THE JOURNAL OF ASIA TEFL Vol. 6, No. 2, pp. 121-148, Summer 2009

How Many Words Do You Need to Know to Understand TOEIC, TOEFL & EIKEN? An Examination of Text Coverage and High Frequency Vocabulary

Kiyomi Chujo
Nihon University, Japan

Kathryn Oghigian
Waseda University, Japan

In order to understand the meaning of a written text, it is generally accepted that a reader should understand an average of nineteen out of twenty words, and this is referred to as "95% coverage." Given the popularity and importance of proficiency tests in second language acquisition, this study explores how much vocabulary a learner needs to know in order to be able to read and understand TOEIC, TOEFL and EIKEN proficiency tests. The vocabulary from several retired tests was compared to the vocabulary on three criterion lists: a high-frequency word list from the British National Corpus, the Standard Vocabulary List, and Nation's 14K word-family list. We were able to determine that in order to gain 95 percent coverage on TOEIC, a reader would need a minimum vocabulary size of 4,000 words, or 3,000 word families. TOEFL requires a 4,500-word vocabulary, or 3,500 word families, and a 5,500-word vocabulary, or 4,500 word families, is needed for EIKEN Pre-1st Grade. We also found that recent (2005/2006) versions of these tests require a smaller vocabulary compared to earlier versions. In addition, the SVL appears to provide a more accurate range of vocabulary for the lower level EIKEN tests but that a high correlation among results for the three criterion lists indicates these are stable and

effective tools for determining text coverage vocabulary size.

Key words: text coverage, vocabulary size, proficiency tests, criterion vocabulary lists

INTRODUCTION

Since foreign language acquisition requires the continuation of learning over the long term, it is useful to have some kind of benchmark to indicate the development of language proficiency. In Japan, English proficiency tests such as the Test of English for International Communication (TOEIC), the Test of English as a Foreign Language (TOEFL), and the EIKEN Test in Practical English Proficiency (EIKEN) are currently enjoying a surge in popularity in both the workplace and universities. Many colleges and universities award credits in English courses to students who meet test score requirements, and an increasing number of companies are using their scores as one criterion for selecting new recruits and for promoting employees. For example, the Japanese Ministry of Education, Culture, Sports, Science and Technology announced the target scores for Japanese English teachers as TOEIC 730, TOEFL 550, and EIKEN Grade Pre-1 or beyond (Tanabe, 2004).

There are a number of factors that influence test scores including reading ability, listening ability, grammatical knowledge, writing ability, and the vocabulary size of the learner. Beglar and Hunt (2005, p. 7) remind us that "vocabulary acquisition is a crucial, and in some senses, the central component in successful foreign language acquisition" and as educators we know that vocabulary is the heart of a language. Learners depend on vocabulary as their first resource (Huckin & Bloch, 1993) and a rich vocabulary makes the skills of listening, speaking, reading, and writing easier to perform (Nation, 1994). Therefore, how many words you need to know in order to do certain things is important in second language acquisition (Miura, 2005; Nation, 2006). It is from this context that we have explored how many words learners need to know to understand proficiency tests such as TOEIC, TOEFL, and EIKEN.

Understanding Text Coverage

How many words in a text must a reader know in order to understand what is being read? Historically, experienced teachers such as West (1926) suggested the guideline that one unknown word in every fifty words would be the minimum threshold necessary for the adequate comprehension of a text. This means 98% of the total number of words should be familiar. Hu and Nation (2000) stated this is the minimum desired level for comprehending a written narrative. Hatori (1979) considered 95% 'coverage,' or one unknown word in every twenty words, to be the threshold, a conclusion later supported by many contemporary researchers in the field of vocabulary teaching and learning (Clay, 1991; Hayashi, 2002; Honig, 2001; Juel, 1994; Laufer, 1989; Laufer, 1992; Nation, 2001; Read, 2000; Schmitt & McCarthy, 1997; Tono, 1997). Knowing that learners should be able to understand 19 of every 20 words in a text is a useful guide for educators, and applying text coverage indices to learners' texts and tests is important to ensure these materials are at the appropriate level.

Text coverage is calculated by counting the number of the known words in the text, multiplying this number by 100 and then dividing by the total number of words in the text. The idea of using text coverage to determine the optimal ratio of known words in a text has been commonly used since 1936 when H. E. Palmer selected 3,000 words for the *Interim Report on Vocabulary Selection* (see Schonell et al., 1956 or Takefuta & Chujo, 1993). We know that text coverage is more stable when the criterion vocabulary list is larger (Chujo & Utiyama, 2005), and with advances in corpus linguistics, larger criterion vocabulary lists are available which provide more stable text coverage than earlier lists such as the 2,000-word *A General Service List of English Words (GSL)* (West, 1953) and the 570-word family Academic Word List (Coxhead, 2000).

In 2004, Chujo developed a 13,994-word lemmatized high frequency word list (HFWL) from the British National Corpus (hereafter BNC HFWL) and she established a means for assessing the vocabulary levels of various texts and tests through text coverage. A vocabulary level can be defined in terms of the number of words counted from the top of a ranked criterion vocabulary

list, such as the BNC HFWL, which would be needed to achieve 95% coverage of the targeted text or test. For example, the vocabulary levels of the top selling junior and senior high school textbooks in Japan were rated at around 3,000 words and those of TOEIC were approximately 4,000 words, meaning a learner would need to know about 3,000 BNC HFWL words to successfully navigate the school textbooks, and about 4,000 to navigate TOEIC. The particular words would be those ranked from 1 to 3,000, or 1 to 4,000, from the top of the list as the most frequent words in English. Other large-scale criterion vocabulary lists are the 12,000-word Standard Vocabulary List (SVL) developed by ALC (2001) which is based on several word lists and corpora, and Nation's (2006) fourteen 1,000-word-family lists (supplemented by Cobb in 2007).

