Articles

The *Eiken* Vocabulary Section[1]: An Analysis and Recommendations for Change

Tsuyuki Miura
Temple University Japan
David Beglar
Temple University Japan

Although the Eiken is one of the most widely taken English proficiency tests in Japan, little empirical research has been conducted on the test. In this study, the vocabulary sections of all levels of the Eiken administered from 1998 to 2000 were analyzed. There were five principal findings: (a) successive levels of the Eiken vocabulary section do not increase in difficulty in a smoothly graduated fashion, (b) some test forms appear more difficult than others, (c) item options from widely differing frequency levels are sometimes used on the same item, (d) the assumed vocabulary sizes of targeted examinees frequently bear little relation to the difficulty of the items included in the vocabulary section, and (e) the sentence stems in the vocabulary section and the reading passages impose a similar lexical load. A number of suggestions for addressing the shortcomings of the vocabulary section are proposed.

実用英語検定試験(英検)は、日本で最も広く受験されている英語熟達度判定テストの一つであるにもかかわらず、実証的調査はほとんどなされていないのが実情である。本稿では１９９８年から２０００年の間に実施された英検全級の語彙問題分析を行い、結果として主に以下五点を挙げる。(a)各級間の難易度変化は均等ではない、(b)テストにより難易度に差がある、(c)一つの項目の選択肢に、頻度が大きく異なる語彙の使用が見られる場合がある、(d)実施者が想定する各級受験対象者の語彙力と、語彙問題における項目難易度は関連が薄い、(e)語彙問題の項目基幹部分と、長文読解問題の引用文は、語彙レベルにおいて近似している。さらに本稿では、語彙問題における問題点に対し、数多くの提案を掲げている。

JALT Journal, Vol. 24, No. 2, November, 2002

One of the most important English proficiency tests in Japan is the *Jitsuyo Eigo Ginou Kentei Shiken* (*Eiken*), which is developed and administered by the *Nihon Eigo Kentei Kyokai* (*Eikyo*). Nearly three million people took a version of the *Eiken* in 2001, and since the test's introduction in 1963, more than 61 million people have sat the exam. The *Eiken*, which is currently made up of seven different level-specific tests beginning with the fifth level and increasing in difficulty through pre-second, second, and pre-first to the first level, was characterized by MacGregor (1997) as being "highly respected in social, educational, and employment circles..." (p. 24) in Japan. This statement is supported by the fact that *Eiken* certification is accepted in lieu of sitting an entrance examination by some Japanese high schools, vocational schools, junior colleges, and universities, and passing particular levels of the test carries university credit in some institutions. In addition, more than one-third of the prefectures in Japan are currently using the *Eiken* as one way to determine the language proficiency of prospective English teachers (see www.eiken.or.jp for further details). Passing higher levels of the *Eiken* also enhances a person's chances to be hired and/or promoted in some companies.

Notwithstanding the *Eiken's* position of importance in Japan, there is a lack of published research that illuminates fundamental testing concerns such as reliability, validity, and test washback. Our investigation of Japanese and English-language educational and language testing journals uncovered surprisingly few investigations of the *Eiken*, and none directly related to the topic of this study. In an early study, Murakami (1972) questioned the Eiken's reliability and the quality of some items. A quarter of a century later, an exploratory examination of a pre-second level form of the *Eiken* was reported by MacGregor (1997), who arrived at five main conclusions. First, the test content appeared to match the intended group of test takers (second and third year high school students), a feature that MacGregor characterized as the test's greatest strength. However, MacGregor's other comments were critical, and they were derived from a cluster of reasons. Foremost among them was the charge that there is reason for concern about the test's reliability and validity. An additional related issue materialized as a result of an item analysis that she conducted. Approximately half of the items on the test were found to have unacceptable item discrimination values (a measure of how well an item differentiates high and low scoring examinees), a factor that would directly contribute to the fairly low reliability coefficient she found for the test form she investigated. Fourth, the context provided for some

items was unclear and even occasionally illogical, another characteristic that can adversely affect test reliability and validity. Finally, MacGregor argued that *Eikyo* should provide published reports of studies on item construction, reliability, and validity, a common practice of large testing companies such as Educational Testing Service in the United States.

Despite the criticisms of MacGregor's study raised by Henry (1998), her work represents an important initial attempt to illuminate the major strengths and weaknesses of the *Eiken*. In contrast to MacGregor, who chose to examine overall test functioning of one level of the test, we will begin a more focused line of research by investigating the *Eiken* vocabulary section. Our primary purpose is to undertake a preliminary analysis of the vocabulary section of all levels of the *Eiken* in order to determine the types of words being tested and to make recommendations for improving that section.

We have chosen to focus on the vocabulary section for three reasons. First, unlike some sections of the *Eiken*, a vocabulary section is included on each level of the test. Thus, unlike some other areas, it is one tested at all proficiency levels. Second, a number of studies conducted in the past decade have highlighted the importance of lexical knowledge for aural language processing (Miller & Eimas, 1995; VanPatten, 1996), speech production (Altman, 1997; de Bot, 1992; Levelt, 1993), reading (de Bot, Paribakht, & Wesche, 1997; Durgunoglu, 1997), and writing (Engber, 1995; Laufer & Nation, 1995). Third, we believe that research on the vocabulary section in particular is needed. The first author's experience and her discussions with other Japanese who have taken several levels of the *Eiken* suggest that the difficulty of the vocabulary section does not increase in smoothly graduated steps. Instead, the informal consensus is that the vocabulary sections of the pre-first and first level tests present unusually severe challenges in comparison with both the vocabulary sections of other levels of the test and with other test sections. Finally, the perception that some editions of the test (same level but appearing at different times) are easier than others, contributes to the feeling that the tests are not entirely fair.

