Reading in a Foreign Language
Volume 15, No. 2, October 2003
ISSN 1539-0578

At what rate do learners learn and retain new vocabulary from reading a graded reader?

Rob Waring
Notre Dame Seishin University

and

Misako Takaki
Okayama Board of Education

Abstract

This study examined the rate at which vocabulary was learned from reading the 400 headword graded reader *A Little Princess*. To ascertain whether words of different frequency of occurrence rates were more likely to be learned and retained or forgotten, 25 words within five bands of differing frequency of occurrence (15 to 18 times to those appearing only once) were selected. The spelling of each word was changed to ensure that each test item was unknown to the 15 intermediate level (or above) female Japanese subjects. Three tests (word-form recognition, prompted meaning recognition and unprompted meaning recognition) were administered immediately after reading, after one week and after a three month delay. The results show that words can be learned incidentally but that most of the words were not learned. More frequent words were more likely to be learned and were more resistant to decay. The data suggest that, on average, the meaning of only one of the 25 items will be remembered after three months, and the meaning of none of the items that were met fewer than eight times will be remembered three months later. The data thus suggest that very little *new* vocabulary is retained from reading one graded reader, and that a massive amount of graded reading is needed to build new vocabulary. It is suggested that the benefits of reading a graded reader should not only be assessed by researching vocabulary gains and retention, but by looking at how graded readers help develop and enrich *already known* vocabulary.
Keywords: guessing vocabulary from context, vocabulary acquisition, graded readers, occurrence rate, vocabulary decay, vocabulary attrition, extensive reading

Introduction

Second language and reading development

It is received wisdom that people learn most of their vocabulary from reading (e.g., Sternberg, 1987). Others take this a little further. Krashen, for example, states that "reading is good for

you. The research supports a stronger conclusion, however. Reading is the only way, the only way we become good readers, develop a good writing style, an adequate vocabulary, advanced grammar, and the only way we become good spellers" (1993:23). In either case, reading is seen to be beneficial for foreign language learning and especially for vocabulary building. There are now quite a number of studies which have looked at how much vocabulary is learned from reading in a foreign language. Examples include, Day, Omura and Hiramatsu (1991); Dupuy and Krashen (1993); Grabe and Stoller (1997); Hayashi (1999); Horst, Cobb and Meara (1998); Mason and Krashen (1997) and Pitts, White and Krashen, (1989) among others.

The general picture that emerges from these studies is that learners do learn vocabulary from their reading but not very much. In many of the studies, typically the gains in scores after reading are only just significant and not much better than random guessing on the tests. Table 1 provides a representative sample of some of the more commonly cited research that has looked at the amount of vocabulary learned from reading in a foreign language. These studies seem to show modest but positive gains in vocabulary acquisition from reading in a foreign language. However, most studies are rather short and the text was often quite difficult.

Table 1: A representative sample of often cited studies of vocabulary growth from reading in a foreign language

Study	Population	Exposure	Materials read	Type of test used	Vocabulary gains
Pitts, White and Krashen (1989) Experiment 1	35 ESL learners	6700 words	2 chapters of *Clockwork Orange* with 123 *nadsat* words	Multiple-choice test	6.4%
Pitts, White and Krashen (1989) Experiment 2	16 ESL learners	6700 words	2 chapters of *Clockwork Orange* plus 2 scenes of the video	Multiple-choice	8.1%
Day, Omura and Hiramatsu (1991)	92 High school EFL learners and 200 university EFL learners	1032 words	Short story	Multiple-choice	1/17 words (5.8%) and 3/17 words (17.6%) (scores above controls who did not see the text)
Dupuy and Krashen (1993)	42 ESL learners	15 page of text	French text plus watched a video	Multiple-choice	6.6 words above controls
Hulstijn (1992)	65 EFL learners	907 words	Advertisement in Dutch	State the meaning of 12 words	1 of 13 words (7.6%)
Horst, Cobb and Meara (1998)	34 EFL learners	21,232 words	A full native speaker novel	Multiple-choice and a word association test	4.62 of 23 MC the words unknown before reading (20.0%) 1.8 of 13 (16%) of the word associations

Several researchers have commented on the inadequacies of the body of research (not only those cited above) purporting to show positive benefits for vocabulary acquisition (among other issues) within foreign language reading. For example, Coady (1997:226), referring to oft-cited research, says that "there appears to be a serious methodological problem with these studies." Nation (1999:124) says that many studies "generally lacked careful control of the research design." Horst, Cobb and Meara (1998) also point out that some of the incidental learning from exposure experiments are "methodologically flawed" (1998: 210). Waring (in preparation), in a meta-analysis of some 28 studies of reading in a foreign language, also found that many of these studies lacked careful control.

Issues in study design

This section will briefly cover some of the research design issues that are common in the body of research into vocabulary gains from reading addressed in this study.

Lack of retention data. It is quite rare in research into vocabulary gains from reading to ask how long these gains will last. The vast majority of the studies on learning from exposure are cross-sectional tests whereby data are gathered only after the reading. To our knowledge, only one second language study has systematically attempted to gather data on how much learning was retained over time (Yamazaki, 1996). One problem with this is that new words will be fresh in the mind for an immediate post test, thus these scores will be higher than if the test were taken some time later. The real test of whether a word has been learned is whether the meaning of a word is retained over time. Following this, we could suggest that cross-sectional studies do not measure the subjects' real amount of learning from reading and therefore we should be wary of claims made only on immediate post test data because the gains would probably have been over-estimated.

Depth of word knowledge. Further, it is rather safe to assume that broadly there are stages, levels or degrees of word knowledge. These could range from knowing only that you have seen or heard the word-form without being able to recall the meaning, to a full understanding of the word and its various nuances and use in a variety of contexts both receptively and productively. In most learning from context research (and most vocabulary research in general), only one test is given which means we can only look at one type of word knowledge gains. This is a shame because it limits us to a one-dimensional picture of what is happening as a result of the treatment. The problem for researchers attempting to create a meta-analysis of the results of many studies is compounded when different researchers use different test types.

Issues in instrumentation. Another problem with many previous studies concerns the types of test used to assess gains in vocabulary from reading in a foreign language. The level of difficulty of the test will have a significant effect on the amount of learning that can be demonstrated and therefore it will affect the gains that can be shown (Waring, 1999; Nation 2001). A test that allows the subject to demonstrate even a small amount of information about the word such as a simple word-form recognition test "have you seen this word before?" will be easier than one that demands that the subject demonstrate more detailed knowledge such as shades of meaning, or the differences from similar words. This implies that researchers who are collecting "learning from context" data should be aware that the type of test that is selected will have a great bearing on the apparent results.

Most studies looking at gains from context when reading extensively have used a multiple-choice test (e.g., Day, Omura and Hiramatsu, 1991; Dupuy and Krashen, 1993; Pitts, White and Krashen, 1989). The complications associated with multiple-choice tests as accurate measures of word knowledge are legion and well documented elsewhere (e.g., Anderson and Freebody, 1981; Meara and Buxton, 1987; Wesche and Paribakht, 1996; Waring, 1999). This type of test is not necessarily the most suitable for assessing how much vocabulary has been learned for several reasons. Firstly, random guessing will affect gain scores. Secondly, multiple-choice tests only assess prompted meaning recognition, not the unprompted meaning recognition one needs for normal reading. Thirdly, multiple-choice tests are notoriously difficult to construct reliably. Therefore, it seems wise when conducting this type of research to use several different tests to determine what types of word knowledge are learned from reading.