Other than Chujo's 2004 study and those of her research team (2003, 2004, 2005, 2006), very little has been reported on how much vocabulary is needed for attaining a successful understanding of the vocabulary on proficiency tests. One exception is Ishida (2004) who investigated the text coverage of TOEIC, TOEFL, and EIKEN, using the SVL as a comparative list, and reported that TOEIC and EIKEN Grade Pre-1 are on an almost similar level, and that TOEFL surpasses EIKEN Grade 1 in terms of required vocabulary knowledge. In other words, a learner would need to know a similar number of words for EIKEN Grade Pre-1 and TOEIC, but many more words for TOEFL than for EIKEN Grade 1. In contrast, Chujo (2004) reported very different results. She found that EIKEN Grade 1 requires more vocabulary than both TOEIC and TOEFL. One way to account for the discrepancy in the results between these two studies may be that Chujo used the BNC HFWL as the criterion vocabulary list in her study, as opposed to Ishida's SVL. In addition, the Chujo study used a 2002 EIKEN test and the Ishida study used a 2003 version. The discrepancy might be indicative of the criterion vocabulary lists or the publication dates of tests used. Another possibility might be that Ishida's findings seem to be based on the observation of the number of lower rated words rather than a 95% coverage level calculation. Thus, a difference in methodology may also explain the difference in findings.

Aims of the Present Study

From the studies discussed earlier, we know that text coverage has often been used to measure the number of known words in a text, and that the current thinking in the field of vocabulary teaching and learning puts the threshold of meaningful input at 95% coverage. We also know that how the text coverage is calculated and the criterion vocabulary list used is likely to affect the text coverage results, as will the publication year of the tests used. Building on the studies discussed in the previous section, the aim of this current study is to continue to define some of the parameters used in text coverage calculations, specifically regarding how variables such as different types of criterion vocabulary lists and test versions affect the stability of text coverage, as well as to investigate vocabulary levels of the TOEIC, TOEFL, and EIKEN tests. More specifically, the research questions of this study are:

(1) How many words are needed to cover 95% of the vocabulary in TOEIC, TOEFL, and EIKEN? In other words, how many vocabulary words would a reader need to know in order to understand 95% of the vocabulary (an average of 19 out of 20 words) on TOEIC, TOEFL and EIKEN?
(2) Is there any difference in the vocabulary level between the original tests and the recently revised versions?
(3) How do the different criterion vocabulary lists (BNC HFWL, SVL, and Nation's 14K) affect the vocabulary level of the tests? In other words, what differences, if any, occur in text coverage calculations between the three criterion lists?

METHOD

Understanding Criterion Vocabulary Lists

All three criterion vocabulary lists used in this study to measure the number of words at the 95% text coverage level of the targeted tests were

based on the British National Corpus (BNC) to a greater or lesser degree. The BNC is one of the largest electronically-accessible corpora consisting of over 100 million words in British English (Burnard, 2000; Kennedy, 2003; Leech, Rayson, & Wilson, 2001). It consists of an approximately 90 million-word written component of informative and imaginative texts, and a 10 million-word spoken component. The three criterion vocabulary lists are described in detail below.

The British National Corpus High Frequency Word List (BNC HFWL)

The BNC HFWL is a list of 13,994 lemmatized words representing 86 million BNC words that occur 100 times or more (Chujo, 2004). It was created by: (a) using the CLAWS7 tag set to extract all base forms; (b) lemmatizing by inflectional form; (c) deleting any low frequency or unusual words (those appearing fewer than 100 times in this lemmatized list); and (d) identifying all proper nouns and numerals by their part of speech tags and deleting manually.

The Standard Vocabulary List 12000 (SVL)

The SVL is a list of 12,000 words specifically developed for Japanese learners of English by the publisher ALC. They emphasize high-frequency words for both native speakers' usefulness and their importance for Japanese learners. The SVL is based on various word lists and corpora including the BNC, along with a special consideration for Japanese learners of English. There are 12 levels of 1,000 words, and they are available on the web at http://www.alc.co.jp/goi/PW_top_all.htm. Although the SVL is also a lemma list, it includes some proper nouns and numerals. In order to be comparable with the BNC HFWL, we deleted the proper nouns and numerals from the SVL, and this reduced the list to 11,312 words.

Nation's 14K

Nation's 14K is a list of fourteen 1,000 word-families. Although the BNC HFWL and Nation's 14K were developed from the same BNC, they are different lists because they use a different counting system; the former is based on lemma and the latter is based on word-families. Nation (2006, p. 63) explains

> For example, the word-family of *abbreviate* contains the following members: *abbreviate*, *abbreviates*, *abbreviated*, *abbreviating*, *abbreviation*, *abbreviations*. This family consists of two lemmas: the *abbreviate* lemma with four members and the *abbreviation* lemma with two members. Word-families include several lemmas and so the frequency, range, and dispersion figures for the lemmas are underestimates of what the figures would be for word-families.

This criterion vocabulary list now has in excess of 20 levels, although at the time of our study, only 14 were available. The words are grouped into 1,000 word families, and the words within each group are not ranked. This criterion vocabulary list comes with dedicated software called Vocabulary Profiler (Cobb, 2008; Nation & Heatley, 2002) currently available at http://www.lextutor.ca/vp/bnc/.

Proficiency Tests Examined

To examine the vocabulary levels, three types of English proficiency tests were collected. The TOEIC, the TOEFL, and the EIKEN are used extensively in Japan and many Japanese learners who study English are likely to encounter at least one if not all of these tests. The sources for these tests are listed in the Appendix.

In this study, word lists were created for each of these tests. To create the word lists, all the text data from the collected tests were scanned into a computer and were proofread. Next, proper nouns and numerals were excluded from each test manually. Finally, for each test, the inflectional variants were listed under a base form of a word or the same lemma, i.e., the

conjugation of verbs, and the declension of nouns, pronouns and adjectives, and were collated into a word list using software programs we developed (see http://www5d.biglobe.ne.jp/~chujo/eng/index.html).