The Importance of High Frequency and Academic Vocabulary

The notion that particular groups of words are of special importance has been largely inspired by corpus-based research undertaken in the past by researchers such as West (1953) and continued in the present in corpora such as Collins' COBUILD Bank of English Corpus (http://titania. cobuild. collins.co.uk/). Such corpora have consistently shown that

a small number of words account for a high percentage of the words met receptively and used productively. For instance, the 2,000 high frequency word families as represented by the headwords in West's (1953) *General Service List* (*GSL*) provide coverage of up to 75% of fiction texts (Hirsh, 1993), 90% of non-fiction texts (Hwang, 1989), and 80% of academic texts (Nation, 2001).

In addition, the 570 general academic word families included in the *Academic Word List* account for an average of about 10% of the running words in academic texts (Coxhead, 2000). Together, these approximately 2,600 word families (i.e., 2,000 high frequency and 570 academic word families) are crucial for academic success in English-language settings as shown by the fact that they accounted for 86% of the vocabulary in Coxhead's 3.5 million word academic corpus, and they constitute the majority of the 3,000 word families that are needed for learners to reach what Laufer (1992) has referred to as "the turning point of vocabulary size for reading comprehension" (p. 130).

In this study, the vocabulary appearing in the *Eiken* vocabulary section are compared with word lists of high frequency and academic vocabulary, the expected proficiencies of the targeted examinees, and the vocabulary on the reading comprehension section of the test. This analysis is an attempt to shed light on precisely what vocabulary is being tested on all levels of the *Eiken*, and the results should be instructive to the test's designers, teachers preparing students to take the *Eiken*, and the examinees themselves.

In addition to the general purpose stated above, we posed three specific research questions:

> *Research question 1:* What is the lexical composition of the multiple-choice vocabulary options (i.e., the correct answer and three distractors) on each level of the *Eiken* in terms of high frequency, academic, and low frequency vocabulary? How consistent is the lexical composition from one administration to the next?

In order to answer these questions, we examined all of the correct answers and distractors of all *Eiken* vocabulary tests administered from 1998 to 2000. These original tests are available in a series of seven books titled *Eiken Zenmondaishu* (e.g., *Eiken*, 2001a, b & c).

> *Research question 2:* To what degree are the results of the first research question in accord with the targeted profi-

> ciencies of the examinees and the vocabulary size for each level that is suggested by *Eikyo*? How appropriate are the targeted proficiencies identified by *Eikyo*?

The purpose of these questions was to investigate whether the vocabulary items in each level of the *Eiken* are consistent with the targeted vocabulary sizes specified by *Eikyo* (2001) and *Monbu-kagaku-sho* (Ministry of Education, Culture, Sports, Science, and Technology), as specified in *Gakushu Shidou Youryou* [(Foreign language in secondary school:) The Course of Study] (*Monbu-kagaku-sho*, 2001).

> *Research question 3:* How does the vocabulary of the item sentences (i.e., the stems and correct answer) in the vocabulary section compare with the vocabulary of the reading comprehension passages for each level?

The objective here was to compare the lexical load of the vocabulary section with that of the reading section. For this analysis, all levels of the *Eiken* administered in June 2000 were examined.

Method

The Range Program

All analyses were conducted with Range (Nation & Heatley, 1996), a PC program that is freely available at the University of Victoria at Wellington's web site (http://www.vuw.ac.nz/lals/). This software compares the words in a text or several texts with the words in three base lists and can be used to find the coverage of a text using preset word lists.

As noted above, Range detects and classifies three categories of words. The first is made up of the 1,000 most frequent words in English (3,126 types or 999 word families) and the second is comprised of the second 1,000 most frequent words (2,721 types or 986 families). The source of these words is *A General Service List of English Words* (West, 1953). Together these 1,985 word families constitute what is commonly referred to as the high frequency vocabulary of English.

The third category is made up of words not found among the high frequency words described above, but which frequently occur in upper secondary school and university textbooks across a wide range of academic subjects (2,540 types or 570 families). The source of these words is the *Academic Word List* (AWL) (Coxhead, 2000).

Range employs three types of units to count words. *Tokens* are tallied

by simply counting every word form in a spoken or written text. If the same word form occurs more than once, each occurrence is counted. *Types* are tallied by counting every unique word form only once. Additional occurrences are not counted. Let us look at one concrete example to help illustrate the idea. In the sentence, *Scientists know that the volume of the moon is the same as the volume of the Pacific Ocean*, there are 18 tokens (i.e., 18 words in the sentence) but only twelve types (i.e., twelve unique word forms). The final type of unit, *word families*, consists of a headword, its inflected forms, and its closely related derived forms. For example, *know* (headword), *knows* (inflected form), and *unknown* (closely related derived form) are all part of the same word family (Bauer & Nation, 1993). Although all three counts serve useful but distinct purposes, in this study we emphasized *types* because we were primarily interested in the occurrence of unique word forms. Finally, in addition to the three categories of words described above, Range indicates which words in a text are not covered by any of the above lists. Thus, a fourth category of low frequency vocabulary is automatically created by the program.

The Eiken Vocabulary Test Section

As noted above, all levels of the *Eiken* include vocabulary items in the first section of the test. The same multiple-choice, minimal context format is used for all levels, but the number of items on each level varies (see Table 1).

