Language Education & Technology, 45 (2008)
pp. 35–52

A Comparison of Aural and Written Vocabulary Size of Japanese EFL University Learners

MIZUMOTO Atsushi
University of Marketing and Distribution Sciences

SHIMAMOTO Taiko
Kansai Gaidai University

This study attempts to compare aural and written vocabulary knowledge (size) of Japanese university EFL (English as a Foreign Language) learners and investigate their relationship to listening and reading abilities, and overall English proficiency. To this end, two types of vocabulary size tests were developed using a word list tailored to Japanese EFL learners, and they were administered to a total of 332 university students. The results indicated that the aural and written format of the tests produced a differing result, although the overall tendency was the same. Item analysis was then carried out to further investigate the differences. Analysis of correlation coefficients of proficiency measures and the two formats of the vocabulary size tests revealed that both the aural and written vocabulary size of the participants correlated strongly with the proficiency measures. Based on the findings, practical implications are discussed for the effective use of vocabulary size tests in educational settings.

Keywords: vocabulary size test, aural and written vocabulary size, vocabulary knowledge, listening and reading abilities

1. Introduction

It has been widely recognized that lexical competence is a crucial factor in almost all aspects of L2 proficiency, and it is certainly considered to be "at the heart of communicative competence" (Meara, 1996). Previous research, for instance, has demonstrated the positive and significant relationship between vocabulary knowledge and reading comprehension (e.g., Coady, Magota, Hubbard, Graney, & Mokhtari, 1993; Hu & Nation, 2000; Laufer, 1992; Qian, 1999), whereas fewer empirical studies have yet examined how vocabulary is related to other skills such as speaking or listening. Assessing lexical knowledge of learners is essential for teachers and researchers in the field of second language acquisition. However, vocabulary tests devised for various purposes tend to vary depending on how the test designers define lexical knowledge (Laufer & Goldstein, 2004). Consequently, lexical knowledge has been defined in various ways by different researchers. Among them, vocabulary knowledge frameworks (e.g., Nation, 1990, 2001; Richards, 1976) have often been quoted to subcategorize the lexical knowledge. Nation (2001), for example, lists three primary components of knowledge including Form, Meaning and Use under which nine different aspects of knowledge, such as associations, grammatical functions, collocations and so on, are subcategorized. Each aspect is further broken down into receptive and productive knowledge, giving a total of 18 subcomponents of knowledge. We will go back to Nation's framework in detail later in this section.

A broad distinction between *breadth* and *depth* of vocabulary knowledge has frequently been made in recent L2 vocabulary research (e.g., Anderson & Freebody, 1981; Read, 1993, 1996; Wesche & Paribakht, 1996). Vocabulary breadth or size is generally interpreted as the number of words for which a learner has at least some minimum knowledge of meaning; whereas vocabulary depth refers to how well the learner knows the word, or a quality of knowledge of various aspects of a given word. There have been a growing number of studies on depth of vocabulary knowledge in recent years. (e.g., Read, 1993, 1996; Schmitt, 1998; Schmitt & Meara, 1997; Shimamoto, 2000; Wesche & Paribakht, 1996). There are also other researchers who perceive the lexical competence in different theoretical frameworks or models. Henriksen (1999) proposes three separate but related dimensions of lexical competence: (a) partial to precise knowledge, (b) depth of knowledge, and (c) receptive to productive use ability. She argues that vocabulary learning processes need to be described in terms of both item learning and system learning. According to her definition, vocabulary size measures are primarily concerned with item learning in L2 lexical development. The distinction between item learning and system learning, she explains, goes back to Aitchison's (1994, p. 307) three tasks involved in the process of acquiring word meaning: *labeling, packaging* and *network building*. Item learning is referred to as the first two terms, namely, mapping meaning onto form and discovering the range of meaning for a given word. She argues that L2 vocabulary studies tend to focus mainly on the item learning and disregard network building, the "semantization process." Nation (2001) also distinguishes between item knowledge and system knowledge by explaining that Aitchison's three stages correspond to the three divisions in the Meaning section of his word knowledge frameworks: form and meaning, concept, and associations.

Alternatively, Meara and Wolter (2004) suggest a more global view of L2 lexical competence by proposing *size* and *organization* as an alternative to breadth and depth. *Organization* can be interpreted as "the organizational links between the words that make up the L2 lexicon" (Meara & Wolter, 2004, p. 89). They claim that the breath/depth dichotomy is unfortunate on the ground that measures for depth of vocabulary knowledge tend to focus in too much detail on knowledge of individual words and disregard the whole picture. Depth measures for a large number of words would seem impractical as a test battery. Meara and Wolter (2004) state that "a better approach to vocabulary development would be to look at features which are characteristic of a learner's whole lexicon, rather than features which are characteristics only of single words" and that "*vocabulary size* is not a feature of individual words: rather it is a characteristic of the test taker's entire vocabulary" (p. 87). It would seem that they attempt to perceive learners' vocabulary sizes as a characteristic of their mental lexicon, not just as a sum of words that they know. Meara (1996) also suggests that "Vocabulary size is probably the only dimension of any real importance as long as we are dealing with a small lexicon" (p. 45), implying that there might be a critical size threshold where organization becomes increasingly important. Read (2000) also claims that "despite the fact the size test may seem superficial, they can give a more representative picture of the overall state of the learner's vocabulary than an in-depth probe of a limited number of words" (p. 115).

We have so far attempted to briefly summarize some current views on lexical knowledge. In sum, while comprehensive approaches are essential for assessing lexical knowledge, vocabulary size seems to be a significant concept to represent the whole picture of learner's vocabulary knowledge at different developmental stages.