TOEIC

The TOEIC test, developed in Japan in 1979 by the Educational Testing Service (ETS), is designed to evaluate learners' communication skills in business and daily life situations. It is a paper-and-pencil test composed of 200 multiple-choice questions. The TOEIC Bridge test, also administered by the ETS, is a simplified version of the TOEIC that targets beginner and intermediate learners and is composed of 100 multiple-choice questions. On the newer TOEIC tests (as of 2005), some of the question types have been changed but the difficulty range is the same. The test now "reflects global business communication styles and emphasizes authentic language contexts" (ETS). In this study, two practice tests included in the Official Guide for the original (2002) and new (2005) TOEIC were used. There is no available Official Guide with recent practice tests for the TOEIC Bridge.

The types (number of different words) and tokens (total number of running words) were counted, and are presented in Table 1. Both the types and tokens of the TOEIC Bridge are less than half of TOEIC. We see from Table 1 that the number of tokens in the new TOEIC tests has increased while the number of different words has not changed. The reason for the increase in tokens may be a reflection of longer reading sections on the test.

TABLE 1
TOEIC Types and Tokens, 2002 & 2005

Tests			Types	Tokens
TOEIC Official Guide	2002	Test 1	1,411	7,642
		Test 2	1,552	7,035
	2005	Test 1	1,560	9,463
		Test 2	1,595	9,924
TOEIC Bridge Official Guide	2002	Test 1	637	2,358
		Test 2	643	2,399

TOEFL

The TOEFL, also developed by the ETS, is an admission tool used by colleges and universities in Canada and the United States. This test is also a popular means of measuring a student's practical English proficiency, and many colleges in Japan grant English credits to students who meet TOEFL score requirements. In this study, six versions of older tests and two versions of newer tests (called TOEFL iBT) were collected from retired tests that are available to the public. It should be noted that Test 2 listed in the table is not a retired test, but a practice test published by a non-ETS publisher[1]. The types and tokens appearing in these TOEFL tests were counted. They are presented in Table 2. We see from Table 2 that the tokens in the new TOEFL tests increased, however, the number of different words has not changed greatly.

TABLE 2
TOEFL Types and Tokens, 1998/9 & 2006

Tests			Types	Tokens
TOEFL Practice Tests	1998/9	Test 1	1,401	7,078
		Test 2	1,324	5,768
		Test 3	1,502	7,540
		Test 4	1,476	7,374
		Test 5	1,410	7,176
		Test 6	1,496	7,784
TOEFL iBT	2006	Test 1	1,540	11,205
		Test 2	1,156	9,096

EIKEN

The EIKEN tests have been developed by the Society for Testing English Proficiency (STEP). STEP is a Japanese nonprofit organization established in 1963 in cooperation with the Japanese Ministry of Education, and the EIKEN test

[1] Although not a retired test, Test 2 is noted as "authentic" by its author (Phillips) and publisher (Longman), who note "it is based on the most up-to-date information available on the iBT" (p. xi).

was developed as a measure of language proficiency. (For more information, see http://stepeiken.org/.) There are seven EIKEN tests, which are, in order of increasing difficulty: Grade 5, Grade 4, Grade 3, Grade Pre-2, Grade 2, Grade Pre-1 and Grade 1. The EIKEN Grade 2 test is generally considered to be a desirable target level of English proficiency for high school graduates. Many Japanese learners of English aim at passing the EIKEN Grade 1 test and the EIKEN website indicates this level is on par with most graduate and undergraduate degree programs (STEP, 2007). In this study, two test versions for each grade were collected from retired tests available to the public. As for Grade 1, Grade Pre-1, and Grade 2, both old (2002) and new (2005) EIKEN tests were collected. The types and tokens appearing in these EIKEN tests were counted, and are presented in Table 3. We see from Table 3 that the types have remained fairly constant, but there has been a slight increase for the new EIKEN tests in the number of tokens for most but not all levels.

TABLE 3
EIKEN Types and Tokens, 2002 & 2005

Tests			Types	Tokens
Eiken Grade 1	2002	Test 1	1,780	7,307
		Test 2	1,740	7,249
	2005	Test 1	1,619	7,179
		Test 2	1,717	8,021
Eiken Grade Pre-1	2002	Test 1	1,493	6,087
		Test 2	1,397	5,990
	2005	Test 1	1,472	6,029
		Test 2	1,445	6,383
Eiken Grade 2	2002	Test 1	833	4,057
		Test 2	849	4,351
	2005	Test 3	940	4,869
		Test 4	930	5,189
Eiken Grade Pre-2	2005	Test 1	698	3,777
		Test 2	720	3,789
Eiken Grade 3	2005	Test 1	458	2,630
		Test 2	482	2,689
Eiken Grade 4	2005	Test 1	348	1,967
		Test 2	324	1,847
Eiken Grade 5	2005	Test 1	200	708
		Test 2	197	660

Calculation of Vocabulary Level

The next step was to assess the vocabulary levels of each test shown in Tables 1, 2 and 3 by using the three criterion vocabulary lists: the BNC HFWL, the SVL and Nation's 14K. As discussed earlier, leading researchers echo Nation's (2001, p. 114) emphasis that "learners would need at least a 95% coverage of the running words in the input in order to gain reasonable comprehension and to have reasonable success at guessing from context." Therefore, this coverage level was chosen as the target, and in this study, coverage is defined as being synonymous with vocabulary level. Using each criterion vocabulary list, we calculated the text coverage as follows:

BNC HFWL

We counted how many words from the top of the BNC HFWL that a learner would need to know in order for that learner to achieve an approximate 95% coverage of the targeted tests. In other words, each targeted test vocabulary level was defined in terms of the number of words counted from the top of BNC HFWL that account for 95% or more of the running words in that test. For the calculation, we used our own software program.

SVL

The SVL consists of twelve levels of 1,000-word sub-lists, each arranged in alphabetical order. We wanted to have more detailed vocabulary level distinctions than that provided by the 1,000-word levels. We had already reorganized the SVL into the same lemma-unit list as the BNC HFWL, and we accepted the alphabetical order within each sub-list as a type of rank order. In this way, the words in each SVL sub-list are grouped into 1,000 lemma units, and the words within each group are ranked. We counted how many words from the top of the SVL that a learner would need to know in order for that learner to achieve an approximate 95% coverage of the targeted tests. For

this calculation, we used the same program.