The effect of frequency of occurrence on incidental word learning. Few studies (Horst, Cobb and Meara, 1998 is an exception) have looked at what types of words are learned in the reading. Usually a single figure is given that reflects the total number of words learned irregardless of whether the words learned words appeared frequently or not. This would be valuable data because if studies controlled for occurrence rate, we would be able to determine how many times a word needs to be met in the reading for it to be learned in fluent reading.

Research questions

From the above, we can see that there is a need to discover how much vocabulary is learned from reading in a foreign language to answer the following research questions:

A. How many new words are learned from reading a graded reader and retained over time?
B. Are words that appear frequently in the text more likely to be learned than words which appear less frequently?
C. At what rate are the words forgotten (i.e., how many of the words known at a previous test time were not known later)?
D. Do different test formats yield different gain scores?

A study was undertaken to answer these and other questions.

Method

Overview

In this study, 25 words that appeared with different occurrence frequencies were selected from the graded reader, *A Little Princess*, and were changed into substitute words. The subjects read the book and were tested on their recall of the words on three types of test, over three test periods.

Participants

Fifteen 19 to 21 year old Japanese female subjects from a university in Western Japan were the subjects in this experiment. All the participants volunteered to take part. Almost all the subjects were members of the university's English Club. Twelve of them were English majors but all the subjects were at the lower-intermediate level or above. This was determined by their class assignments and ongoing teacher evaluations.

Materials

In order that learning can take place, there should be a good balance of known and unknown words. If the text is too difficult, successful guessing will be hard to achieve. When constructing research of this nature, several points have to be kept in mind for successful

guessing to occur. Firstly, it is well known that until the learner reads at a very high level of text comprehension and text coverage (i.e., one unknown word in 50 or so) little new vocabulary can be guessed from context (Liu and Nation, 1985; Laufer and Sim, 1985; Bensoussan and Laufer, 1984; Hu and Nation, 2000). The optimal rate seems to be between 96 to 99% coverage of known words. Secondly, learners need to meet an unknown word many times before it is learned (Nagy, Herman and Anderson, 1985; Nagy, Anderson and Herman, 1987; Shu, Anderson and Zhang, 1995; Nation 2001: 237). This rate appears to be about ten to fifteen times or more but depends on the word itself and many other factors.

One way to achieve the desired coverage rate would be to use a text the subjects would normally meet in their level studies. Several words could then be selected from this text and tested after reading. This presents two problems. Firstly, we would have to know that the words were indeed unknown prior to reading which would mean a pre-treatment intervention test of the chosen items. However, doing this would highlight the selected words for the subjects, which might compromise the study. The second problem is that we cannot be sure that the non-test items (i.e., the surrounding co-text) were all known. If we selected say 40 items to test, there may be another 40 items the subjects did not know. Eighty unknown items in a text would make the text rather lexically dense in terms of unknown words, thus lowering the known/unknown coverage rate. These two points mean that the text would have to be screened to determine that it met the appropriate rate of 96-99% *before* reading.

The preferred alternative framework we selected was to use a graded reader which would be very easy for the subjects, but introduce some test items into the easier text. For example, a 400 headword graded reader should be easy for intermediate subjects to read and should present no great problems lexically. In this way, we could be reasonably assured that the surrounding co-text for the test items would be known and we can thus look to see what rate of acquisition takes place based solely on the test items.

Twelve graded readers were selected as candidates for this study. Each book was scanned and converted into digital text and then analyzed by computer for the frequency of occurrence of its words. The aim of this was to find a text with a suitable range of occurrence rates which would be easy to read for level subjects. In the end, *A Little Princess* was selected as it met the criteria (see below). *A Little Princess* text is one of Oxford University Press's graded readers Level 1, and has 400 different headwords.

We then had to decide how to present the test items within the surrounding easy text. One choice was to use synonyms for the test items. For example, *curtains* could have been changed to *drapes* and so on. Unfortunately, this raised two potential problems. Firstly, the synonym might already be known, and we would have to introduce a pre-test of the items to ascertain this, which again may prejudice the test. Moreover, it would have caused us problems if some of the subjects knew some words but not others. The second problem with using synonyms is that not all words have synonyms. For example, there are no obvious synonyms for our test items *year* or *snow*. We were not able to select different test items because *A Little Princess* was the only one of the twelve graded readers that were analyzed that had a decent spread of test items according to our occurrence rate criteria (see below). Furthermore, the use of synonyms would

often require a change at the phrase or sentence level to accommodate the relevant collocational and colligational changes, which may compromise the comprehension of the text.

To solve this problem it was decided that we should change the spelling of the test items. These words are henceforth called *substitute words.* It is important to note that these are not *nonsense* words as they are sometimes referred to in the literature. A nonsense word would be one coined for learning that did not previously exist, for example the word for an imaginary six-legged horse, or a yellow and black striped tomato. The use of substitute words refers to the change in spelling of an already known, very common concept. There is no reason why (apart from largely unknown historical reasons) the symbol in English for the large glowing yellow object in our sky should be called *sun* or *blund* or *smalt* or indeed any other letter combination. As words are symbols of meanings, a change in the symbol (its spelling), provided it conforms to normal spelling conventions, has face validity.

The main advantage of a change in spelling is that it allows us to ensure that the words would not be known before reading (although the word's concept would be). A further reason for changing the spelling of the words was to ensure that the words would not be met after the reading in the subjects' normal studies. If they had met the words later, it would have affected their recall on the delayed post-tests.

The substitute words were constructed to look like plausible English words and take on English spelling conventions. For example, we changed *house* into *windle*, *yes* into *yoot*, *name* into *parrow*, *week* into *prink* and so on. These words had already been tested for plausibility by native speakers (Waring, 1999). The substitute words were also checked by five second language learners, who were not part of the experiment, to ensure that that they could pronounce them fairly well so it would not slow their reading. Implausible words, and words difficult to pronounce, were discarded. The substitute words were not highlighted in the text by making them bold, colored or underlined in any way, and were left unmarked for natural reading. We ensured that the words fitted smoothly into the text, which on occasions meant making some words plural by using the unmarked "add an *s*" rule. No definitions or glosses were given at any time.

In order to get reasonably reliable data, we needed to test at least 25 words which the subjects would have to guess from context. And, in order to answer research question B, we needed to select words of differing frequencies of occurrence. However, we also needed to decide what types of words we should select. Nouns and adjectives were chosen because they are generally easier to guess than adverbs. Verbs were not selected because they appear with their inflections which can make it difficult to decide whether the word is "known" and to determine how frequently the word has occurred.

After looking at the occurrences of words in *A Little Princess*, we made six groups of words. These were a group of 5 words appearing one time, a group that included five words appearing 4 to 5 times, and so on for an 8 to 10 group, a 13 to 14 group, the 15 to 20 group and the 21 to 31 group. The sum of the number of occurrences of all the words in the 6 categories that need to be learned was 480. Unfortunately, this made for only 91.9% coverage as *A Little Princess* has only 5872 words (618 types). This meant that the chance of guessing would be low and would be

potentially too difficult for the learners and may affect the study as it did not meet the criteria of 96-99% set earlier. Therefore, we deleted the group of five test items which appeared 21 to 31 times in the text, and we were left with 221 occurrences for the 25 items in the remaining five groups with a "known words" coverage of 96.2% of the running words and a 96.0% coverage by types.