1.1 *Word Form and Meaning*

What is a vocabulary size test supposed to measure? Vocabulary breadth or size was originally defined by Anderson and Freebody (1981) as "the number of words for which the person knows at least some of the significant aspects of meaning" (p. 92). In Nation's knowledge frameworks (2001), the concept of vocabulary size can be broadly covered by aspects of *Form (spoken, written, word parts)* and *Meaning (form and meaning)*. Thus, if a person can perceive the form of a given word and retrieve the significant aspects of meaning attached to the word form, it can safely be assumed the word is part of his/her vocabulary. If learners can only recognize either the form or the meaning of a word and cannot connect the two, they would not be able to comprehend the word. In brief, unless the form and semantic representations of a word are tightly connected in a learner's mental lexicon, the meaning cannot be automatically retrieved when seeing a word or hearing it. Therefore, how quickly L2 learners can associate the word form with its meaning in L2 reading or listening seems crucial. There appears to be a gap in knowledge and/or automaticity between the form and the meaning in L2 learners. It can be argued that size tests are concerned primarily with this linkage between word form and its meaning. Form and meaning can be considered the most important aspects of vocabulary knowledge and the form-meaning link is central to whatever is being tested (Laufer, Elder, Hill, & Congdon, 2004). As is clearly shown in Nation's framework, knowledge of *Form* should include both the spoken form and the written form of a word. This means that a learner needs to know what the word sounds like as well as what the word looks like. Moreover, the learner needs to know what meaning these word-forms signal in order to comprehend the word (Nation, 2001, p. 27).

1.2 *Vocabulary Size Tests*

Various types of vocabulary size tests have been devised and implemented in the field of ESL/EFL. The most well known one would be the Vocabulary Levels Test (Nation, 1990; Schmitt, Schmitt, & Clapham, 2001). It was originally devised as a diagnostic vocabulary test for classroom use by teachers, but recently it has been widely used as an assessment tool for both L2 pedagogy and research. A new version of the Vocabulary Levels Test (Nation 2001; Schmitt, et al., 2001) is a short multiple choice test which contains 30 items in 10 clusters at 5 different frequency levels. It gives frequency profile information instead of a single figure for overall vocabulary size (Schmitt, et al. 2001, p. 60). Another common type of size measure is a check list (or yes-no) test devised by Meara and his colleagues, in which test-takers are simply asked to indicate whether they know the word or not. (Meara & Jones, 1990; Meara & Milton, 1993). A computer assisted new version, X_Lex (Meara & Milton, 1993) can provide an estimate of learners' vocabulary size in most frequent 5,000 word-bands, which has shown to be linked well with general language exams (Milton & Hopkins, 1996). There are other types of size tests such as supplying an L1 translation for each L2 target word (Nurweni & Read, 1999). Also in Japan, a few size tests suitable for Japanese EFL learners have been devised (e.g., Aizawa, 1998; Mochizuki, 1998; Sato, 2003). Among them, the Mochizuki Vocabulary Size Test (Mochizuki, 1998) has been widely used as a size test for learners in beginner to intermediate levels.

Most size tests including the ones mentioned above, however, are generally presented in a written rather than a spoken/aural format: they attempt to measure the vocabulary knowledge of written forms

rather than spoken forms of the words. As we have seen in Nation's framework, knowing a word refers to knowing what the word sounds like (its spoken form) or looks like (its written form) and its meaning (Nation, 2001). Furthermore, learners need to connect the form (written or spoken) of a word and its meaning. It can be assumed that there might be a difference in the vocabulary knowledge of written and spoken forms of words, as is often the case with some Japanese EFL learners, where a word can be understood when it is seen, but not when it is heard. Therefore, the link between the spoken form and its meaning also needs to be assessed in size tests. Unfortunately, there seems to be very few aural size tests that have been developed for EFL learners. For instance, a dictation test of vocabulary based on frequency levels was probably the only known aural vocabulary test used to test listening vocabulary (Fountain & Nation, 2000; Nation, 1990, 2001).

A new aural vocabulary size test, called Aural Lex (Milton & Hopkins, 2005) has recently been developed. It is a computer-delivered Yes/No test, which is comparable to X_Lex (written version). Test-takers are presented with 120 words one by one and required to indicate whether they know each word. The 20 target words are randomly selected from each 1,000 word-frequency band, and 20 pseudo words are included in the test in order to be adjusted for guessing and overestimation. Milton and Hopkins (2006) compared phonological and orthographic vocabulary sizes of Greek and Arabic EFL learners using X_Lex and Aural Lex as measuring tools. They claim that the vocabulary knowledge of Arabic-speaking learners tends to be more phonological in form than orthographical since they often experience problems with English orthography due to the different Arabic writing system, while Greek-speaking learners do not usually have this kind of orthographic difficulty. Therefore, it was assumed that vocabulary tests that are in written format might underestimate the real vocabulary knowledge of Arabic learners. Their results indicated that the two groups showed very similar scores on X_Lex (written version), but they differed quite noticeably on Aural Lex. The Arabic group's score was significantly higher than the Greek group's score ($p < .01$). Therefore, the study suggested that the relationship between phonological and orthographic vocabulary knowledge might be related in some way to learners' backgrounds and their first language and that a phonological size test along with an orthographic one can tap this kind of difference in L2 learners' lexical knowledge. We would suspect that the relationship between written and aural vocabulary size may also be related to learners' reading and listening abilities, and also to overall English proficiency. It would be interesting to know how these two types of knowledge, of written and spoken form, compare in Japanese EFL learners with different English proficiency levels or how they relate to their reading or listening skills. We assume that Japanese EFL learners would generally tend to have more difficulty in recognizing an English word in spoken form than in written form due to the pedagogical practice in Japan, where reading tends to be more stressed than oral communication. It is also possibly due to an interference from so many English loan words in Japanese, which sound quite different from English pronunciation. According to a study by Yamauchi (2005), which used a web-based vocabulary test based on the Vocabulary Levels Test (Nation, 1990), the vocabulary size of Japanese EFL learners for reading comprehension was found to be significantly larger than that for listening comprehension. The study also indicated that the correlations between reading vocabulary size and reading ability ($r = .49$) was higher than that of listening vocabulary size and listening ability ($r = .32$). Alternatively Katagiri (2001), using Mochizuki's vocabulary test (Mochizuki, 1998) and targeting senior high school students as participants, found that the listening vocabulary tests seemed to be more difficult than the written vocabulary. His