Nation's 14K

As mentioned earlier, Nation's 14K is equipped with a dedicated program, the Vocabulary Profiler (VP), to calculate coverage. We simply submitted the text data for each test, which already excluded proper nouns and numerals, to the VP program. The VP software calculated each 1,000-word base-word's text coverage over the targeted tests, and we looked at the number of 1,000-word base-words needed until the total coverage reached 95%. For example, when the total coverage reached the second 1,000-word base-words, it was denoted as 2K or 2,000 word-families. Since Nation's 14K is based on 'word families' units as mentioned above, and this is different from the BNC HFWL and SVL lemma units, we report the results as the required number of 'words' (lemmas) or 'word families.'

RESULTS AND DISCUSSION

How Many Words are Needed to Cover 95% of the TOEIC Vocabulary?

The BNC HFWL, the SVL, and Nation's 14K were used to identify the graduations among the diverse vocabulary levels contained within each of the three types of English proficiency tests. Table 4 shows the vocabulary levels of the TOEIC and the TOEIC Bridge tests investigated in this study. Each score shows how many words are required to obtain 95% coverage for each test. Two sets of each test were examined to increase reliability. As is shown in Table 4, each set indicated similar values, considering the inherent level differences between two sets. We can also see that the vocabulary level of the same tests of the same year is similar, indicating consistency within the sets.

We can see from Table 4 that a learner would need to know between 3,000 word-families and 4,000 words to understand a current TOEIC test, and

between 2,000 word-families and 2,500 words to understand the TOEIC Bridge.

TABLE 4
Vocabulary Size Necessary for 95% Coverage for TOEIC

	Tests		BNC HFWL (words)	SVL (words)	Nation's 14K (word-families)
TOEIC	2002	Test 1	3,879	4,436	3,000
		Test 2	4,139	4,436	4,000
		Average	4,009	4,436	3,500
	2005	Test 1	3,325	3,850	3,000
		Test 2	3,474	4,285	3,000
		Average	3,400	4,068	3,000
TOEIC Bridge	2002	Test 1	2,477	2,347	2,000
		Test 2	2,582	1,863	2,000
		Average	2,530	2,105	2,000

How Many Words are Needed to Cover 95% of the TOEFL Vocabulary?

Table 5 shows the vocabulary levels of the TOEFL tests investigated in this study. Six sets of older versions and two sets of recent versions were examined to increase reliability. As is shown in Table 5, each set indicated similar values, and we can also see that the vocabulary level of the same tests of the same year is similar, indicating consistency within two sets.

We see from Table 5 that there were a greater number of words needed to understand older versions of TOEFL than for recent versions. Recent versions of TOEFL require between 3,500 word-families and 4,500 words. This is substantially fewer words — for example, in looking at the BNC HFWL coverage, the figure drops from more than 6,000 to about 3,900, and similar drops are noted for the SVL and Nation's 14K. We might speculate that the tests are now using more high frequency words.

TABLE 5
Vocabulary Size Necessary for 95% Coverage for TOEFL

	Tests		BNC HFWL (words)	SVL (words)	Nation's 14K (word-families)
TOEFL	1998/9	Test 1	6,540	5,918	5,000
		Test 2	5,660	6,251	5,000
		Test 3	6,540	6,537	5,000
		Test 4	6,306	6,346	5,000
		Test 5	6,177	6,019	5,000
		Test 6	6,540	6,381	5,000
		Average	6,294	6,242	5,000
TOEFL	2006	Test 1	3,871	4,719	4,000
		Test 2	3,879	4,170	3,000
		Average	3,875	4,445	3,500

How Many Words are Needed to Cover 95% of the EIKEN Vocabulary?

Table 6 shows the vocabulary levels of the EIKEN tests investigated in this study. Two sets of each test were examined to increase reliability. As is shown in Table 6, each set indicated similar values, and again we can also see that the vocabulary level of the same tests of the same year is similar, indicating consistency within two sets.

Looking at the last line of Table 6, we see that to successfully navigate a recent (2005) EIKEN test, a learner would need to know an average of the top 900 SVL words to the top BNC HFWL 2,400 words for the least difficult level test (Grade 5). Not surprisingly, as the tests increase in difficulty, looking upward from the bottom (from Grades 5 to 4 or higher), the number of words increases. At the most advanced level (Grade 1) of the 2005 versions, a learner would need approximately 4,000 word-families to 6,200 words to maintain 95% coverage.

TABLE 6
Vocabulary Size Necessary for 95% Coverage Level for EIKEN

	Tests		BNC HFWL (words)	SVL (words)	Nation's 14K (word-families)
Grade 1	2002	Test 1	8,967	9,071	6,000
		Test 2	8,897	8,840	7,000
		Average	8,932	8,956	6,500
	2005	Test 1	4,985	5,820	4,000
		Test 2	5,992	6,668	4,000
		Average	5,489	6,244	4,000
Grade Pre-1	2002	Test 1	7,046	6,150	5,000
		Test 2	6,178	6,447	5,000
		Average	6,612	6,299	5,000
	2005	Test 1	5,916	5,109	5,000
		Test 2	5,017	5,258	4,000
		Average	5,467	5,184	4,500
Grade 2	2002	Test 1	3,219	2,772	3,000
		Test 2	2,787	2,697	3,000
		Average	3,003	2,735	3,000
	2005	Test 1	3,865	2,602	3,000
		Test 2	2,341	2,480	2,000
		Average	3,103	2,541	2,500
Grade Pre-2	2005	Test 1	2,526	1,650	2,000
		Test 2	2,994	1,828	3,000
		Average	2,760	1,739	2,500
Grade 3	2005	Test 1	1,915	1,076	2,000
		Test 2	1,881	901	2,000
		Average	1,898	989	2,000
Grade 4	2005	Test 1	2,043	893	2,000
		Test 2	2,526	893	3,000
		Average	2,285	893	2,500
Grade 5	2005	Test 1	2,455	893	2,000
		Test 2	2,266	893	2,000
		Average	2,361	893	2,000

We can also see that there was a considerable drop in the number of words required for the 2002 EIKEN Grade 1 tests and the more recent 2005 versions — from 9,000 words to about 6,000 words, or a drop from 6,500 word-families to about 4,000 word-families. The STEP website reports "TOEFL 610 was the mean score reported by examinees passing the First Stage of EIKEN Grade 1" (http://stepeiken.org/benefits/comparison-toefl.shtml). Presumably if a learner knew as many as 9,000 words (required for the 2002 EIKEN Grade 1 version) and could conceivably score 610 on the TOEFL, can the same claim now be made for the revised 2005 version which would only require two thirds of those words?