Table 1: Items Included in the Analyses

Test level	# of items per test	# of items inspected	# of items deleted/test	Total # of items deleted	Total # of items analyzed
First	30	180	6-7	38	142
Pre-first	30	180	6-7	38	142
Second	25	150	15	90	60
Pre-second	25	150	13-15	86	64
Third	20	120	8-11	56	64
Fourth	20	120	7-11	61	59
Fifth	15	90	5-9	46	44

The following is one item from the first level test administered in June 2000:

> After her pleasant first flight, the woman realized that her fear of flying had been ( ).
>
> 1. undaunted
> 2. unfounded
> 3. unabashed
> 4. unscathed
>
> (*Eiken*, 2001a, p. 14)

Each test that we examined also included a number of items testing knowledge of English idioms and grammar. For instance, the following item from the pre-first level tests idiomatic knowledge:

> The Internet stock's value grew ( ) soon after it was offered to the public. It rose 20% in one month.
>
> 1. out and about
> 2. by leaps and bounds
> 3. above and beyond
> 4. in bits and pieces
>
> (*Eiken*, 2001b, p. 19)

A typical fourth level grammar item is:

> George ( ) his friend in the park yesterday.
>
> 1. sees
> 2. will see
> 3. saw
> 4. seen
>
> (*Eiken*, 2001c, p. 28)

Because it is not possible to analyze multi-word units such as phrasal verbs and idioms with Range, these items, as well as the items testing

grammatical knowledge, were eliminated from the data set by both researchers working in consultation. Table 1 summarizes the number of items deleted from the analysis and the number of items remaining after the deletions. For instance, at the first level, 180 items were inspected (6 test forms x 30 items/per form), and depending on the specific form, six or seven items were deleted. This resulted in 38 total deletions. The remaining 142 items were used in the analyses.

The remaining multiple-choice options in all six administrations of the *Eiken* from 1998 to 2000 were then entered into Microsoft Word 2000 (2000). The files were then saved in text format so that they could be read by Range. Data files for each level consisted of the four multiple-choice options for each question, including the correct answers (e.g., *unfounded* in the first example test item above) and the three distractors (e.g., *undaunted*, *unabashed* and *unscathed*) for each item. The data from the six test forms were entered into separate test files so that we could investigate differences between the test forms.

The second set of data that were collected was for the item sentences (stems) in the vocabulary section along with the correct options (incorrect options excluded). Items that were excluded in the previous analysis were also excluded here.

The third data set was made up of the reading passages from the first to the fourth levels of the *Eiken* administered in June 2000[2]. The passages were entered into Microsoft Word 2000, converted to text format and then submitted to Range.

Results

The Multiple-Choice Vocabulary Options

The initial analysis concerned the multiple-choice options in the vocabulary section. Columns 4 to 7 in Table 2 summarize the results of the Range analysis. It can be seen, for example, that of all the types appearing in the fifth level test forms under examination, 95, or 81.2%, appear on Range's list of the 1,000 most frequent words. In general, the amount of higher frequency vocabulary decreases and the amount of lower frequency vocabulary increases as the tests move from the easiest (fifth) level to the most difficult (first) level, at which point over 90% of the vocabulary options are low frequency words.

Table 2: Targeted Examinees, Assumed Vocabulary Sizes, and Coverage of the Multiple-choice Options in the Vocabulary Section

Test Level	Targeted Examinees	Targeted Size	First 1,000 Types (%)	Second 1,000 Types (%)	AWL Types (%)	Low Frequency Types (%)
First	Four-year college grads	10,000 -15,000	8 (1.4)	11 (1.9)	23 (4.1)	523 (92.6)
Pre-first	Two-year college grads	7,500	32 (5.7)	55 (9.7)	133 (23.4)	346 (61.1)
Second	HS seniors	5,100	77 (33.5)	64 (27.8)	58 (25.2)	31 (13.5)
Pre-second	HS first & second year	3,600	143 (61.1)	57 (24.3)	17 (7.3)	17 (7.3)
Third	JHS third year	2,100	167 (77.7)	39 (18.1)	3 (1.4)	6 (2.8)
Fourth	JHS second year	1,300	161 (81.3)	34 (17.2)	2 (1.0)	1 (0.5)
Fifth	JHS first year	600	95 (81.2)	20 (17.1)	0 (0.0)	2 (1.7)

Note. HS = High school; JHS = Junior high school.

Variation among the lexical profiles of different administrations of the same test level was also investigated. The results for the first, pre-first, and second level test forms are displayed in Table 3. The second column shows the six administrations of the highest three levels of the *Eiken* included in this study, and columns 3 through 6 show the four lexical categories reported by Range. As can be seen, different versions of the same level test are not entirely consistent. For instance, the profiles of the June 1998 and the October 2000 administrations of the pre-first level show considerable variation, particularly where the second 1,000 word frequency level (column 4) and low frequency words (column 6) are concerned.

Table 3: Variation in the Lexical Distribution of Item Options on the First, Pre-first, and Second Level Test Forms

Test Level	Administration Date	First 1,000 Types (%)	Second 1,000 Types (%)	AWL Types (%)	Low frequency Types (%)
First	Oct. 2000	1 (1.1)	0 (0.0)	3 (3.2)	91 (95.8)
	June 2000	0 (0.0)	0 (0.0)	2 (2.1)	94 (97.9)
	Oct. 1999	0 (0.0)	0 (0.0)	2 (2.1)	94 (97.9)
	June 1999	2 (2.1)	4 (4.2)	5 (5.2)	85 (88.5)
	Oct. 1998	4 (4.3)	5 (5.4)	3 (3.3)	80 (87.0)
	June 1998	1 (1.1)	2 (2.2)	8 (8.7)	81 (88.0)
Pre-first	Oct. 2000	6 (6.3)	6 (6.3)	16 (16.7)	68 (70.8)
	June 2000	4 (4.2)	7 (7.3)	21 (21.9)	64 (66.7)
	Oct. 1999	3 (3.1)	10 (10.4)	23 (24.0)	60 (62.5)
	June 1999	5 (5.2)	8 (8.3)	23 (24.0)	60 (62.5)
	Oct. 1998	4 (4.3)	6 (6.5)	29 (31.5)	53 (57.6)
	June 1998	10 (10.9)	18 (19.6)	21 (22.8)	43 (46.7)
Second	Oct. 2000	15 (37.5)	17 (42.5)	3 (7.5)	5 (12.5)
	June 2000	16 (40.0)	13 (32.5)	10 (25.0)	1 (2.5)
	Oct. 1999	10 (25.0)	8 (20.0)	14 (35.0)	8 (20.0)
	June 1999	10 (25.0)	11 (27.5)	14 (35.0)	5 (12.5)
	Oct. 1998	15 (37.5)	7 (17.5)	15 (37.5)	3 (7.5)
	June 1998	15 (37.5)	9 (22.2)	7 (17.5)	9 (22.5)