results also indicated that the written version of the vocabulary test had higher correlations with general English tests (TOEIC; the Test of English for International Communication) than the listening version (written, r = .72; listening, r = .63) did. A small number of previous studies seem to suggest that there might be a difference between aural and written vocabulary size of L2 learners with different proficiency levels. Also, it can be assumed that using different testing measures would be likely to result in different findings.

2. The Present Study

The purposes of this study are as follows: (a) to assess the knowledge of spoken and written forms of English words among Japanese EFL learners by using both aural and written formats of vocabulary size tests; (b) to compare the results of both the aural and written formats of vocabulary size tests; and (c) to examine the relationship between aural and written vocabulary sizes of learners and their language skills.

Specifically, the present study addresses the following research questions:

1. Is there a significant difference between learners' aural and written vocabulary size? If there is a difference, how do they differ? Do they differ in separate frequency bands?
2. Is there a significant difference between aural and written vocabulary size in learners with different proficiency levels? If there is any difference, how can we account for it?
3. How do aural and written vocabulary sizes relate to listening and reading abilities, and overall English proficiency?

2.1 *The Development of Aural and Written Vocabulary Size Tests*

In creating the aural and written versions of the vocabulary size test, a few considerations were made. First of all, the aural and written versions of the test were designed to be exactly the same in the number of testing items, the selection of them, and the test format (multiple choice). The only difference was the mode of delivery: the aural vocabulary size test was elicited by sound stimuli, whereas the written version was given in written format. Secondly, since we were using an audio CD for the aural version of the test, the format of the test needed to be very simple because test takers were not able to go back and listen to multiple choice questions again. Another factor to be considered was that in order to increase the validity and reliability of the test, it was necessary for a vocabulary size test to cover as many sample words (testing words) as possible, including target words and options. In other words, the sampling rate needed to be high. In addition, it was preferable that the test form was in four-choice questions so that we could utilize elaborate data analysis methods such as item analysis established in language testing. With these considerations in mind, the test format was decided as follows.

1. 静かな, 穏やかな

(A) successful (B) quiet (C) strange (D) true

* For the aural version, choices (A) ~ (D) were not written out but only given by sound stimuli recorded on the CD.

It should be noted that the four English option words in the multiple choice question were not written out

in the aural version of the test. In each four-choice question, the definition of a target word was given in Japanese (L1) on the test paper so that examinees could easily understand it. The first and the second (sometimes the third) meaning senses of the target word were given for its definition. The task for the aural vocabulary size test was to listen to the four optional words recorded on an audio CD and then to choose the target word which matches the Japanese definition written on the test. On the other hand, in the written version, participants were able to see all four options as they marked the answers.

JACET8000 (JACET, 2003) was used as a vocabulary list for selecting target words and option words in the test. It was mainly because the participants of this research were all Japanese college students, and that JACET8000 was developed particularly for Japanese EFL learners. JACET8000 was compiled based on the frequency of the British National Corpus with a sub corpus useful for Japanese EFL learners. It consists of eight levels of vocabulary (from the 1,000 to the 8,000 word levels) according to frequency counts. It employs word item counting rather than word family counting (where a word is counted as a word family consisting of a base form together with the derived forms and some familiar inflected forms such as -able , -er, -tion and so on).

First, 20 target words were randomly selected from each of the eight levels of JACET8000, for a total of 160 target words (20 target words × 8 levels). Target words were all content words including nouns, verbs, adjectives and adverbs (in addition, any possible set of these). The percentage of these word classes in the test was based on the one appearing in each level of JACET8000. The option words in each multiple choice question were also randomly sampled from the same level of the target word, and they also fell into the same word class as that of the target word. The percentage of English loan words in this test was 17%. The loan words were not omitted but included in the test because it seemed appropriate based on a finding by Daulton (1998) that of the 2,000 most frequent words of English, up to 38% are loan words in Japanese. Using the multiple-choice test format, we can assess a total of 640 sample words (80 testing words × 8 levels). Table 1 shows the distribution of word classes in the new vocabulary size test (the actual test items are available at www.mizumot.com/let2008vocsizetest.doc).

Table 1

Distribution of Word Classes in the Vocabulary Size Test Used in the Current Study

Level	1000	2000	3000	4000	5000	6000	7000	8000
Sample words	80	80	80	80	80	80	80	80
Nouns	28	40	44	40	44	40	36	36
Nouns/Verbs	24	12	4	4	4	4	8	8
Verbs	12	8	12	12	12	12	8	4
Adjectives	8	12	12	16	16	16	20	20
Adjectives/Nouns	4	4	4	4	0	4	4	4
Adverbs	4	4	4	4	4	4	4	8

Note. In the 80 sample words, 20 target words are included.

The aural vocabulary size test was digitally recorded onto a CD by an American male English teacher with a clear voice. The pauses between each word and those between each multiple choice question were timed evenly (14 seconds for each question) using audio filing software. The eight levels were randomly placed in the test for a counterbalance. The total length of time recorded in the CD was approximately 40

minutes including the directions for the test and an example question. The written format of the test was the same as the aural test mentioned above except that the option words were written out on the test paper (the duration of the test was approximately 40 minutes). The task was to choose an appropriate target word among the four written optional words to match the Japanese definition. In this way, we were able to assess the participants' ability to aurally and visually recognize the words.