Finally, it is interesting that for the easier tests (Grades 3, 4, and 5), there is a large difference in the number of required words calculated from the SVL, as opposed to the BNC HFWL and Nation's 14K. In fact, they are more than 50% fewer in most cases. As noted earlier, the SVL, although based partly on the BNC, includes not only other corpora, but has been further refined to meet the particular needs of Japanese students. The BNC-based lists are British English words, so we might speculate that the lower number of SVL words reflect the publisher's adaptations. From this we can see the impact of a particular criterion list on text coverage and surmise that the SVL may be more accurate in identifying vocabulary level requirements for the lower level EIKEN tests. In other words, the range given in Table 6 of 900 words (SVL) and 2,400 words (BNC HFWL) or 2,000 word-families (Nation's 14K) for the lower EIKEN tests clearly illustrates the difference demonstrated by the three criterion lists. All three lists, however, seem to similarly identify vocabulary ranges for the higher level tests.

Is there any Difference in Vocabulary Level Between the Original Tests and the Recently Revised Versions?

In Table 7, we included the average scores from Tables 4, 5, and 6 and showed the vocabulary level differences between the 1998/1999/2002 tests and the 20005/2006 tests. We can see from Table 7 that all three criterion

vocabulary lists provide a similar pattern in changes in the number of words or word families. In all cases except for the EIKEN Grade 2 in comparison with the BNC HWFL, there was a drop in vocabulary level. This ranges from about 200 words to almost 3,500 words.

TABLE 7
Vocabulary Coverage Levels for 1998/99/2002 and 2005/06 Versions of TOEIC, TOEFL & EIKEN

Tests		BNC HFWL (words)	SVL (words)	Nation's 14K (word-families)
TOEIC	2002	4,009	4,436	3,500
	2005	3,400	4,068	3,000
	Difference	-609	-368	-500
TOEFL	1998/9	6,294	6,242	5,000
	2006	3,875	4,445	3,500
	Difference	-2,419	-1,797	-1,500
EIKEN Grade 1	2002	8,932	8,956	6,500
	2005	5,489	6,244	4,000
	Difference	-3,443	-2,712	-2,500
EIKEN Grade Pre-1	2002	6,612	6,299	5,000
	2005	5,467	5,184	4,500
	Difference	-1,145	-1,115	-500
EIKEN Grade 2	2002	3,003	2,735	3,000
	2005	3,103	2,541	2,500
	Difference	+100	-194	-500

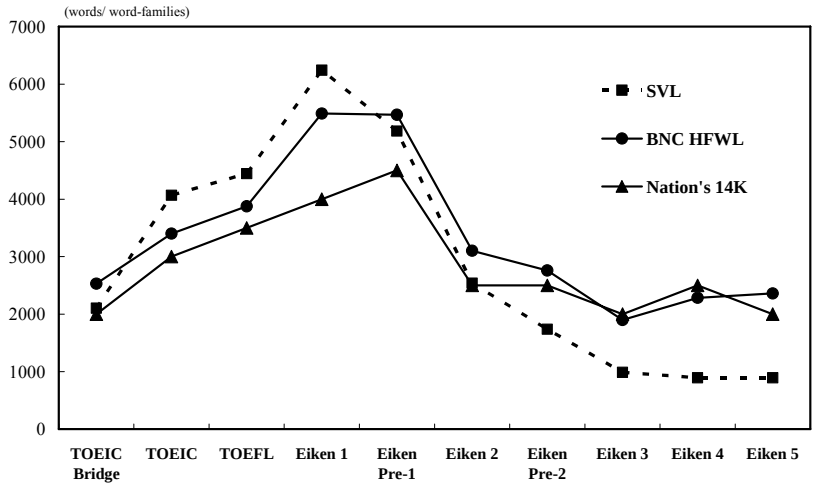
So far there has been no research-based or official explanation for this decrease in vocabulary level. We could speculate that the vocabulary levels of the new tests such as the TOEFL iBT might not be stabilized yet and in fact it is true that we weren't able to collect a larger number of retired tests. This however would not explain the drop for other tests. One possibility may be the influence of the rapid development of corpus linguistics. Since the BNC became available outside EU countries in 2000, the application of corpus results has greatly impacted not only theories and application in second language acquisitions but the development of dictionaries, vocabulary

and grammar books, and course books. Some dictionaries were compiled from self-built corpora such as the *Collins Cobuild English Dictionary* (Sinclair, 2001) and the *Wisdom English-Japanese Dictionary* (Inoue & Akano, 2003), and other dictionaries referred to the BNC for their compilation. Frequency word lists are easily available from various types of corpora such as the American National Corpus and the International Corpus of English. With the support of ETS, Biber and his associates (2004) constructed a corpus of *The TOEFL 2000 Spoken and Written Academic Language Corpus (T2K-SWAL)* to investigate the real-life academic language use. It is reasonable to speculate that the comment on the ETS website that recent versions of the TOEIC "…emphasize[s] authentic language contexts" (ETS) may be an indication of test makers' recognition of high frequency words found in corpora and their incorporation in the tests. We believe that to accurately reflect authentic language, recent corpus results should be reflected in the creation of English proficiency tests.

How do the Different Criterion Vocabulary Lists Affect the Vocabulary Level of the Tests?