The Multiple-choice Vocabulary Options and their Relationship to the Examinees

The results pertaining to research question 2 are displayed in Table 2. The targeted examinees are shown in the second column, and the targeted vocabulary sizes of the examinees are shown in the third column. These can be compared to the lexical composition of the different test levels. For instance, at the fourth level, second year junior high school students are expected to have a vocabulary of approximately 1,300 words. The test options match this target well as they are taken primarily from the first and second 1,000 most frequent words of English.

A Comparison of the Vocabulary and Reading Comprehension Sections

Our final research question concerned the degree of consistency between the vocabulary and reading comprehension sections of the *Eiken*. The results are shown in Table 4.

Table 4: Coverage of the Vocabulary Section Options, Vocabulary Item Sentences and Reading Section Passages

Test Level	Word List	Options: Types (%)	Sentence Stems: Types (%)	Reading Passages: Types (%)
First	1st 1,000	8 (1.4)	182 (58.5)	526 (51.4)
	2nd 1,000	11 (1.9)	39 (12.5)	112 (10.9)
	AWL	23 (4.1)	28 (9.0)	141 (13.8)
	Low Frequency	523 (92.6)	62 (19.9)	244 (23.9)
Pre-first	1st 1,000	32 (5.7)	172 (62.3)	385 (54.3)
	2nd 1,000	55 (9.7)	33 (12.0)	81 (11.4)
	AWL	133 (23.4)	22 (8.0)	80 (11.3)
	Low Frequency	125 (61.1)	49 (17.8)	163 (23.0)
Second	1st 1,000	77 (33.5)	107 (79.3)	311 (73.0)
	2nd 1,000	64 (27.8)	14 (10.4)	38 (8.9)
	AWL	58 (25.2)	4 (3.0)	29 (6.8)
	Low Frequency	31 (13.5)	10 (7.4)	48 (11.3)
Pre-second	1st 1,000	143 (61.1)	106 (80.9)	252 (78.5)
	2nd 1,000	57 (24.3)	11 (8.4)	21 (6.5)
	AWL	17 (7.3)	2 (1.5)	9 (2.8)
	Low Frequency	17 (7.3)	12 (9.2)	39 (12.1)
Third	1st 1,000	167 (77.7)	113 (75.3)	184 (82.9)
	2nd 1,000	39 (18.1)	15 (10.0)	20 (9.0)
	AWL	3 (1.4)	1 (0.7)	0 (0.0)
	Low Frequency	6 (2.8)	21 (14.0)	18 (8.1)
Fourth	1st 1,000	161 (81.3)	75 (81.5)	136 (86.1)
	2nd 1,000	34 (17.2)	7 (7.6)	10 (6.3)
	AWL	2 (1.0)	0 (0.0)	1 (0.6)
	Low Frequency	1 (0.5)	10 (10.9)	11 (7.0)

Item options for all test levels are shown in the third column. A comparison of the percentages found under Sentence Stems % (column 4) and Reading Passages % (column 5) shows that they are relatively close to each other throughout all test levels and for all word categories. In the first through pre-second levels, the sentences have a slightly greater proportion of high frequency vocabulary. This situation is reversed on the third and fourth levels where the vocabulary in the reading section appears to be slightly easier.

Discussion

The Multiple-choice Vocabulary Options

Five main points are deserving of comment. First, the degree of difficulty of the first level vocabulary section is now clear. More than 90% of the item options at the first level are low frequency words. Although low frequency words should be tested at this level, the gap in difficulty between the pre-first and first levels is quite large, as can be seen by the increase (61.1% to 92.6%) in low frequency vocabulary (Table 2).

Second, the largest jump in difficulty occurs between the second and pre-first levels. At the second level, high frequency vocabulary accounts for 61.3% of the distractors and low frequency vocabulary only 13.5%. However, when we move to the pre-first level, these numbers are effectively reversed: high frequency vocabulary has fallen to 15.4% and low frequency vocabulary has risen sharply to 61.1%. This sudden shift validates the subjective experience voiced by many Japanese examinees: The pre-first and first level vocabulary sections are far more difficult than the vocabulary found at other levels of the test.

Third, despite the fact that the first level is a test of low frequency vocabulary and the pre-first level a test of low frequency and academic vocabulary, high frequency words account for 3.3% (1.4% + 1.9%) of the options in the first level and 15.4% (5.7% + 9.7%) in the pre-first level. It is inappropriate to include such options on the highest two levels of the test. In order to illustrate the reason for this, let us look at one example from a pre-first level test administered October 18, 1998.

The politician got upset when he found his views had been ( ) by the journalist's misleading article.