2.2 *Participants*

The participants of this study were Japanese university and junior college students in different academic years, mostly majoring in language studies such as English, French, and Spanish. There were also some engineering majors. A total of 332 learners took both the aural and written versions of the vocabulary size test.

2.3 *Administration of the Tests*

After a pilot study to check the practicability of the tests with 25 participants, the aural format of the test was administered to a total of 332 students in nine different classes in early December 2005. All the test papers were collected after the test so that the participants could not review them before the written format of the test was given. About a month later, at the beginning of January the students took the written format of the test. Most of the students took TOEIC (the Test of English for International Communication) early in December, but those who did not were asked to report their previous TOEIC score. TOEIC consists of a listening section (100 items) and a reading section (100 items). The full score for each section is 495, making 990 the total score. According to the Education Testing Service (2006), the test developer, "TOEIC has been used to measure the English proficiency of nonnative English-speaking people." Thus, the three types of TOEIC scores (the total score, the listening score, and the reading score) were used in the analysis to measure their English proficiency, listening ability, and reading ability respectively.

3. Results

3.1 *Descriptive Statistics*

Table 2 and Table 3 show the descriptive statistics for the aural version and the written version of the vocabulary size test (henceforth, aural vocabulary size and written vocabulary size) for participants who took both the aural and written versions of the test (N = 332). Cronbach alpha for the total items in both tests were high; the aural version (α = .93) and written (α = .94). The results show a gradual decrease in the mean scores in the eight levels.

The estimated vocabulary size was calculated based on the following formula by Nation (1990, p. 76): Number of correct answers × 8000 (total number in the word list) / 160 (the number of items in test). The mean vocabulary size of the aural version of the test was 5,511, and the written version yielded 5,877, approximately 3,000 words in word family counting (Laufer, 1992).

Table 2

Descriptive Statistics for the Aural Vocabulary Size Test (N =332)

Level (JACET8000)	1000	2000	3000	4000	5000	6000	7000	8000	Total
Number of Items	20	20	20	20	20	20	20	20	160
Mean	19.12	17.55	15.53	12.86	11.37	11.62	11.27	10.90	110.22
Standard Deviation	1.72	2.26	3.60	3.32	3.14	3.27	2.80	2.85	18.07

Table 3

Descriptive Statistics for the Written Vocabulary Size Test (N =332)

Level (JACET8000)	1000	2000	3000	4000	5000	6000	7000	8000	Total
Number of Items	20	20	20	20	20	20	20	20	160
Mean	19.41	17.99	17.17	13.4	12.65	12.32	12.58	12.05	117.55
Standard Deviation	1.46	2.46	2.64	3.37	3.27	3.26	3.15	3.19	18.33

3.2 *Differences Between the Test Formats and Levels*

Figure 1 shows the mean scores of both the written and the aural vocabulary size tests. On the whole, mean scores for the written vocabulary size test were higher than those for the aural vocabulary size test. In order to examine the interaction between the test format and eight levels of JACET8000, a two-way ANOVA with repeated measures was conducted using *SPSS 14.0*. Prior to the analysis, we checked sphericity using Mauchly's test, and it was significant ($p < .05$), indicating that the sphericity assumption was violated. Thus, we used the Greenhouse-Geisser corrected tests (Field, 2005, p. 430). The results revealed that there was a significant main effect of the test format, $F(1, 331) = 243.03$, $p < .01$, partial $\eta^2 = .42$ and the eight levels of the test, $F(5.88, 1944.51) = 1354.95$, $p < .01$, partial $\eta^2 = .80$. Also, there existed a significant interaction effect between them, $F(6.22, 2059.94) = 15.44$, $p < .01$, partial $\eta^2 = .05$. Among these results, the main effect of the test format shows the two types of the tests resulted in statistically different scores.

Following these results, we compared the differences of the mean scores in each level (e.g., the 1,000 level of the aural version against the 1,000 level of the written version) with the Bonferroni adjustment, and as a result they were significant in all the eight levels ($p < .01$), with medium to large effect sizes from the 1,000 to the 8,000 level ($r = .25, .56, .21, .47, .24, .44, .40$). Focusing on the differences in the levels, it was found that the differences between the mean scores for any pairs of levels 5,000-8,000 in each mode of the test (aural and written respectively) were not significant, except the pair of levels 6,000-8,000 in the aural version of the test, meaning that at this point (the 5,000 level) the participants' vocabulary size seemed to reach its limit. It can thus be assumed that for the participants of the current study the levels over 5,000 were almost the same in difficulty. Considering this leveling-off phenomenon, we need to pay further attention to the difference between the aural and written vocabulary size tests up to the 4,000 level. In view of this, the mean score difference in level 3,000 requires further investigation since it shows a large gap between the aural and written versions of the vocabulary size tests.