Figure 1 shows the vocabulary levels of ten pairs of the recently revised (2005/06) tests (TOEIC Bridge, TOEIC, TOEFL and the seven-level EIKEN tests) each measured by the three criterion vocabulary lists. The line graph shows how many words from each criterion vocabulary list are required to obtain 95% coverage for each test. Two sets of each test were examined to ensure reliability. For the sake of simplicity, we averaged together the two vocabulary levels of each pair of tests and dotted these in the graph below. The vocabulary levels of the ten pairs of tests measured by the BNC HFWL are shown by round dots and a line, the SVL by square dots and a dotted line and those of Nation's 14K are shown by triangular dots and a line.

FIGURE 1
The Vocabulary Coverage Level Distribution of Each Test Based on the BNC HFWL, SVL, and Nation's 14K



The graduation seen in Figure 1 indicates several interesting results. We can see clearly that the TOEIC Bridge requires less vocabulary than the TOEIC tests, and the TOEFL tests require more vocabulary than the TOEIC tests. It is notable that the vocabulary levels of the EIKEN Grade 1 and Pre-1 are higher than the TOEFL tests by all three criterion vocabulary lists. In addition, it is remarkable that the vocabulary level difference between the EIKEN Grade 1 or Pre-1 and the EIKEN Grade 2 is so large (and bigger than the TOEIC Bridge and TOEFL) that it is understandable for Japanese students to not be able to pass beyond the EIKEN Grade 2 test without significantly expanding vocabulary size.

The SVL calculations ranked the seven EIKEN tests in approximately the same order of the proficiency level as reported by the authors: EIKEN Grade 1, EIKEN Grade Pre-1, EIKEN Grade 2, EIKEN Grade Pre-2, then EIKEN Grade 3, and finally EIKEN Grade 4 and 5, although the distinction of the last two levels are blurred. However, the BNC HFWL and Nation's 14K did

not show a distinctive rank between EIKEN Grades 3, 4, and 5. Such a result would indicate that the SVL, by its own claim, was developed with a special consideration for Japanese learners of English, suggesting that it is more suitable for estimating the vocabulary level of materials made for Japanese learners of English such as Japan-made EIKEN tests, while the BNC HFWL and the Nation's 14K were made from the BNC which 'represented formal, adult, British language' (Miura, 2005, p. 12), indicating the distinction of the vocabulary levels of materials such as those words lower than the 2,000 most common words might be difficult to grasp. As Nation explains in Miura (2005, p. 12).

> I tried to make my own list from the British National Corpus because it is one of the largest well-organized corpora of English. I wanted to see if whether using that list would be a way of making a new *GSL*. But it became clear after I made the first 3,000 words that that wasn't the way to do it. One of the reasons for that was that the corpus didn't represent the needs of second or foreign language learners.

It is also interesting to note that Cobb (2008) recently added 'VP-Kids' to his program to identify lexical growth in young children, especially in the important K-2 phase.

Looking at the line graphs overall, we can see that the zigzag lines of the three criterion vocabulary lists go up and down almost in unison, corresponding to the proficiency levels of each test. This indicates that the coverage requirement for each criterion vocabulary list (BNC HFWL, the SVL and Nations' 14K) falls in a similar range, or that all three criterion vocabulary lists provide a similar pattern in the coverage required for each test. We see that the fluctuation of the SVL is the largest of the three, while that of the Nation's 14K, whose vocabulary levels are at 1,000-word segments and based on word-families, is the smallest. The fact that all three lines demonstrate the same pattern also indicates there is some correlation among the comparisons.

As we saw from Figure 1, there is a considerable degree of agreement

among the vocabulary levels measured by the three criterion vocabulary lists. In order to understand the extent of the similarity among them, we next analyzed the data using the Spearman rank-order correlation coefficient for a set of paired rankings of each two criterion vocabulary lists. The raw scores, i.e. vocabulary levels from all tests, for the combined data of 34 tests (6 TOEIC tests + 8 TOEFL tests + 20 EIKEN tests) were converted to ranks from small size to larger size such as rank 1, 2, 3, 4, …and 34, and the differences between the ranks of each observation on the two criterion vocabulary lists were calculated.

Table 8 shows that the relationship between the BNC HFWL and the SVL is significant ($r\gamma^2$=0.9461, 32 d.f., ρ <0.000001); that the BNC HFWL and Nation's 14K is significant ($r\gamma^2$=0.9576, 32 d.f., ρ <0.000001); and that the SVL and Nation's 14K is significant ($r\gamma^2$=0.9159, 32 d.f., ρ <0.000001). It is not surprising that the BNC HFWL and Nation's 14K had the highest correlation because they are created from the same BNC list. The three vocabulary levels measured by the three criterion vocabulary lists highly correlated with each other. This level of correlation indicates that the three measures are closely related, given the three measured vocabulary levels of all proficiency tests.

TABLE 8
Spearman Rank Correlation between Criterion Vocabulary Lists

	BNC HFWL	SVL	Nation's 14K
BNC HFWL	—		
SVL	0.9461	—	
Nation's 14K	0.9576	0.9159	—

CONCLUSION

From this study we know that the vocabulary levels in all three proficiency tests have dropped in the last few years, and that in order to reach a 95% coverage level of proficiency tests, a learner would need to know between

3,000 word-families and 4,000 words for TOEIC, between 3,500 word-families and 4,500 words for TOEFL, and 4,500 word-families and 5,500 words for EIKEN Grade Pre-1. Data analyzed for different versions of the same type of test are within the same range, indicating that there is consistency within tests; and the vocabulary level of each test measured by the three large-scale criterion vocabulary lists showed a similar graduation and a high correlation to each other, indicating that all three criterion vocabulary lists were useful for estimating the vocabulary levels of the tests examined here. Because it is specifically tailored to Japanese learners, the SVL may be the best predictor of vocabulary range for the lower level EIKEN tests.