1 adopted

2 distorted

3 implied

4 proclaimed

(*Eiken*, 2001b, p. 106)

Because of the frequency-sensitive nature of second language vocabulary acquisition, the higher the frequency level of a particular word, the higher the probability it is known.[3] In the above item, option 1 (*adopted*) is one of the most frequent 1,000 words of English, options 2 (*distorted*) and 3 (*implied*) are part of the AWL, and option 4 (*proclaimed*) is a low frequency item. This mixing of words from very different frequency levels increases the likelihood that a relatively high frequency option such as *adopted* will not function effectively as a distractor in the presence of lower frequency vocabulary because many examinees will be able to eliminate it relatively easily, or, if it is the correct option, choose it with little difficulty (see Haladyna, 1994 for an extensive review of multiple-choice item functioning and distractor analysis).

The fourth point concerns the similarity of the lexical profiles of the third, fourth, and fifth levels. Although each of these levels is appropriately focused on high frequency vocabulary, the lack of a shift in emphasis from the first to the second 1,000 word families suggests that there is no significant change in difficulty from one level to the next given the well-known influence of word frequency on lexical acquisition. We investigated this possibility more closely by randomly selecting 25 words each from the third, fourth, and fifth level vocabulary options and checking the precise frequency of those words with the Carroll, Davies, and Richman (1971) word frequency list. The fifth level test form was essentially a test of the 500 most frequent words of English and, as such, was easier than the third and fourth level tests. However, the composition of the third and fourth level tests was extremely similar in terms of word frequency. In addition, when all of the third and fourth level options were compared, it was found that 22.8% (38 out of 167 types) were included on both test levels. This degree of overlap is troubling on tests that are purported to be aimed at different proficiency groups.

Fifth, the major difference at the first level concerns a change made by *Eikyo* between the June and October 1999 administrations. As shown in Table 3, the 1998 and June 1999 administrations display consistent profiles, but the test writers appear to have made the test more difficult beginning with the October 1999 administration, at which time the test becomes almost entirely composed of low frequency vocabulary.

Inconsistencies also appear in the pre-first and second level test forms. For instance, the June 1998 pre-first level test appears to be considerably easier than the October 2000 administration based on the amount of low frequency vocabulary tested on each form—46.7% versus 70.8%. Furthermore, 80% (37.5% + 42.5%) of the vocabulary on the October 2000 second level test form is made up of high frequency vocabulary whereas the same vocabulary levels comprise only 60% of the June 1998 second level form.

The Multiple-choice Vocabulary Options and their Relationship to the Examinees

Our second research question concerned the targeted examinees, their assumed vocabulary sizes, the degree to which the *Eiken* vocabulary section is in accord with the assumed sizes, and the appropriateness of those assumptions. Table 2 shows the targeted examinees by educational level (column 2) and their assumed vocabulary sizes (column 3) as stated by *Eikyo* (2001). Vocabulary size is assumed to increase as grade level rises.

Let us first turn to the question of the degree to which the *Eiken* vocabulary sections are in accord with the target vocabulary sizes shown in Table 2. Answering this question is not entirely straightforward for two reasons. First, we do not know which words *Eikyo* counts as the targeted vocabulary because they do not disclose the word list(s) that they are using. Secondly, although *Eikyo* does not publicly disclose how it counts words, an *Eikyo* representative informed us that the test makers count words "like in a dictionary" (anonymous *Eikyo* representative, personal communication, February 24, 2002). This suggests that *Eikyo* may be counting words in a manner that is similar to our focus on word types. This is an important issue because word counts change significantly depending on what counts as a word. For instance, the first 1,000 high frequency words of English can be counted as 3,126 types or 999 word families.

Because of the large number of interrelationships between the cells in Table 2, we will highlight only a few of the more important points by focusing on the third column (targeted size) and the four columns that show the word type breakdowns for the four types of vocabulary (columns 4 to 7). *Eikyo* assumes that examinees taking the second level of the *Eiken* have a receptive vocabulary of approximately 5,100 words. However, if this is the case, it makes little sense to test the high frequency

words of English, and our data show that high frequency words account for approximately 60% (33.5% + 27.8%) of the words tested at the second level.

A second example of an apparent mismatch can be found at the pre-second level, for which *Eikyo* has stated that examinees should have a receptive vocabulary of approximately 3,600 words. Although *Eikyo* probably intends this figure to be an approximation, it is puzzling that 61.1% of the vocabulary options that we sampled from six different pre-second level tests were chosen from the first 1,000 words of English. These words should present no challenge to a learner with anything approaching a 3,600-word vocabulary.

One final example concerns the fifth through third levels. In spite of the fact that, as noted above, the examinees' vocabulary sizes are expected to increase from 600 words at the fifth level to 2,100 words at the third level, the actual data show that the three sections are made up of broadly similar items: The first 1,000 word level accounts for 81.2% of the words at the fifth level and 77.7% of the items at the third level. The second 1,000-word level accounts for 17.1% (fifth level) and 18.1% (third level) of the items. Thus, expected rises in examinees' receptive vocabularies are not mirrored by changes in the lexical profiles of the items on the test. In sum, we can only conclude that the items on the tests administered from 1998 to 2000 and the assumed vocabulary knowledge of examinees have at best a weak relationship with one another.

The second part of research question 2 asked about the appropriateness of the proposed vocabulary sizes shown in the third column of Table 2. For instance, is it reasonable to expect a third year junior high school student to have a 2,100-word receptive vocabulary? Although we have considerable unpublished data showing that this figure is quite high, there is little published research available to answer this question. However, we believe that the figures proposed in Table 2 are unrealistic in terms of the language acquisition of the average Japanese student. Barrow, Nakanishi, and Ishino (1999) reported that the Japanese learners in their study had receptive vocabularies of approximately 2,400 words on average after six years of formal English education. In other words, first year university students had vocabularies only slightly larger than the 2,100-word vocabulary proposed by *Eikyo* for third year junior high school students.