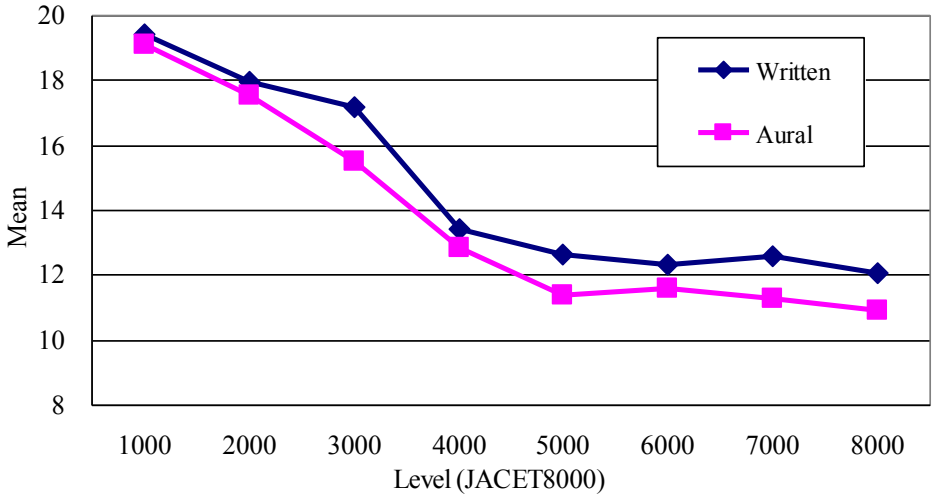
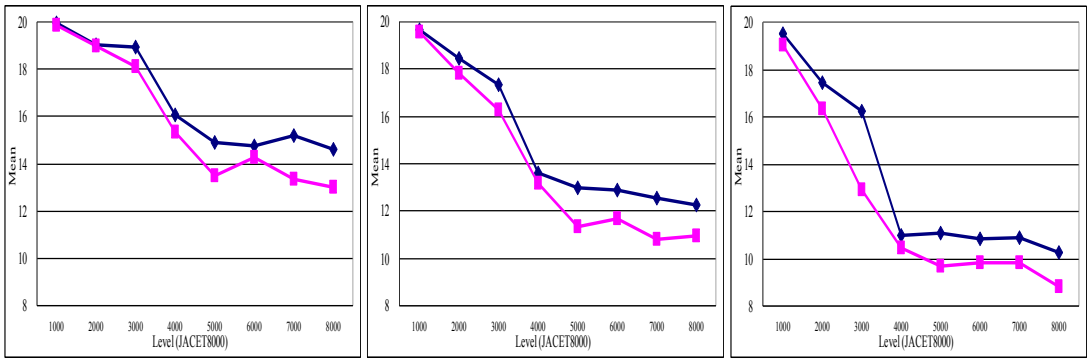


Figure 1. Mean scores of the written and aural vocabulary size tests.

3.3 *Learners with Different Proficiency Levels*

In order to closely investigate the differences of two versions of vocabulary size tests, participants were divided into three groups depending on their proficiency measure—the total TOEIC score (n = 209, TOEIC Mean = 471.05, SD = 127.41). Based on the group dividing method used for item analysis in language testing (Nakamura 2002), about 27 percent (about ± 1.2 SD from the mean) of the high and low scoring groups were screened, and the rest (about 46 percent) were assigned as the middle group. The breakdown of all participants with TOEIC scores was as follows: the higher group (n = 54, TOEIC Mean = 637.86, SD = 90.90), the middle group (n = 99, TOEIC Mean = 460.20, SD = 42.93), and the lower group (n = 56, TOEIC Mean = 327.69, SD = 48.32). Figure 2 displays the mean scores of the written and aural vocabulary size test for these three levels of learners.



Higher (n = 54) Middle (n = 99) Lower (n = 56)

Figure 2. Mean scores of the written and aural vocabulary size tests for three proficiency levels. ◆shows the means of the written version, and ■ is those of the aural test.

Again, the analysis showed the overall tendency that the mean scores for the written version of the test were higher. Figure 2 also exhibits that as the learners become more proficient, their vocabulary size gets larger both in aural and written versions of the vocabulary size test. Nevertheless, the large gap between the aural and written vocabulary size observed at the 3,000 level is more apparent in the low level group as shown in Figure 2—the mean score of the aural version drops sharply at the 3,000 level.

3.4 *Item Analysis*

For the purpose of scrutinizing what caused the differences of the aural and written vocabulary size tests, item analysis was conducted using *Test Data Analysis Program 2.0* (Nakamura, 2002) and *ITEMAN 3.6*. The means of the item difficulty index and the discrimination power index for each of the levels in the aural and written vocabulary size test are shown in Table 4 and 5. The values in the item difficulty index have numbers between 0 and 1, with 1 the easiest and 0 the hardest. Item difficulty index is defined as the percentage of test takers who answered the question correctly. Hence, it is evident in Table 4 and 5 that the percentage of correct answers gradually decreases until it becomes almost the same after the 5,000 level.

Table 4

Item Difficulty Index and Discrimination Power Index for the Aural Vocabulary Size Test

	Item difficulty index		Discrimination power index	
Level	*M*	*SD*	*M*	*SD*
1000	0.96	0.04	0.38	0.21
2000	0.87	0.12	0.36	0.12
3000	0.77	0.13	0.46	0.08
4000	0.64	0.18	0.37	0.10
5000	0.57	0.20	0.34	0.10
6000	0.58	0.17	0.35	0.08
7000	0.56	0.26	0.34	0.07
8000	0.54	0.25	0.33	0.11

Note. $N = 332$; Number of items = 20; *M* = Mean, *SD* = Standard Deviation

Table 5

Item Difficulty Index and Discrimination Power Index for the Written Vocabulary Size Test

	Item difficulty index		Discrimination power index	
Level	*M*	*SD*	*M*	*SD*
1000	0.97	0.02	0.40	0.18
2000	0.89	0.12	0.44	0.12
3000	0.85	0.14	0.41	0.11
4000	0.66	0.20	0.39	0.10
5000	0.63	0.21	0.38	0.09
6000	0.61	0.19	0.36	0.13
7000	0.62	0.19	0.36	0.10
8000	0.60	0.24	0.37	0.10

Note. $N = 332$; Number of items = 20; *M* = Mean, *SD* = Standard Deviation

This result indicates that learners in this study seemed to perceive the difficulty of the words above the 5,000 word level in the same way. The discrimination power index, which shows how well the test can separate (discriminate) higher-level learners and lower-level learners, was more than 0.30. Since point biserial correlations were used for this analysis, a discrimination power index over 0.30 can be considered satisfactory (Nakamura, 2002).