In calculating this figure of 96.2%, we worked on the assumption that all the other words in the book would be known as it was at a reading level far below what learners of their ability should have been capable of. Clearly, this would not be true for all learners, and for all words, but was the best assumption we could make without having the learners read and underline every word they did not know before reading the text. It must be noted of course, that as the subjects read the book, many of the words will be recognized and learned as they read, thus the coverage rate will *increase* as they progress through the book. Following the discussion above, we could safely assume then that the learners would be reading at about an *i*+1 level, (i.e., at the 96% to 99% recommended by Hu and Nation (2000) above for successful guessing from context). The list of words and their substitutes appears in Table 2.

Table 2: The list of English words and their substitute word equivalent and the number of occurrences in the text

English word	Substitute word	Number of occurrences in the text	***Test word group***
house / s	windle / s	17	15-18 Group
yes	yoot	17	
face	mand	18	
mine	brench	18	
money	mear	15	
good	mork	14	13-14 Group
night	cadle	13	
beautiful	smorty	13	
new	tantic	13	
window	bettle	14	
name	parrow	9	8-10 Group
year / s	jurg / s	10	
dead	molden	8	
rich	tring	8	
bread	toker	8	
head	nase	4	4-5 Group
late	bick	4	
week	prink	5	
snow	sind	4	
winter	greal	4	
sun	blund	1	One occurrence Group
special	palk	1	
moment	tance	1	
wrong	vack	1	
world	rimple	1	

In order to answer research question D, we needed to have several tests because we wanted to test different types of word knowledge. We selected three tests, which were: 1) a word-form recognition test; 2) a multiple-choice (prompted recognition) test; and, 3) a meaning by translation (unprompted recognition) test. The three tests were extensively piloted with a group of eight subjects of similar ability and background. These subjects were not part of the main study. The aim of the piloting was to confirm that words were easy pronounceable by Japanese subjects, that the tests contained enough words and the text was not too long and could be read in about one hour at a reasonable reading speed.

The word-form recognition test required the subjects to circle any words they recognized from the text. The test contained the twenty-five substitute words that they had met in the text, plus an additional seventeen distractors to investigate the level of guessing. Piloting had shown that seventeen distractors were enough to provide us with a reliable check for errant guessing. Data

were collected for the number of correct recognitions and the number of false recognitions (i.e., when they circled a substitute word that did not appear in the text.). The test is in Appendix A.

The multiple-choice recognition test is a standard prompted recognition four-choice test with the correct meaning and three distractors. An *I do not know* option was added to allow subjects to indicate when they did not know an item so as to reduce the effect of guessing. The subjects were asked to circle the words they thought were nearest to these words in meaning. These choices are the same part of speech. For example, the substitute word *mand* means *face*. *Face* is a concrete noun, so the four choices are concrete nouns. Care was taken to ensure that the distractors were from different semantic sets to allow for small amounts of knowledge to be demonstrated. The test appears in Appendix B.

The meaning (translation) test presented the twenty-five substitute words in a list. The subjects were asked "What do these words mean? Write the meaning in Japanese." Second and third answers options were given to give them a chance to provide plausible alternative answers to discover whether the subjects could provide a near rather than an exact synonym. The test appears in Appendix C.

The tests were given in a strict order. The word-form recognition test, (the one requiring the least amount of word knowledge) was given first. If the subjects had taken the meaning (translation) test after the multiple-choice recognition test they would have been able to remember some meanings and taken this information to the translation test. Therefore the meaning (translation) test was given second, and the multiple-choice recognition test was given last.

As this research tried to establish how much vocabulary was learned from natural reading, it was decided not to test the words in context. This was because if context had been used (whether contrived, or from the text itself), the subjects may have been able guess what the word meant by working it out at test time. Therefore, we would not know if the word was learned from reading, or guessed at test time.

Procedure

The full text of *A Little Princess*, with the substitute words, was printed for each subject and was put into book form along with a test booklet. The subjects were asked to "read this story as usual and enjoy it." They were told there would be a test after their reading, but no detailed information was given about the test. The subjects were not told that there would be any unfamiliar words in the text. During the test, the researchers checked that the subjects were not turning to look at the test words, and ensured that the subjects did not look at the text when taking the tests. Subjects were not allowed to use a dictionary. As soon as each subject finished the text, she was required to take the tests in the strict order outlined above.

The time taken to read the book was collected for each subject. After reading, each subject was asked if she thought the text was a) easy to read, b) a little difficult, or c) very difficult. Each subject was also interviewed informally after the reading. During the reading, there were some

comments from the subjects, such as "there are a lot of unknown words," but they were only told "please enjoy it."

Seven to ten days after the first test administrations, the subjects took the tests a second time. The aim of this was to determine the rate of forgetting that had occurred from the learning from context. Approximately three months after the reading, the subjects were tested again. This allowed us to test research question C. Both of these test administrations were unannounced. The subjects took the same tests without reading the story again and they never met the words again. At each test administration, the order of presentation of the test items on each test was changed, but the order in which the tests were given was retained. Table 3 has a summary of the types of tests given at each administration.

Table 3: Summary of the procedure

Test Time	**Test Type (in order of presentation)**
	Reading the text
Immediately after reading (n = 15)	1. *word-form recognition test* 2. *meaning (translation) test* 3. *multiple-choice recognition test*
One week later (n = 15)	1. *word-form recognition test* 2. *meaning (translation) test* 3. *multiple-choice recognition test*
3 months later (n = 14)	1. *word-form recognition test* 2. *meaning (translation) test* 3. *multiple-choice recognition test*

Marking

The correct answers on the word-form recognition test were counted as one point. False recognitions were also counted to determine how much guessing was happening. One point was awarded for correct answers and one point was awarded for each of the words "selected in error". These are the "correct" and the "selected in error" scores found in Table 6. On the meaning (translation) test, correct answers were given one point and a word with a similar meaning was given a half point. For example, if the test word's correct answer was *shame*, one point was given, but if the subject supplied *pity*, or *regrettable*, because it is a near synonym, a half point was awarded. A total of only 15 response items were given a half point for the 15 subjects over the three tests over the three test administrations and would not have significantly affected the results overall. On the multiple-choice recognition test only correct scores were counted.

Results

In this section the results of the experiment will be presented. The data for only 14 subjects were collected in the third administration because one subject had left the country by that time.

Time and difficulty data

The average length of time to read the graded reader was 56.3 minutes (*s.d.* 15.5) at an average rate of 104.8 words per minute. Twelve of the 15 subjects took less than 60 minutes and the remainder just over 60 minutes. The average rating of text difficulty was given as 1.5 (*s.d.* 0.6) which shows that the text was not difficult for the subjects. Only one subject rated the text as difficult but her results were about average on all tests.