We can also analyze the appropriateness of the *Eiken* vocabulary section by comparing it with the vocabulary sizes that are endorsed by *Monbu-kagaku-sho*, as specified in *Gakushu Shidou Youryou* (*Monbu-*

kagaku-sho, 2001). In this document, *Monbu-kagaku-sho* suggests vocabulary learning goals for junior and senior high school students. These guidelines state that up to 900 words should selectively be taught during three years of junior high school, including basic vocabulary that relates to aspects of daily life such as seasons, months, days of the week, time, weather, ordinal and cardinal numbers, and the family. Furthermore, the Ministry sets a target of learning an additional 1,800 words for high school students. Thus, Japanese students are expected to learn approximately 2,700 words after six years of formal education. When we compare the *Monbu-kagaku-sho's* suggested vocabulary learning goals and the vocabulary test items on the *Eiken* test, it is difficult to identify a clear relationship between the two, a problem that is particularly acute at the higher levels of the *Eiken*.[4]

A Comparison of the Vocabulary and Reading Comprehension Sections

As noted in the Results section, the percentages found under Sentence Stems % and Reading Passages % in Table 4 show broad similarities for all test levels and word categories. This is appropriate because both sections should be targeted on the same proficiency level. Large differences would suggest that at least one section is not appropriate for the targeted examinees.

Two additional findings appear in Table 4. First, the multiple-choice options (column 3) at the first and pre-first test levels are composed of more difficult vocabulary than the sentence stems (column 4) and reading passages (column 5). While low frequency vocabulary makes up 92.6% of the options at the first level, it comprises only 19.9% of the sentence stems and 23.9% of the reading passages. At the pre-first level, low-frequency vocabulary accounts for 61.1% of the options, 17.8 % of the sentence stems and 23% of the reading passages. Thus, the multiple-choice vocabulary options in the first and pre-first levels pose the greatest lexical challenge for test takers at those levels.

The second finding concerns the relationship between the options and sentence stems at the third and fourth levels. Some sentence stems appear to be made up of more difficult vocabulary than the options. For instance, at the third level, 14% of the word types in the sentences are low frequency vocabulary, whereas only 2.8% of the options are low frequency. As a result, the sentence stem, whose purpose is to provide context for choosing the correct option, may sometimes be less comprehensible than the options themselves, and examinees may miss an

item not because they lack knowledge of the targeted vocabulary, but because they did not understand the sentence context.

Recommendations for Improving the Eiken Vocabulary Section

Our intention from the beginning of this study has been to investigate the *Eiken* vocabulary section, identify problematic areas, and make specific suggestions for improving the section. It is to this last goal that we now turn.

Our first finding was that the different levels of the *Eiken* vocabulary section do not increase in difficulty in a smoothly graduated fashion, and the difficulty levels of different test forms at the first, pre-first and second level are not consistent (see Table 3). The third and fourth levels show virtually no change and there are large gaps between the second and pre-first and the pre-first and first levels of the test (see Table 2). Although *Eikyo* has chosen this design based on "teachers' opinions and guidance from *Monbu-kagaku-sho*" (name withheld, personal communication, October 12, 2001), the result is an overall design that is at best clumsy and at worst ineffective. One way to remedy this problem would be to apply the following guidelines: (a) high frequency words should not be tested or included as distractors at the first, pre-first, and second levels, (b) the number of items sampled from the *AWL* should be increased at the pre-first, second, pre-second, and third levels, and (c) the first 1,000 words should be gradually deemphasized and the second 1,000 words gradually emphasized as the test moves from the fifth to the third level. *Eikyo* could implement this suggestion by utilizing software such as Range and by consulting multiple word frequency lists of written English when choosing words for inclusion on the tests. A second, and in our opinion, more elegant solution to this problem could be implemented through the proper use of item response theory (IRT). Although *Eikyo* informed us that they are using a form of IRT to analyze the tests (name withheld, personal communication, September 12, 2001), we see little evidence that they have taken advantage of the strengths of IRT. The Rasch model, which is a latent trait measurement model that places person ability and item difficulty on a single log linear scale, would permit *Eikyo* to produce vocabulary sections that sensitively measure lexical knowledge, avoid the gaps that we found at the higher levels of the test, equate test forms relatively easily, make shorter yet more reliable tests, and deliver the tests in a computer-adaptive format (see Bond & Fox, 2001 and Wright & Stone, 1979 for details regarding how these objectives can be achieved with the Rasch model). Using the Rasch

model and word frequency information to model reading development has been undertaken with considerable success in the United States by Lexile (www.lexile.com). This work could serve as a useful model for *Eikyo*.

Our second main finding concerned the use of multiple-choice options from widely varying frequency levels. We recommend that the four options for any single question be drawn from similar word frequency levels. As outlined earlier, the influence of word frequency effects is so pervasive that higher frequency distractors can be comprehended relatively easily and either eliminated or chosen as the correct option. By using options from similar frequency levels, the effectiveness of the distractors can be enhanced and the possibility of successful guessing minimized. This could best be implemented by consulting multiple word frequency lists when selecting vocabulary item options.

Our third major suggestion concerns our finding that the assumed vocabulary sizes of the targeted examinees frequently bear little relation to the difficulty of the items included in the vocabulary section. One clear example of the current mismatch can be found in the third level test. The assumed vocabulary size is 2,100 words, yet the third level vocabulary section is primarily testing the first 1,000 high frequency words of English. If *Eikyo* insists on using vocabulary size figures such as the ones reported in Table 2, then they should construct the different levels of their tests to more closely match those figures.