On a practical level, we might speculate that a drop in required vocabulary for more recent versions might be indicative of a trend by test publishers to use more high frequency words. For students, understanding the differences in required levels might help them plan tests taking strategies. For example, a student first taking TOEIC, then TOEFL, then the EIKEN Pre-1 will understand this reflects a graduation of the number of required words and comparative test scores might be more informative. Also knowing that a huge increase in vocabulary is necessary to be successful on EIKEN Pre-1 or Grade 1 from Grade 2 might better prepare students, teachers and the creators of test study materials. Finally, to provide access to the lemmatized TOEIC, TOEFL and EIKEN vocabulary lists, we have posted them on http://www5d.biglobe.ne.jp/~chujo/eng/index.html.

In terms of measuring text coverage, Nation's 14K might be the easiest list to use for most educators, since it is equipped with a program and all that is required is to type or paste text into this program and click "submit." However, the observations suggest that the SVL is possibly the most suitable list for estimating materials made for Japanese language learners. Although the disadvantage of the SVL is that it is not equipped with software for measuring vocabulary sizes, a freeware vocabulary profiling program such as AntVocabCheck 1.01 (2008) might be useful. This is also true for the BNC HFWL.

Given the importance of vocabulary in language learning, studies such as this one underscore the value of high frequency vocabulary lists available from various corpora, and understanding text coverage enables educators to better guide learners to a more effective and efficient means for learning.

THE AUTHORS

*Kiyomi Chuj*o is an Associate Professor at the College of Industrial Technology, Nihon University, Japan. She completed her Ph.D. on vocabulary selection for English education at Chiba University, Japan in 1991. Her current research interests are vocabulary learning and the pedagogical applications of corpus linguistics such as Data-Driven Learning. She was the recipient of JAECS Award in 2008 (Japan Association for English Corpus Studies).
Email: chuujou.kiyomi@nihon-u.ac.jp

Kathryn Oghigian teaches at Waseda University, Center for English Language Education in Science and Engineering. She completed her master's degree in Modern Language Education at the University of British Columbia, Vancouver, Canada. She is currently involved in research in applied corpus linguistics and in the development of SLA material for elementary education in Japan.
Email: k_oghigian@aoni.waseda.jp

REFERENCES

ALC (2001). *Standard vocabulary list (SVL)* 12000. Retrieved 5/4/2006 from the World Wide Web http:// www.alc.co.jp/goi/PW_top_all.htm

AntVocabCheck 1.01 (2008). [Computer software]. Available from http://www.antlab.sci.waseda.ac.jp/software.html

Beglar, D., & Hunt, A. (2005). Six principles for teaching foreign language vocabulary: A commentary on Laufer, Meara, and Nation's "ten best ideas." *The Language*

Teacher, 29(7), 7-10.

Biber, D., Conrad, S., Reppen, R., Byrd, H. P., Helt, M., Clark, V., Cortes, V., Csomay, E., & Urzua, A. (2004). Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus. *ETS TOEFL Monograph Series MS-25*, Princeton, NJ: Educational Testing Service. Available from http://www.ets.org/research/researcher/RM-04-03.html

Burnard, L. (2000). *Reference guide for the British National Corpus (World Edition)*. Retrieved February 10, 2009 from the World Wide Web http://www.natcorp.ox.ac.uk/World/HTML/thebib.html

Chujo, K., & Nishigaki, C. (2003). Bridging the vocabulary gap: From EGP to EAP. *JACET Bulletin, 37*, 73-84.

Chujo, K. (2004). Measuring vocabulary levels of English textbooks and tests using a BNC lemmatized high frequency word list. In J. Nakamura, N. Inoue, & T. Tabata, (Eds.), *English corpora under Japanese eyes* (pp. 231-249). Amsterdam: Rodopi.

Chujo, K., & Genung, M. (2004). Comparing the three specialized vocabularies used in 'business English,' TOEIC, and British National Corpus spoken business communications. *Practical English Studies*, *11*, 49-63.

Chujo, K., & Genung, M. (2005). Utilizing the British National Corpus to analyze TOEIC tests: The quantification of vocabulary-usage levels and the extraction of characteristically used words. *TOEIC Research Report, 3*, 1-23.

Chujo, K., & Utiyama, M. (2005). Understanding the role of text length, sample size and vocabulary size in determining text coverage. *Reading in a Foreign Language, 17*(1), 1-22. Available from http://nflrc.hawaii.edu/rfl/

Chujo, K., & Hasegawa, S. (2006). An investigation into the star-rated words in English-Japanese learner's dictionaries. *International Journal of Lexicography*, *19*(2), 175-195.

Clay, M. (1991). *Becoming literate: The construction of inner control*. Portsmouth, NH: Heinemann.

Cobb (2006). Web VP (Version 2.6) [Computer software]. British National Corpus lists version (Nation's BNC 14K). (2006). Retrieved February 10, 2009 from the World Wide Web http://www.lextutor.ca/vp/bnc/ (Now Web VP / BNC-20 v 3.0 British National Corpus lists version available. Retrieved March 31, 2007 from the World Wide Web http://www.lextutor. ca/vp/bnc/

Cobb, T. (2008). Web Vocabprofiler (Version 2.6) [Computer software]. Accessed 03/17/08 from http://www.lextutor.ca/vp/

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213-238.

ETS. (2007). New TOEIC. Retrieved March 31, 2007 from the World Wide Web

http://www.ets.org/portal/site/ets/menuitem.22f30af61d34e9c39a77b13bc3921509/?vgnextoid=1b5ed87dcc9ae010VgnVCM10000022f95190RCRD

Hatori, H. (1979). *Eigo shidouhou handbook (4) Hyouka-hen* [A handbook for English teaching (4) evaluation]. Tokyo: Taishukanshoten.

Hayashi, H. (2002). *Eigo no goi shidou* [Teaching English vocabulary]. Hiroshima: Keisuisha.

Honig, B. (2001). *Teaching our children to read*. Thousand Oaks, CA: Corwin Press.