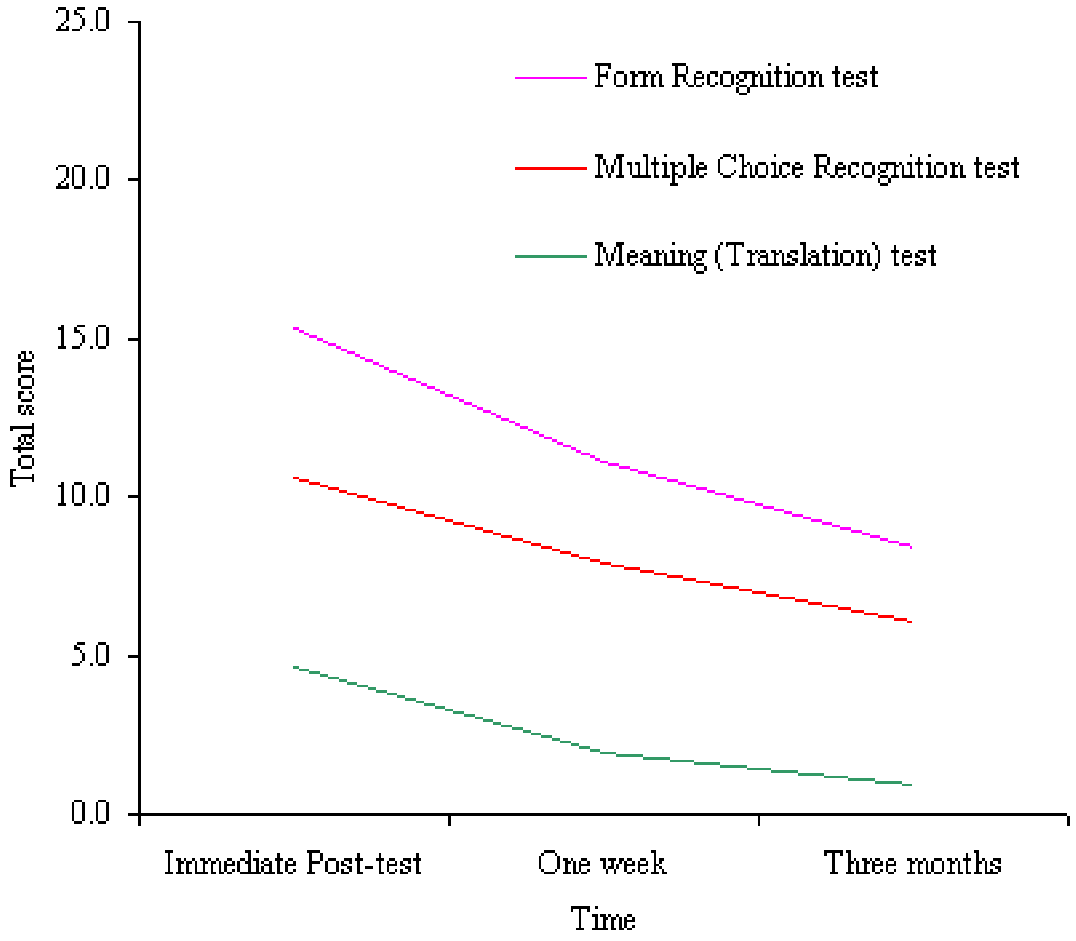
Overall results

The data in Table 4 and Figure 1 show the total scores by test over the three test administrations. The data are shown graphically in Figure 1. This is reported in detail below.

Table 4: The mean scores by test type for the three test administration (Max = 25, n = 15)

	Administration 1 (Immediate post-test)		Administration 2 (One week delay)		Administration 3 (Three months delay) (n = 14)	
	mean	*s.d.*	mean	*s.d.*	mean	*s.d.*
word-form recognition test	15.3	*(3.3)*	11.1	*(5.5)*	8.4	*(4.3)*
multiple-choice recognition test	10.6	*(4.0)*	7.9	*(5.4)*	6.1	*(4.2)*
meaning (translation) test	4.6	*(3.5)*	1.9	*(1.7)*	0.9	*(1.1)*

Figure 1: Mean scores by test over the three test administrations (Max = 25, n = 15)



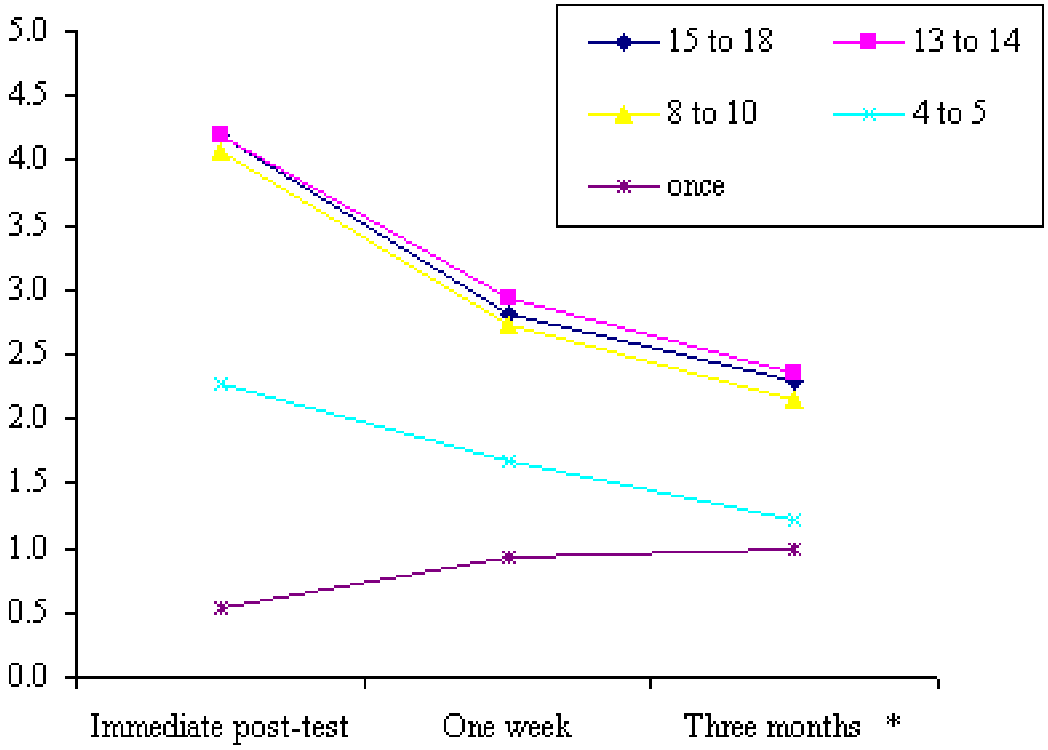
Word-form recognition test data

Table 4 and Figure 1 show that the mean score on the word-form recognition test at the immediate post-test was 15.3 *(s.d. 3.3)* of the 25 items which decreased to 11.1 *(s.d. 5.5)* after one week and to 8.4 *(s.d. 4.4)* after three months. The standard deviations, as a percentage of their means, all increased over time. Table 5 and Figure 2 show the mean scores by occurrence rate of the three test times on the word-form recognition test. The mean scores decreased over time from 4.2 to 2.3 (maximum 5) over the three months for the most frequent items, and the items which occurred only once stayed about the same. The more frequent items had higher recognition rates than the less frequent items.