Fourth, we have criticized the proposed vocabulary sizes summarized in Table 2 as being largely divorced from reality. Our recommendation, which we direct at *Eikyo*, *Monbu-kagaku-sho*, and second language researchers in Japan, is that more empirical investigations of the lexical knowledge of Japanese learners at all levels of the formal education system are needed. When *Eikyo* suggests that specific levels of the *Eiken* are appropriate for learners in a particular grade in school, those figures and the amounts of lexical growth associated with them should be based on empirical studies that suggest what amount of lexical growth is challenging yet generally achievable. In this regard, we would like to pose three broad research questions to *Eikyo* and independent researchers suggested by the data in Table 2: (a) For what percentage of Japanese students are the vocabulary size figures accurate? (b) What rate of lexical growth do Japanese students show throughout their junior high school, senior high school, and university studies? (c) To what degree do published figures such as those shown in Table 2 influence Japanese learners? This last question concerns test washback and is related to our belief

that the vocabulary learning goals established by *Monbu-kagaku-sho* for junior and senior high school students are too low.

A New Eiken? A New Eikyo?

Although we believe that the *Eiken* would be improved if the above suggestions were implemented, our recommendations may be analogous to repairing an old car: the repairs help, but what is really needed is a new car. What form might the "new *Eiken*" and the "new *Eikyo*" take? Our list of wishes is long, but we will discuss only three.

First, we would like to see *Eikyo* undertake a reconceptualizion of the entire vocabulary section based on what is currently known about text processing and the second language lexicon on one hand and item response theory (IRT) on the other. As with every other professional language testing organization, *Eikyo* must constantly strive to better understand the underlying construct that they wish to test. At a minimum, this would involve the careful study of recent theories of lexical knowledge and its interaction with text comprehension (e.g., Kintsch, 1998), the second language lexicon (e.g., Pavlenko, 1999), and vocabulary test validation (e.g., Perkins & Linville, 1987). The second base upon which a reconceptualized *Eiken* would rest is statistical theory. As stated earlier, the appropriate use of IRT would permit *Eikyo* to design, refine, and administer the vocabulary section more effectively and circumvent many of the problems we have pointed out.

Our second, and more radical suggestion, is that *Eikyo* should carry out detailed empirical investigations of test functioning that would reveal whether an independent vocabulary section is even needed. A number of studies conducted over the past three decades have consistently shown that vocabulary knowledge is the primary factor underlying reading comprehension. As a result, it may be redundant and therefore inefficient to include both reading comprehension and vocabulary sections on the test. Moreover, current approaches to language testing in general (Chapelle, 1998) and vocabulary testing in particular (Read & Chapelle, 2001) suggest that placing lexical items in rich contexts is the most valid way in which to test examinees' lexical knowledge. In addition, this testing format would overcome the negative washback associated with the vocabulary section of the *Eiken*. Books (e.g., the six volume *Eiken Pass Tanjyukugo*, 1998) and Internet sites (e.g., http:/www19.big.or.jp/~hmnomura/eikenbbs2/eikenbbs2.cgi) dedicated to helping Japanese learners successfully pass the *Eiken* consistently

promote a heavily decontextualized approach to vocabulary learning despite the fact that studies on lexical acquisition (e.g., Prince, 1996) have shown that the overuse of decontextualized vocabulary study can result in learners who cannot break away from a reliance on translation, are unable to exploit the lexicon effectively for production, and have slow and effortful processing of L2 syntax and word identification.

Our final wish is that as a socially responsible corporation, *Eikyo* should be more forthcoming about test functioning. Validation studies need to be undertaken for every section of the *Eiken*, and the results of these studies published so that language testing professionals, teachers, and test takers can examine them in detail. In addition, a test booklet disclosing section and test reliabilities, intercorrelations among test sections, and other quantitative and qualitative data should be made publicly available. One of the best examples of this practice in the field of second language testing is Educational Testing Service, which has long published information about the functioning of the TOEFL test in articles written for the general public and technical research reports that disclose the results of detailed investigations into specific sections of the test (see www.toefl.org for general information and a large number of technical research reports available online). This is all the more important because independent studies (e.g., MacGregor, 1997 and this study) have arrived at the same general conclusion: the *Eiken* has potentially serious reliability and validity problems. In addition to the employees of *Eikyo*, a potentially large number of language testing professionals both inside and outside of Japan could lend their expertise to the development of improved tests.

Conclusion

In this study we have made suggestions for improving the vocabulary section of the *Eiken* based on an analysis of the lexical categories of the item options, sentence stems and reading passages on all seven levels of the *Eiken* administered over a three year period. It is our hope that further studies on the *Eiken* will be undertaken both by independent researchers and by researchers working together with *Eikyo* in order to improve what is unarguably one of the most important proficiency tests in Japan. The Japanese students and adults who take future versions of the test deserve no less.

Acknowledgments

We are grateful to George Curuby for his careful reading of the original manuscript and his thoughtful suggestions and encouragement, Bill Hogue for his insightful and detailed comments, and two anonymous JALT Journal *reviewers whose suggestions resulted in a much-improved paper.*

Tsuyuki Miura is currently working towards a Master of Education degree at Temple University Japan's Osaka campus. Her primary interests are curriculum development and teaching methodology.

David Beglar is an Associate Professor at Temple University Japan. His research interests are vocabulary acquisition and language testing.

Notes

1. We have called *Daimon 1* (section 1 of the written part) the vocabulary section because the majority of the items test knowledge of single words, two-word verbs, or idioms.
2. The fifth level of the *Eiken* does not have a reading test section.
3. Although a large number of factors, such as concreteness, phonological and orthographic regularity, part of speech and pronuncibility influence word difficulty, a considerable amount of research evidence from the field of language testing (e.g., Miller & Lee, 1993; Read, 1988; and Schmitt, Schmitt, & Clapham, 2001) and second language lexical acquisition (e.g., Kirsner, 1994 and Ellis, 1994, 2001) has shown that word frequency is the primary factor underlying lexical difficulty.
4. One reviewer raised the point that other factors, such as the role of cram schools, affect the lexical acquisition of Japanese learners. If *Eikyo* considers these factors, it is their responsibility to describe how such factors are accounted for and how they influence decisions about test construction.