As is clear from the results showing the mean scores in Figure 1 and Figure 2, items in level 3,000 required an item analysis in order to further examine the reasons causing the differences in the aural and written vocabulary size of participants in the study. The proportion of the correct answers and the discrimination index were checked, and three patterns emerged as a result. Table 6 summarizes the first group of words. Note that in this analysis the lower discrimination index shows that the item cannot distinguish learners with lower scores from those with higher scores. The written version of the test resulted in a considerably higher proportion of correct answers, thus making the discrimination index lower. It can be surmised that being able to see the spelling of the words made it easier for the test takers to understand their meanings and not to endorse other distractors. Furthermore, there is a possibility that the participants used word parts knowledge (roots, prefixes, and suffixes) because words such as *signature*, *uncertainty*, *hidden*, and *efficiency* in Table 6 include these features.

Table 6

Words with High Proportion of Correct Answers in the 3,000 Level (*N* = 332)

Word	Aural Version		Written Version	
	Proportion of correct answers	Discrimination index	Proportion of correct answers	Discrimination index
signature	0.74	0.50	0.92	0.22
uncertainty	0.72	0.56	0.93	0.19
grip	0.79	0.51	0.90	0.21
hidden	0.73	0.56	0.86	0.37
efficiency	0.64	0.46	0.94	0.19
landing	0.82	0.43	0.93	0.20

The second group of words is shown in Table 7. The difference between the aural and written versions for each of the three words in question (*upper*, *concrete*, and *shelf*) is not big, but the discrimination index varied. From the reasons for the differences in the discrimination index it can be inferred that in the aural version of the test, the words *upper* and *shelf* were more difficult for lower level learners than for higher level learners, thus making the discrimination index higher, whereas in the written version, this was not the case. Both lower and higher levels of learners performed in the same manner with the help of the spelling of those words in the test booklet. Exactly the same interpretation can be also applied to the fact that low-level learners could not perform well in the aural version of the test. One exception was the word *concrete*. A further investigation into which option the test takers endorsed revealed that in the aural version, fewer high-level learners and more low-level learners chose the correct answer, and in the written version vice versa. This might be due to the fact that the word *concrete* has an equivalent loan word used as a noun. Hence, low-level learners made full use of this advantage when listening to the sound of it;

however, they could not associate its sound with the spelling and the word meaning when looking at it.

Words such as *silk*, *celebration*, and *rat* were placed into the third group, and these words showed a very high proportion of correct answers (above 0.95) in either aural or written version of the vocabulary size test. Emergence of this distinctive group can be explained with the same logic argued above: using knowledge of English loan words in Japanese. Since these words have varying degrees of familiarity in usage in Japanese as loan English words, they might have attracted test takers.

Table 7

Words with a Characteristic Discrimination Index ($N = 332$)

Word	Aural Version		Written Version	
	Proportion of correct answers	Discrimination index	Proportion of correct answers	Discrimination index
upper	0.90	0.27	0.97	0.06
concrete	0.52	0.39	0.50	0.59
shelf	0.84	0.38	0.88	0.22

3.5 *Correlations with the Proficiency Measures*

The correlations of the two types of vocabulary size tests (aural and written) and TOEIC, a proficiency measure used in this study, were examined. Participants who reported their TOEIC total score, listening score, and reading score were chosen for this analysis. The number of the individuals in this particular analysis, their mean scores, and standard deviations are illustrated in Table 8.

Table 8

Means and Standard Deviations of the Variables Used for the Correlation Analysis ($n = 209$)

	M	Possible Score Range	*SD*
TOEIC	473.56	10 – 990	128.55
TOEIC Listening	278.06	5 – 495	72.57
TOEIC Reading	195.67	5 – 495	66.69
Aural Vocabulary Test	111.78	0 – 160	17.22
Written Vocabulary Test	120.15	0 – 160	15.56

Presented in Table 9 are the correlation coefficients of the TOEIC total score, the TOEIC listening score, the TOEIC reading score, the aural vocabulary size, and the written vocabulary size. As can be seen in Table 9, both the aural and written vocabulary size tests have high correlation coefficients with the TOEIC total, listening, and reading scores (Aural = .70, .60, .69, Written = .69, .56, .72, respectively). The correlations between the written vocabulary size and TOEIC reading scores (.72) was higher than that of aural vocabulary size and the listening scores (.60). Looking at the relationship between the TOEIC listening and the two types of vocabulary size figures, it was found that the aural vocabulary size test has

slightly higher correlation coefficients (.60) than that of the written vocabulary size test (.56). Interestingly, the opposite was the case for the correlation between the TOEIC reading and the aural vocabulary size test (.69) and the written vocabulary size test (.72). These results may be interpreted as showing that each vocabulary size test reflects the skill being tested, i.e. listening and reading. The correlation coefficients of the aural vocabulary size test and written vocabulary size test were considerably higher (.89). Even though these two were regarded as counterparts and we had expected that the behaviors of these two types of tests would be different, the very highly correlated figure proved that it was not the case.