Table 5: The mean test scores by occurrence rate for the three test administrations on the word-form recognition test (Max = 5, n = 15)

	15 to 18 times	13 to 14 times	8 to 10 times	4 to 5 times	once only
Immediate post-test	4.2	4.2	4.1	2.3	0.5
One week	2.8	2.9	2.7	1.7	0.9
Three months (n=14)	2.3	2.4	2.1	1.2	1.0

Figure 2: Mean scores by occurrence rate on the word-form recognition test over time (Max = 5, n = 15)



* n = 14 at three months

Table 6 and Figure 3 compare the number of word forms which were selected correctly or selected in error. For example, the test item *bettle* represented the English word *window* and appeared in the text. However, the substitute word *stoll* did not represent an English word found in the text and therefore should not have been selected on the word-form recognition test. The columns "selected in error" in Table 6 indicates how often the subjects were guessing. The data by subject are shown in Appendix D.

Table 6 and Figure 3 show that the mean number of incorrect substitute words selected increased from 1.3 items (8.5% of the correct score) on the immediate post-test to 2.8 items (25.2%) after

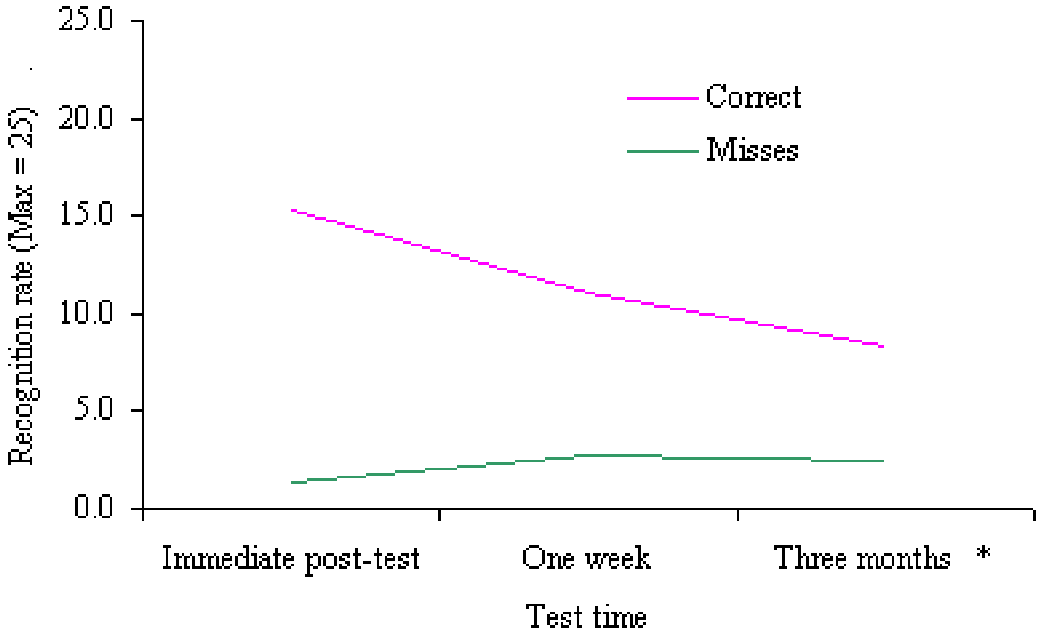
one week and to 2.4 (28.5%) after three months. After correction, the mean scores were 14.0 at the immediate post-test, 8.3 after one week and 6.0 after three months.

Table 6: Mean scores and number of errors on the word-form recognition test for the three test administrations (Max = 25, n = 15)

	Administration 1 (Immediate post-test)		Administration 2 (One week)		Administration 3 (Three months) (n = 14)	
	Correct	Selected in error	Correct	Selected in error	Correct	Selected in error
mean	15.3	1.3	11.1	2.8	8.4	2.4
s.d.	*(3.3)*	*(1.8)*	*(5.5)*	*(3.5)*	*(4.3)*	*(1.6)*
Adjusted means *	14.0		8.3		6.0	

* These were calculated by subtracting the number of incorrect items from the number of correct items.

Figure 3: The rate of correct and missed recognition on the word-form recognition test over time (Max = 25, n = 15)



* n = 14 at three months

Multiple-choice recognition test data

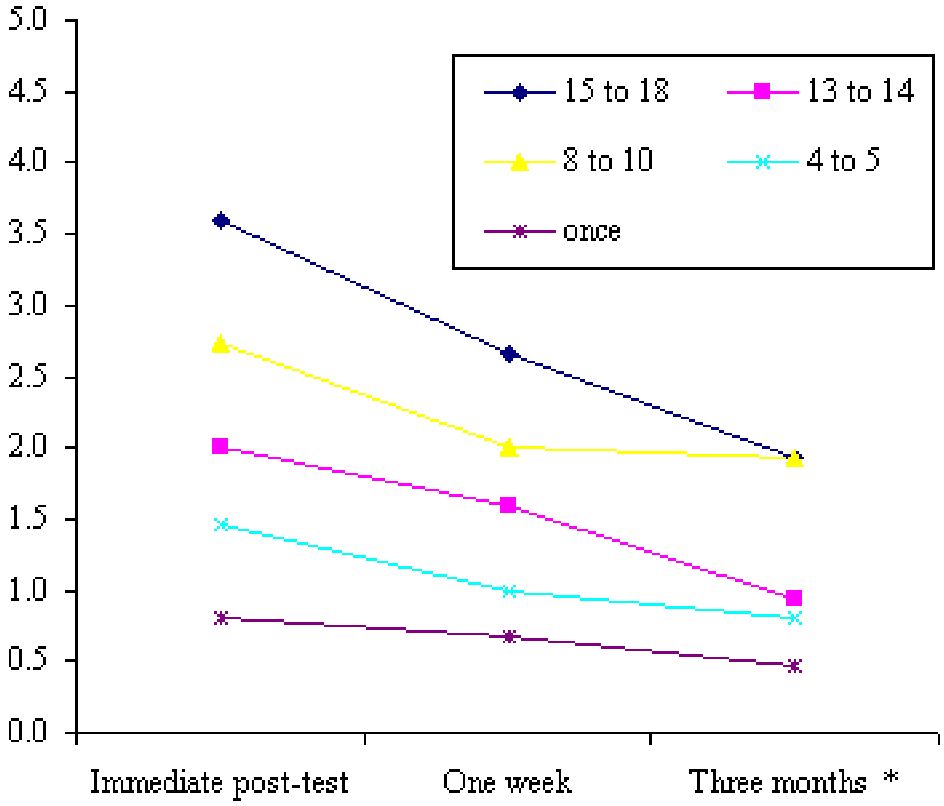
The data for the multiple-choice recognition test are shown in Tables 4 and 7 and Figure 4. Table 7 shows a steady decline in mean scores from the immediate post-test of 10.4 (s.d. 4.0) to 7.9 (s.d. 5.4) after one week and to 6.1 (s.d. 4.2) after three months.

Table 7 and Figure 4 show the mean scores by occurrence rate for the three test administrations. On the immediate post test, there is a drop in scores by occurrence rate from 3.6 of the 5 test items (72%) for the 15-18 group to 0.8 (16%) for words which occurred once. Similar patterns are found after one week (2.7 to 0.7) and after three months (1.9 to 0.5). The data by subject and their standard deviations can be found in Appendix E.

