even though the ‘level’ interpretation may not provide the whole picture of a learner’s vocabulary, it still showed that the many of the subjects in this study have already mastered several thousand high-frequency words that could be large enough to understand the reading passage used in the TOEIC test.

In summary, a general tendency was observed in this study, that the strength of the correlation with vocabulary size is modest. With the interpretive methods still under investigation, it is too early to make a judgment as to how strongly vocabulary size affects successful reading comprehension. Despite attempts to introduce several new methods of vocabulary size test interpretation, it was the traditional ‘size’ interpretation that showed one of the highest correlation coefficients with reading proficiency and which held practical usefulness. However, this does not necessarily dismiss other methods of interpretations, such as the ‘level’ and the ‘confidence’ interpretation. The ‘level’ interpretation, especially the most traditional 80-C method, is very useful in educational settings where teachers may want to know the level of vocabulary their students have mastered. This kind of information can only be obtained through the ‘level’ interpretation. As for the ‘confidence’ interpretation, it is still in its infancy, as discussed above. With more research in this direction, especially in the weighing of the answers given, it is hoped that the estimation of learners’ vocabulary size and the subsequent estimation of reading proficiency become more accurate and meaningful in the near future.

Acknowledgement

We would like to express our gratitude to Dr. Meara for providing us with invaluable comments on an earlier draft for this project.
References