Table 9

Correlation Matrix of Proficiency Measures and Vocabulary Size Tests (n = 209)

	TOEIC Total	TOEIC Listening	TOEIC Reading	Aural Vocabulary	Written Vocabulary
TOEIC Total	-				
TOEIC Listening	.93	-			
TOEIC Reading	.92	.71	-		
Aural Vocabulary	.70	.60**	.69	-	
Written Vocabulary	.69	.56	.72	.89	-

4. Discussion

4.1 *The findings for Research question 1: Is there a significant difference between learners' aural and written vocabulary size? If there is a difference, how do they differ? Do they differ in separate frequency bands?*

From the results, it is clear that the two types of tests differ significantly; thus, showing the difference in learners' aural and written vocabulary size. The participants' written vocabulary size turned out to be larger than their aural vocabulary size. Even though the difficulty of the level over 5,000 was almost the same for the participants in the current study, it is still obvious that the aural vocabulary size test was more difficult than its written counterpart in any of the frequency bands. This finding is consistent with the previous studies sampling other Japanese EFL learner groups (Katagiri, 2001; Yamauchi, 2005). Now that we have gained these results along with the findings in the previous studies, we are in a better position to argue that there is a gap in the knowledge of spoken forms and written forms of the words for Japanese EFL learners. What we have shown in this study is that there are cases, especially for Japanese EFL learners, where the phonological forms of some words have not been acquired after their orthographical forms are learned.

4.2 *The findings for Research question 2: Is there a significant difference between aural and written vocabulary size in learners with different proficiency levels? If there is any difference, how can we account for it?*

Further analysis by dividing participants into three proficiency groups revealed that the written vocabulary size was larger than the aural one in all the three levels of learners. The lower group was poor in their

performance of the aural vocabulary size test, especially in the 3,000 level. From this observation, we can argue that aural vocabulary size tests might help unveil this kind of gap in vocabulary knowledge. Subsequent item analysis made it clear that some words are hard for low-level learners, and they cannot retrieve their meanings without the help of written presentation of the words. This may suggest that learners at this level can improve their vocabulary knowledge by learning the pronunciation of the words they have not mastered and becoming able to distinguish the sound of these words better.

4.3 *The findings for Research question 3: How do aural and written vocabulary size relate to listening and reading abilities, and overall English proficiency?*

Both the aural and written vocabulary size of the participants correlated strongly with TOEIC total, reading and listening scores. The fact that the correlations between the written vocabulary size and TOEIC reading scores were higher than that of the aural vocabulary size and the listening scores seems to suggest that factors other than vocabulary size might affect listening more than reading comprehension. The aural vocabulary size test provided slightly higher correlation coefficients with the TOEIC listening score than the written version, whereas with the TOEIC reading score the written version correlated slightly higher than the aural version. This result might be interpreted as each vocabulary size test (aural and written) better reflecting the listening and reading abilities respectively. However, considering the fairly high correlation coefficients of the two types of vocabulary size tests and the fact that both the aural and written vocabulary size tests yielded relatively high correlation coefficients with the TOEIC total score, it would be fair to conclude that the behaviors of the two types of vocabulary size tests are almost the same in predicting the listening and reading abilities as well as overall English proficiency. Katagiri (2001) observed the same tendency and suggested that if predicting the learners' listening ability is the goal, administering the aural vocabulary size test will not be necessarily required from a practical point of view given that they have taken the written version. It can also be argued that we have gathered evidence that irrespective of the aural or written version, vocabulary size tests have a strong relationship with listening, reading, and overall proficiency.

4.4 *Limitations*

Before discussing conclusion and implications, limitations in the current study should be pointed out. First, the proficiency of the participants was homogeneous, with pre-intermediate and intermediate level learners being dominant. A follow-up study will be needed to confirm that the findings of this study will be observed if a more proficient group or beginner level group of learners is included in the sample population. Therefore, studies depending on a different developmental stage will unveil the complete relationship between the aural and written vocabulary knowledge. Secondly, a qualitative research design may be necessary because this study was completely based on quantative data. A closer look at each individual learner will explain the research findings which were not fully interpretable with the data in this type of quantative study. Thirdly, the test length might have been problematic (160 questions); thus, the fatigue factor of the participants may have clouded the results we gained. As Meara and his colleagues have demonstrated (e.g., Meara & Milton, 1993), developing a computer-assisted shorter version will

solve this problem. Finally, due care should be exercised when interpreting the result of the aural format test, because the test format was not a pure aural test. The participants got the translation of the target word, so they could use this as a way of guessing what the target word was going to be. When they heard the target word, they only needed to confirm whether their guess was correct or not. This is not the same as hearing a word without a supporting context. Additionally, presenting the four options aurally is significantly harder than presenting them visually. There is a possibility that the lower scores we got may be a result of this change.

5. Conclusion and Implications

In this study, we attempted to explore the differences and characteristics of aural and written vocabulary size for Japanese EFL learners by developing and administering two types of tests. The results showed that there is a significant difference between learners' aural and written vocabulary size in any frequency band for Japanese university EFL learners. Also, it was found that in comparing different proficiency levels, less proficient learners have difficulty in connecting the sound of some words with its meaning even when they can do so in the written form. Regarding the relationship between vocabulary size tests and listening, reading, and overall English proficiency, it can best be summarized that both the aural and written versions of the vocabulary size tests have high correlations with proficiency measures, indicating that both tests are almost equally effective in predicting the results of proficiency measures.

Based on these findings, pedagogical implications can be drawn. Since the difference in aural and written vocabulary size has been verified, and low-level learners have been detected to contribute to this discrepancy, teachers should incorporate the instruction of pronunciation in vocabulary teaching particularly for less proficient learners. In addition, in the current study, the gradual decrease in the mean scores in both the aural and written versions of the tests up to the 5,000 level indicates that aural and written vocabulary size is a result of a linear vocabulary learning process to a certain degree, in spite of the fact that vocabulary is often learned as a discrete item in Japanese EFL settings. Adolphs and Schmitt (2003) claim that around 5,000 individual words are required to cover about 96% of the words used in every day spoken discourse. Taking this into consideration, EFL learners desirably need to understand the spoken forms of high frequency 3,000 word families and more in order to actively participate in general conversation. It may be necessary, therefore, for teachers to teach high-frequency words with their pronunciations explicitly in their classes. Students should be encouraged to learn the pronunciation, spelling, and the meaning of a word at the same time. Reading aloud, parallel reading (reading aloud while listening), or shadowing of an appropriate text would help to meet this end. Although pronunciation teaching is being emphasized in the *Course of Study* for junior and senior high schools, students are not always taught the phonetic transcription of a word. Introducing the phonetic transcription of words, therefore, might make our students more aware of pronunciation. Also, in order to prepare learners to deal with low-frequency words, teachers should teach vocabulary learning strategies or encourage incidental learning (Nation, 2001; Nation & Waring, 1997).