Table 7: Mean scores by occurrence rate for the multiple-choice recognition test over time (Max = 5, n = 15)

	15 to 18 times	13 to 14 times	8 to 10 times	4 to 5 times	once only
Immediate post-test	3.6	2.0	2.7	1.5	0.8
One week	2.7	1.6	2.0	1.0	0.7
Three months (n = 14)	1.9	0.9	1.9	0.8	0.5

Figure 4: Mean scores for the multiple-choice recognition test by occurrence rate over time (Max = 5, n = 15)



* n = 14 at three months

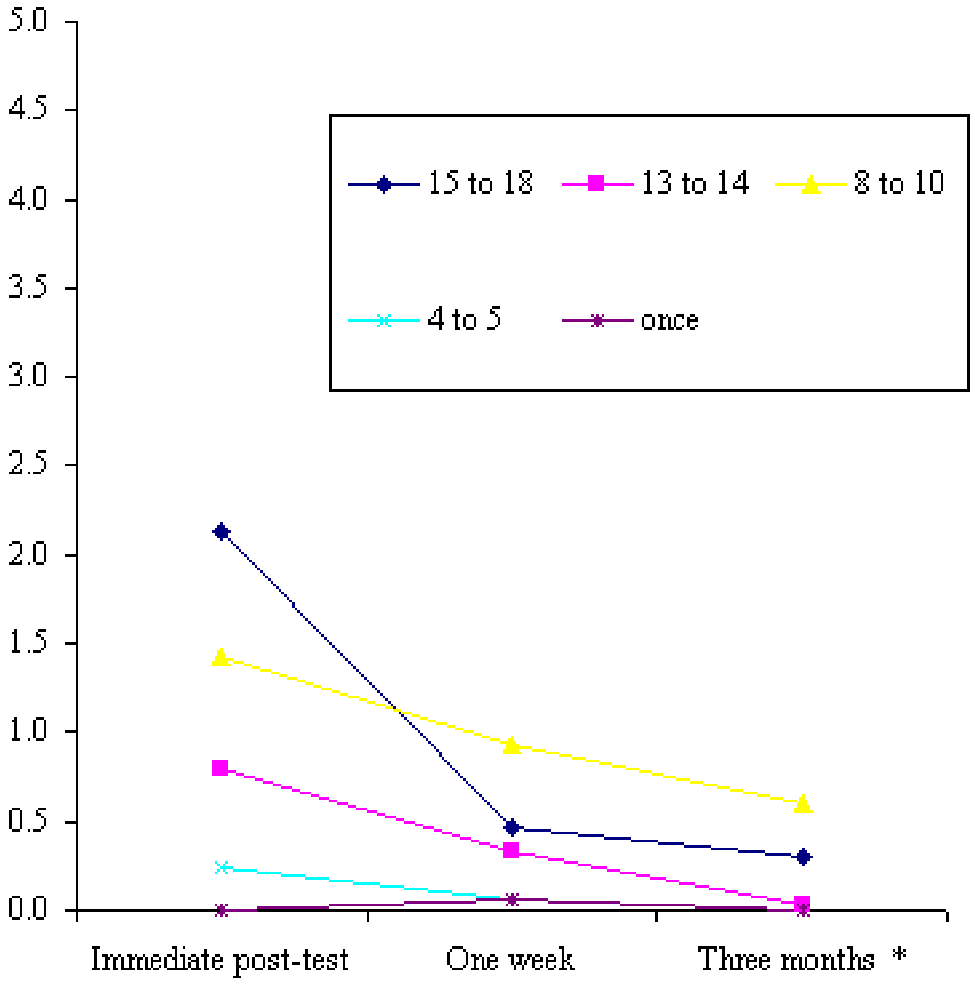
The meaning (translation) test data

The mean scores for the meaning (translation) test by test time are shown in Table 4 and Figure 1. The data show a steady decline in mean scores from 4.6 (s.d. 3.5) of the 25 test items (17.6%) at the immediate post-test to 0.9 (s.d. 1.1) (3.6%) after three months. The data by occurrence rate for the three test administrations on the meaning (translation) test are shown in Table 8 and Figure 5. Table 8 and Figure 5 show an overall decline in mean scores for each of the test times. On the immediate post-test, the items which were met 15-18 times registered a mean score of 2.1 of the 5 items (42%), but after three months this had dropped to 0.3 items (6%). Words which occurred fewer than eight times registered near zero on all tests. The data by subject are shown in Appendix F.

Table 8: Mean scores by occurrence rate for the three test administrations on the meaning (translation) test (Max = 5, n = 15)

	15 to 18 times	13 to 14 times	8 to 10 times	4 to 5 times	once only
Immediate post-test	2.1	0.8	1.4	0.2	0.0
One week delay	0.5	0.3	0.9	0.1	0.1
Three months delay (n = 14)	0.3	0.0	0.6	0.0	0.0

Figure 5: Mean test scores for the meaning (translation) test by occurrence rate (Max = 5, n = 15)



* n = 14 at three months

A repeated measures ANOVA reveals that the three tests produce significantly different scores ($F = 17.11$, $p < .001$) at the immediate post-test. Similar figures between the tests were evident

at the other two test administrations. There are also significant differences between test administration times with each test type. The word-form recognition test produced F = 35.45, p <.001, the multiple-choice test produced F = 22.00, p < .001 and the meaning (translation) test F = 15.41, p = .002.

Discussion

In this section, the results will be discussed. An interpretation of the overall results will be presented first. This will be followed by a detailed analysis of the data from each test and then other related points.

Overall results

The overall results for research question A (How many new words are learned from reading a graded reader over time?) show that the subjects were able to learn some new words from their reading but the vast majority of the new words were not learned. Moreover, the subjects forgot the vast majority of their words they read and learned. On average, the subjects learned (three months retention of the unprompted meaning) one new word from one hour of reading. This is rather a poor rate of return for the effort expended at least as far as the learning of new words is concerned. Research question B asked whether words that appear in the text more frequently were learned before words which appear less frequently. These data suggest that the subjects are able to learn words from context, but they are more successful if they meet the word several times. Research question C asked about the rate at which words are forgotten. The results depend on the test format, but in the main, about half of the words learned in the reading are forgotten after three months. Research question D asked whether different test types yield different test scores, and they did.

Results from the tests

The word-form recognition test assessed the subjects' ability to learn new word forms during their reading. The test did not require them to know the meaning of the word, but just to say if they had recognized the form before. The mean score on the immediate post test of 15.3 of the 25 items dropped to 8.4 after three months (Table 4). This translates into a decay rate of over 45% over three months. On the immediate post test, words met more than 8 times were recognized most of the time, but words met only four or five times were recognized only half of the time (Table 5). However, by three months, nearly half of these gains in word-form recognition were lost. There was also an increase in the number of test items selected in error (Table 6). At the immediate post test 8.5% of the words said to be recognized were incorrect, but by three months this was 28.5%. The standard deviations also increased as a percentage of the means in a similar way. This seems to suggest that the subjects were less and less sure of their knowledge as time passed and they were more willing to guess incorrectly.