Hu, M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension, *Reading in a Foreign Language*, *13*(1). 31, 2007 from the World Wide Web http://www.vuw.ac.nz/ lals/staff/paul_nation/marcella.rtf

Huckin, T., & Bloch, J. (1993). Strategies for inferring word-meanings in context: A cognitive model. In T. Huckin, et al. (Eds.) *Second language, reading and vocabulary acquisition* (pp. 153-180). Ablex, NJ: Norwood.

Inoue, N., & Akano, I. (2003). *The wisdom English-Japanese dictionary*. Tokyo: Sanseido.

Ishida, M. (2004). Eigo kyouin ga sonaeteoku-beki eigo-ryoku [Targeted English proficiency level for Japanese English teachers]. *Eigo Tenbou* (*ELEC Bulletin*) 111:10-17.

Juel, C. (1994). *Learning to read and write in one elementary school*. New York: Springer-Verlag.

Kennedy, G. (2003). Amplifier collocations in the British National Corpus: Implications for English language teaching. *TESOL Quarterly*, *37*(3), 467-487.

Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316-323). Clevedon, UK: Multilingual Matters.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In L. Arnaud & H. Bejoint (Eds.), *Vocabulary and applied linguistics* (pp. 126-312). London: Macmillan.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English*. Harlow, CA: Pearson Education Limited.

Miura, T. (2005). Interview with Paul Nation: The past, present, and future of second language vocabulary acquisition. *The Language Teacher*, *29*(7), 11-14.

Nation, P. (1994). *New ways in teaching vocabulary*. Alexandria, VA: TESOL, Inc.

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

Nation, P., & Heatley, A. (2002). Range: A program for the analysis of vocabulary in texts [Computer software]. Available from http://www.vuw.ac.nz/lals/staff/paul-nation/nation.aspx

Nation, P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, *63*(1), 59-82.

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

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

Schonell, F. J., Meddleton, Y., Shaw, B., Routh, M., Popham, D., Gill, J., et al. (1956). *A study of the oral vocabulary of adults*. Brisbane: University of Queensland Press.

Sinclair, J. (Ed. in chief) (2001). *Collins COBUILD English dictionary for advanced learners*. (Third Edition.) Glasgow: HarperCollins Publishers.

Society for Testing English Proficiency (STEP). (2007). http://stepeiken.org/.

Takefuta, Y., & Chujo, K. (1993). Goi lisuto no kyakkanteki hyouka hikaku no tame no yukoudo shihyou no kaihatsu [The development of text coverage for evaluating word lists quantitatively]. *Working Papers in Language and Speech Science*, *4*, 68-84.

Tanabe, Y. (2004). What the 2003 MEXT action plan proposes to teachers of English. *The Language Teacher*, *28*(3), 3-8. Available from http://www.jalt-publications.org/tlt/articles/2004/03/tanabe

Tono, Y. (ed.) (1997). *Eigo Goi Shuutoku-ron* [*Theories of teaching and learning English vocabulary*]. Tokyo: Kagensha.

West, M. (1926). *Learning to read a foreign language*. London: Longman, Green & Co.

West, M. (1953). *A general service list of English words*. London: Longman, Green & Co.

APPENDIX

Tests Used

TOEIC & TOEIC Bridge

The Chauncey Group International (2002). *TOEIC Koushiki Guide & Mondaishuu* [TOEIC Official Test-Preparation Guide & Practice Tests]. Tokyo: IIBC (The Institute for International Business Communication).

The Chauncey Group International (2002). *TOEIC Koushiki Guide & Mondaishuu Vol.2* [TOEIC Official Test-Preparation Guide & Practice Tests Vol.2]. Tokyo: IIBC.

Educational Testing Service (2005). *TOEIC Test ShinKoushiku Mondaishuu* [TOEIC New Official Test-Preparation Guide & Practice Tests]. Tokyo: IIBC.

The Chauncey Group International (2002). *TOEIC Bridge Koushiki Guide & Mondaishuu* [TOEIC Bridge Official Test-Preparation Guide & Practice Tests]. Tokyo: IIBC (The Institute for International Business Communication).

TOEFL

Educational Testing Service (1998). *TOEFL Practice Tests Volume 1*. Princeton, N.J.: Educational Testing Service.

Educational Testing Service (1999). *TOEFL Practice Tests Volume 2*. Princeton, N.J.: Educational Testing Service.

Educational Testing Service (2006). *THE Official Guide TO THE NEW TOEFL iBT*. New York: McGraw-Hill.

Phillips, D (2006). *Longman Preparation Course for the TOEFL Test: iBT*. White Plains, NY: Person Education, Inc.

EIKEN

Obunsha (2002). *Eiken 1 Kyu Zenmondaishuu* [Practice Tests for Eiken Grade 1]. Tokyo: Obunsha.

Obunsha (2002). *Eiken Jun 1 Kyu Zenmondaishuu* [Practice Tests for Eiken Grade Pre-1]. Tokyo: Obunsha.

Obunsha (2002). *Eiken 2 Kyu Zenmondaishuu* [Practice Tests for Eiken Grade 2].

Tokyo: Obunsha.
Obunsha (2005). *Eiken 1 Kyu Zenmondaishuu* [Practice Tests for Eiken Grade 1]. Tokyo: Obunsha.
Obunsha (2005). *Eiken Jun 1 Kyu Zenmondaishuu* [Practice Tests for Eiken Grade Pre-1]. Tokyo: Obunsha.
Obunsha (2005). *Eiken 2 Kyu Zenmondaishuu* [Practice Tests for Eiken Grade 2]. Tokyo: Obunsha.
Obunsha (2005). *Eiken Jun 2 Kyu Zenmondaishuu* [Practice Tests for Eiken Grade Pre-2]. Tokyo: Obunsha.
Obunsha (2005). *Eiken 3 Kyu Zenmondaishuu* [Practice Tests for Eiken Grade 3]. Tokyo: Obunsha.
Obunsha (2005). *Eiken 4 Kyu Zenmondaishuu* [Practice Tests for Eiken Grade 4]. Tokyo: Obunsha.
Obunsha (2005). *Eiken 5 Kyu Zenmondaishuu* [Practice Tests for Eiken Grade 5]. Tokyo: Obunsha.