Acknowledgments

The authors are grateful to Dr. Paul Meara, University of Wales, for his insightful comments and suggestions on the manuscript. Also, our thanks are due to Dr. Rie Koizumi for her insightful suggestions on data analysis, and Professor Tomoyasu Miyano, Kansai Gaidai College, for data collection.

References

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics*, *24*, 425-438.

Aitchison, J. (1994). *Words in the mind* (2nd ed.). Oxford: Blackwell.

Aizawa, K. (1998). Developing a vocabulary size test for Japanese EFL learners. *ARELE*, *9*, 75-82.

Anderson, C. R., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 77-117). Newark, DE: International Reading Association.

Coady, J., Magota, J., Hubbard, P., Graney, J., & Mokhtari, K. (1993). High frequency vocabulary and reading proficiency in ESL readers. In M. Haynes & J. Coady (Eds.), *Second language reading and vocabulary learning* (pp. 217-228). Norwood, NJ: Ablex.

Daulton, F. (1998). Japanese loanword cognates and the acquisition of English vocabulary. *The Language Teacher*, *22*, 17-25.

Educational Testing Service (2006). *Test of English for International Communication (TOEIC)*. Retrieved October 18, 2006, from http://www.ets.org/toeic

Field, A. (2005). *Discovering statistics using SPSS* (2nd ed.). London: SAGE Publications.

Fountain, L. R. & Nation, P. (2000). A vocabulary–based graded dictation test. *RELC Journal*, *31*, 29-44.

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition*, *21*, 303-317.

Hu, M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*, 403-430.

JACET (2003). *JACET8000: JACET word list of 8000 basic words*. Tokyo: JACET.

Katagiri, K. (2001). *Developing the ten-minute vocabulary tests for quick and approximate estimates of general English ability of Japanese EFL learners*. Unpublished Ph.D. dissertation submitted to the United Graduate School of Education Tokyo Gakugei University.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P. Arnaud & H. Bejoint (Eds.), *Vocabulary and applied linguistics* (pp. 126-132). London: Macmillan.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*, 399-436.

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.

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer & J. Williams (Eds.), *Performance and competence in second language acquisition* (pp. 35-53). Cambridge: Cambridge University Press.

Meara, P., & Jones, G. (1990). *Eurocentres vocabulary size tests 10KA*. Zurich: Eurocentres Learning Service.

Meara, P., & Milton, J. (2003). *X_LeX: the Swansea Vocabulary Levels Test*. Swansea: Lognostics. Available from http://www.swan.ac.uk/cals/calsres/lognostics.htm

Meara, P., & Wolter, B. (2004). V links: beyond vocabulary depth. *Angles on the English-Speaking World*, *4*, 85-96.

Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners. *The Canadian Modern Language Review*, *63*, 127-147.

Mochizuki, M. (1998). Nihonjin eigo gakushusha no tameno goi saizu tesuto. [A vocabulary size test for Japanese learners of English.] *Institute for Research in Language Teaching Bulletin*, *12*, 27-53.

Nakamura, Y. (2002). *Test de gengonouryoku wa hakareruka?* [*Can tests measure language proficiency?*]. Tokyo: Taishukan.

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

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

Nation, P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), V*ocabulary: Description, acquisition and pedagogy* (pp. 6-19). Cambridge: Cambridge University Press.

Nurweni, A., & Read, J. (1999). The English vocabulary knowledge of Indonesian university students. *English for Specific Purposes*, *18*, 161-175.

Qian, D. (1999). Assessing the role of depth and breadth of vocabulary knowledge in reading comprehension. *The Canadian Modern Language Review*, *56*, 282-307.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, *10*, 355-371.

Read, J. (1996). Validating a test of depth of vocabulary knowledge. In A. Kunnan (Ed.), *Issues in language testing research: Conventional validity and beyond* (pp. 41-60). Mahwah, NJ: Lawrence Erlbaum Associates.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Richards, C. J. (1976). The role of vocabulary teaching. *TESOL Quarterly*, *10*, 77-89.

Sato, R. (2003). Shingakushuushidouyouryou wo fumaeta chukouseiwo taishoutoshita reberubetu goisaizutesuto no sakusei. [The development of vocabulary size test for different proficiency levels of junior and senior high school students with the new national curriculum in mind.] *STEP Bulletin*, *15*, 25-38.

Schmitt, N. (1998). Tracking the incremental acquisition of second language vocabulary: A longitudinal study. *Language Learning*, *48*, 281-312.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Schmitt, N., & McCarthy, M. (Eds.). (1997). *Vocabulary, description, acquisition and pedagogy*. Cambridge: Cambridge University Press.

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition*, *19*, 17-36.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the vocabulary levels test. *Language Testing*, *18*, 55-88.

Shimamoto, T. (2000). An analysis of receptive vocabulary knowledge: Depth versus breadth. *JABAET Journal*, *4*, 69-80.

Wesche, M., & Paribakht, T, S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *The Canadian Modern Language Review*, *53*, 13-40.

Yamauchi, Y. (2005). Nihonjin eigogakusyusya no listening goi to reading goi. [Listening vocabulary and reading vocabulary of Japanese EFL learners]. *Bulletin of FLT Association of University of Tokyo*, *9*, 91-103.