The multiple-choice recognition test was a test of prompted meaning recognition. Approximately 40% of the words were remembered at the immediate post test, but after three months, this decreased to one quarter. In other words, of the 10.6 words that were learned, only

6.1 of them were retained, which is a decay rate of over 40%. At the immediate post test 3.6 (72%) of the 5 items at the 15-18 occurrence rate group were recognized when prompted, but this dropped to 1.9 (38%) after three months. Words appearing eight or more times had a 54% chance of being learned but a less than 40% chance of being retained after three months. About one word appearing fewer than 8 times was recognized immediately after reading, dropping to about half of that after three months (Table 6).

On the meaning (translation) test, the subjects had to write a Japanese translation for the test item. At the immediate post test 4.6 (18.4%) of the 25 items were translated correctly, dropping to less than one word (3.6%) after three months. This is remarkable considering the time expended on reading. Only 42% of the words occurring more than 15 times were translated correctly just after the reading. However, within one week this had dropped to 10%, and by 6% after three months. Words met fewer than 15 times had little or no chance of being learned. This suggests that meanings are lost faster than other the types of word knowledge tested here. The rate of forgetting found here largely mirrors that found in memory research that shows that most forgetting occurs soon after learning (Ebbinghaus, 1885 /1913; Baddeley, 1990).

The data from the three tests suggest that while some words were learned, the benefits of the learning were soon lost especially when higher demands were placed on word knowledge (i.e., when meaning was not prompted). Approximately half of the word knowledge gained in the reading was lost after three months. Importantly, the data seem to suggest that there is a critical threshold of the number of times a word has to be met before learning can take place from this type of reading, but it is not clear what this constitutes.

Implications for the type of test used

Research question A asked about the rate at which new items are learned. In natural reading, learners have to be able to recognize words without being prompted (i.e., no definition is given and there is no gloss in the margin). Thus, a new word can be said to be learned only when the subject can connect the form of the word (its spelling) with its meaning. Thus, only the results of the meaning (translation) test can validate whether learning, in terms of unprompted meaning recognition which most closely resembles the knowledge needed for natural reading, has taken place. The recognition of only the form in the word-form recognition test does not constitute learning a new word, but may reflect the recognition of a new word form. Moreover, recognition of the meaning when prompted on the multiple-choice recognition test is not considered to be "learning a new word" in this sense either, but is a demonstration of simple prompted recognition which does not occur in natural reading.

It is our contention, therefore, that a multiple-choice test is not the best way to assess learning new words from context. Similar conclusions were found in Hulstijn (1992). These test types might be better used as pedagogical classroom vocabulary activities than as research tools. It is, of course, recognized that prompted word-form recognition and prompted meaning recognition are important assessment points in the learning of a word. However, they do not show whether a word is known in the sense of being able to make an unprompted form-meaning connection. Having said that, much research in this area has tended to use multiple-choice tests (e.g., Day, Omura and Hiramatsu, 1991; Dupuy and Krashen, 1993; Pitts, White and Krashen, 1989). As

we can see from the data in Table 4, the mean scores on the multiple-choice test are higher than those on the meaning (translation) test. The immediate post-test showed that the multiple-choice test scores were over two times (230%) higher than those of the meaning test scores (10.6 compared to 4.6). However, after three months, this had dramatically increased to nearly seven times higher, or some 677% (6.1 to 0.9). This is rather worrisome for the interpretation of some previous research in this area because this seems to imply that some of the reported vocabulary gains from studies of extensive reading that have used multiple-choice tests may have been over-estimated. The overall results here suggest that a multiple-choice test score will be double that of an unprompted meaning test in terms of what constitutes a known word. This also implies that the vocabulary gains found in the studies mentioned in Table 1 that used multiple-choice tests should be halved to approximate unprompted recognition word knowledge scores.

Number of meetings needed to learn a word

As we saw in the introduction, previous estimates of the number of times it takes to learn a word from reading varied considerably. It is clear from this research that it is very difficult to pin a number on this age-old question. It seems much more complex than a simple single figure. From the results of this experiment, it seems that to have a 50% chance of recognizing a word form again three months later, learners have to meet the word at least eight times. Similar results could be said for prompted recognition. However, for unprompted form-meaning recognition (i.e., word learning) there is only a 10% to 15% chance that the word's meaning will be remembered after three months even if it was met more than 18 times. If the word was met fewer than 5 times, the chance is next to zero. This is rather disappointing because it suggests that we do not learn a lot of new words from our reading even with a 96% coverage rate.

There are several reasons why this might be so. Firstly, the learners are presumably focused on comprehending and enjoying the story rather than on the words themselves. The words were not made explicit by bolding or highlighting the words in any way, as is the case in natural reading. Because of this, the learners are not being forced to notice them and their awareness of the words is not being raised. Some recent research has suggested the noticing of a form is an essential step in word learning (Schmidt, 1990). The question for editors of graded reader series therefore is whether we should highlight words in the text to ensure that certain key words are noticed. It is still an open question whether highlighting the words would help vocabulary acquisition, but many who believe that graded readers should look like books, not like textbooks, would object to this proposal. However, recently some series of graded readers (e.g., Oxford University Press' Dominoes, and Cideb's Black Cat series) have started to do this in their easier level Readers that blend an intensive reading approach with an extensive reading one.

Secondly, the coverage rate in this experiment may have been too high. In other words, little was learned despite the coverage rate being over 96%. It is important to remember that some of the words were learned during the reading, which would have increased the coverage rate and therefore it is not likely that the coverage rate would have been a problem. Moreover, the majority of the subjects rated the book as either easy or not too difficult, thus comprehensibility may have been high and the conditions for successful learning from context seem to have been met.

Thirdly, the reason for low vocabulary rate retention may have simply been that there were too few chances to learn the words. As we have seen, it takes much more than one meeting of a word to learn it from reading. Moreover, even words met more than fifteen times in the text still have only a 40% change of being learned. This seems to suggest that it would take well over 20 or even 30 meetings for most of those words to be learned. If a learner is reading at the 96 to 99% coverage rate as suggested above, and it takes twenty or more meetings with a word to learn it as the data here seem to show, then learners will have to read several hundred or several thousand words in order to learn one new word from their reading. This is a greater amount of reading than is often recommended for foreign language learners reading graded readers. (Nation and Wang [1999] for example suggest a "book a week at their own level of difficulty" as an appropriate rate.) It should also be pointed out that the volume of text that would need to be read to meet an unknown word increases with reading ability level. This is because rarer words are met less frequently and thus more text has to be read to meet an unknown word the required number of times. This also has implications for the amount of text that needs to be read.

Fourthly, the subjects may have found the learning of substitute words more difficult because they may have already known the real English word forms prior to reading. This could potentially mean that when they met a substitute word, they may have been confused because they would have expected the already known word form, not a substitute. To determine if this happened, the subjects were interviewed about the words they met. Many of the subjects said that they were able to guess the meaning of words such as *yoot* (yes) and *molden* (dead) even though they knew the real English word. Further investigation of this revealed that guessing and comprehension were not slowed to any large degree because the real English word was already known. In fact, some of the subjects reported that they assumed it to have a similar meaning to the already known word, which aided their comprehension.