References

Altman, R. (1997). Oral production of vocabulary. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition* (pp. 69-97). New York: Cambridge University Press.

Barrow, J., Nakanishi, Y., & Ishino, H. (1999). Assessing Japanese college students' vocabulary knowledge with a self-checking familiarity survey. *System,* 27, 223-247.

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography,* 6, 253-279.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum.

Carroll, J. B., Davies, P., & Richman, B. (1971). *The American Heritage word frequency book*. New York: American Heritage.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32-70). Cambridge: Cambridge University Press.

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

de Bot, K. (1992). A bilingual production model: Levelt's 'speaking' model adapted. *Applied Linguistics,* 13, 1-24.

de Bot, K., Paribakht, T. S., & Wesche, M. B. (1997). Toward a lexical processing model for the study of second language vocabulary acquisition. *Studies in Second Language Acquisition,* 19, 309-329.

Durgunoglu, A. Y. (1997). Bilingual reading: Its components, development, and other issues. In A. M. B. de Groot & J. F. Kroll (Eds.), *Tutorials in bilingualism* (pp. 255-276). Mahwah, NJ: Lawrence Erlbaum.

Eiken (1998). *Eiken Pass Tanjyukugo.* Tokyo: Obunsha.

Eiken (2001a). *Eiken 1-kyu: Zenmondaishu.* Tokyo: Obunsha.

Eiken (2001b). *Eiken jun-1-kyu: Zenmondaishu.* Tokyo: Obunsha.

Eiken (2001c). *Eiken 4-kyu: Zenmondaishu.* Tokyo: Obunsha.

Eikyo (2001). *Eikyo* Home Page. Retrieved June 6, 2001, from: http://www.eiken.or.jp/

Ellis, N. C. (1994). Vocabulary acquisition: The implicit ins and outs of explicit cognitive mediation. In N. C. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 211-282). San Diego, CA: Academic Press.

Ellis, N. C. (2001). Memory for language. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 33-68). Cambridge: Cambridge University Press.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing,* 4, 139-155.

Haladyna, T. M. (1994). *Developing and validating multiple-choice test items*. Hillsdale, NJ: Lawrence Erlbaum.

Henry, N. (1998). The Eiken test: An investigation? *JALT Journal,* 20, 83-84.

Hirsh, D. (1993). *The vocabulary demands and vocabulary learning opportunities in short novels*. Unpublished master's thesis, Victoria University of Wellington, New Zealand.

Hwang, K. (1989). *Reading newspapers for the improvement of vocabulary and reading skills*. Unpublished master's thesis, Victoria University of Wellington, New Zealand.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: Cambridge University Press.

Kirsner, K. (1994). Second language vocabulary learning: The role of implicit processes. In N. C. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 283-312). San Diego, CA: Academic Press.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P. J. L. Arnaud & H. Bejoint (Eds.), *Vocabulary and applied linguistics* (pp.

126-132). London: MacMillan.
Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics,* 16, 307-322.
Levelt, W. (1993). Language use in normal speakers and its disorders. In G. Blanken, E. Dittman, H. Grimm, J. Marshall, & C. Wallesch (Eds.), *Linguistic disorders and pathologies. An international handbook* (pp. 1-15). Berlin: de Gruyter.
MacGregor, L. (1997). The Eiken test: An investigation. *JALT Journal,* 19, 24-42.
Microsoft Word 2000 [computer software]. (2000). Seattle, WA: Microsoft.
Miller, J. L., & Eimas, P. D. (1995). Speech perception: From signal to word. *Annual Review of Psychology,* 46, 467-492.
Miller, L. T., & Lee, C. J. (1993). Construct validation of the Peabody Picture Vocabulary Test—revised: A structural equation model of the acquisition order of words. *Psychological Assessment,* 5, 438-441.
Monbu-kagaku-sho (2001). *Gakushu shido youryou* [(Foreign language in secondary school:) The Course of Study]. Retrieved June 6, 2001, from: http://www.mext.go.jp/.
Murakami, T. (1972). Eiken ha hyojyun test toshite datou ka [Is the Eiken a valid standardized test?]. *The English Teacher's Magazine*, 20, 41-43.
Nation, I. S. P. (2001). *Learning vocabulary in another language*. New York NY: Cambridge University Press.
Nation, I. S. P., Heatley, A. (1996). VocabProfile, Word, and Range: programs for processing text. [Computer software]. Wellington, NZ: LALS, Victoria University of Wellington.
Pavlenko, A. (1999). New approaches to concepts in bilingual memory. *Bilingualism: Language and Cognition,* 2, 209-230.
Perkins, K., & Linville, S. E. (1987). A construct definition study of a standardized ESL vocabulary test. *Language Testing,* 4, 125-141.
Prince, P. (1996). Second language vocabulary learning: The role of context versus translations as a function of proficiency. *The Modern Language Journal,* 80, 478-493.
Read, J. (1988). Measuring the vocabulary knowledge of second language learners. *RELC Journal,* 19, 12-25.
Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing,* 18, 1-32.
Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behavior of two new versions of the Vocabulary Levels Test. *Language Testing,* 18, 55-88.
VanPatten, B. (1996). *Input processing and grammar instruction.* Norwood, NJ: Ablex.
West, M. (1953). *A general service list of English words.* London: Longman.
Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement.* Chicago: Mesa Press.

(Received May 14, 2002; Revised July 3, 2002)