Individual variation

As one would expect from a study of this type, there is considerable individual variation in the gain scores (see Appendices D, E, and F). This variation is also apparent in the increase in standard deviation scores as a percentage of their mean over time. One possible reason would be the speed at which they read. The faster readers could be assumed to have better language ability. Therefore, the data were reanalyzed looking at the differences between subjects who read quickly compared to those who read slowly. The subjects were broken into two groups – those who finished under 60 minutes and those who finished over. On the word-form recognition test's first administration both groups scored exactly the same score of 15.3. On the multiple-choice test, the faster group scored 10.5 while the slower group scored a similar 11.0. On the translation test, the scores were 4.8 and 3.7. Similar profiles were evident at the other two administration times. The similarity of these figures suggests that the speed of reading does not greatly affect the ability to guess new words from context.

Another reason may have been their ability level and this was also investigated. The authors have known each of the subjects quite well, often for several years. Casual observations seem to suggest that the subjects with above average proficiency for the group scored slightly higher on some tests. However, not all those who appear to be lower ability in English did below average on the tests, some in fact scored above average. This suggests that the variation may be a result

of the reading matching their preferred learning style rather than a manifestation of their ability. This of course warrants further investigation.

Interview data

Although no comprehension test was given, the subjects reported that they had understood the main part of the story and had high levels of comprehension. After the immediate post-test we asked all subjects whether the story was easy to read, difficult or very difficult for them, and also we asked what they thought of the story. The four subjects felt it was difficult and could not enjoy it so much. These subjects got low scores, especially on the meaning (translation) test. By contrast, those who rated the book as easy to read generally showed higher gains, but as mentioned above, this was not always a consistent finding.

Difficulty or ease of learning

In Table 4 there is an obvious increase or bump in scores at the 8-10 occurrence rate for the multiple-choice-choice recognition test and the meaning (translation) test. It appears that several words in this group were easy to learn and retain. For example, in Group 8 to 10 the word *jurg/s* appeared with the meaning of "year/s". In *A Little Princess, jurg/s* always appears after numbers which may have contributed to a high learning rate. An analysis of the item by test data shows that *jurgs* is consistently the highest rated word on both tests and at all three test times. *Molden* (dead) rated almost as highly as *jurg* across the two tests and times. These two words account for approximately 65% of all the correct suppliances at the 8-10 level and thus have disproportionately been better learned than other words at that level. At the 15-18 level *yoot* (yes) is also consistently well recognized on all tests and at all times. *Yoot* alone accounts for approximately 58% of the total score for that level on the meaning (translation) test across all test times.

This suggests two things. Firstly, that some words are easier to learn than others, but secondly that the experimental data may have been compromised in some way. If we assume that words like yes will be easier to spot than more conventional words like head, sun and face, then the scores for these levels would have been much higher. Similarly because *jurgs* is easier to guess than a conventional word because of its strong collocation with numbers, it also may have been disproportionately easy to learn. To examine this, *jurgs* and *yoot* were removed from the data analysis in order to more clearly see the rate of acquisition of more conventional content words. The data were reanalyzed on the basis of a re-calculated mean score at each level based on the mean for the remaining words. By doing so, the learning rate on the meaning (translation) test is cut in half (less 47.5%). Thus, while *yoot* and *jurgs* can be said to be important words to some degree, their ease of learning may have over-estimated the results for the learning of the meanings of normal content words. This suggests that lower learning rates can be expected than those presented here.

Also there are some words which lead the subjects to guess incorrectly. This may be because subjects tried to connect the new form to information they already have, whether it is in English or Japanese. The translation data suggest that some subjects misunderstood a word that sounds similar to their first language, here Japanese, or the sounds of words they know in English. For

example, the substitute word *windle* was confused with the English window or wind. The substitute word *brench* was confused with branch, and *cadle* was confused with candle. *Bick* was confused with big because it appears as a similar sounding loan word in Japanese. The word *tance* was confused with the Japanese *tansu* which means wardrobe in English. However, the total number of confusions was relatively small as a percentage of the total suppliances, which reduces these concerns. Moreover, first language and second language interference is common in learning to read in a foreign language anyway, so it does not seem these confusions would have compromised the data abnormally.

Limitations of the study

The study is limited in several ways. Firstly, there are relatively few subjects and a much larger number is more desirable so that the effect of frequency of occurrence rates (where only 5 items were tested in each group) can be more clearly seen. Secondly, only 5872 words were read and to gather more data on the effectiveness of learning vocabulary from reading in a foreign language it would be best to conduct the same study over a number of texts or with a much longer text. Thirdly, texts would be needed that have a good variety of words of various occurrence rates so that all words are comparatively easy to guess and words such as yes and years need not be selected.

Conclusion

The results of this study point to several things. Firstly, the data support the notion that words can be learned incidentally from context. However, these data suggest that few *new* words appear to be learned from this type of reading, and half of those that *are* learned are soon lost. Secondly, the test type affects the gain scores that are shown from the reading. Therefore, researchers should be particularly cautious about selecting multiple-choice tests to validate the learning of vocabulary. Thirdly, previous research that used a multiple-choice test format rather than a translation test most probably has overstated learning gains. Fourthly, those studies that did not have vocabulary retention data almost certainly will have overstated natural learning too. Thus, the results here suggest that studies that had both these elements in their design, appear to have substantially overstated their natural vocabulary gains. This should be borne in mind when interpreting their results.

However, we have to be cautious when saying that very little vocabulary can be learned from the reading. This study only looked at the learning of *new* words from the reading. This study did not attempt to study a myriad of other forms of word knowledge which include lexical access speed gains; the noticing of collocations, colligations or patterns within text; the recognition of new word forms yet to be learned; an increase in the ability to guess from context; a confirmation that a previously guessed word's meaning is probably correct; recognition of new word associations; the raising of the ability to recognize discourse and text structure; an increase in the ability to 'chunk' text; the development of saccadic eye movements and so on, and so on. The jury is still out on these. Research into what effect reading in a foreign language has on these elements of the reading puzzle is welcomed. However, it is our contention that ultimately learners do not learn a lot of *new* words from graded reading, but in fact graded reading helps to

deepen and consolidate *already known* language. The data presented here should not be interpreted as a case against the need for graded reading.

Also it must not be forgotten that the data have been gathered from the reading of only one graded reader. Clearly, graded readers are not supposed to be used as a one-off learning experience with the vague hope that some new words or language will be learned. Unfortunately, all too often this is the case as graded readers are subjected to the 'supplementary' shelf of the teacher's armory. This research points ever more clearly to the need for repeated and consistent exposure to graded readers if words are going to be acquired and especially if the aim is to learn *new* words.

However, there are clear implications for the recycling of vocabulary for the series editors of graded readers. These data suggest three possible courses that a series editor could take for the rate of repetition within a single graded reader, or within a level, or series. Firstly, a graded reader series editor could largely ignore the vocabulary requirements (in terms of volume of text, rate of repetition, and the chance of learning) in favour of letting the story control the vocabulary. A second option would be relevant if the series editor wishes to aid the learning of *new* vocabulary. This could be done by identifying certain target words which may need to be highlighted and repeated over ten to twenty times within a book, or certainly within a level. This seems like an unnecessarily hard constraint on both authors and series editors of graded readers as the naturalness of a book may be distorted by vocabulary requirements. However, at least from the perspective of learning new words from graded reading, considerable care should be taken to ensure that the headwords chosen for a particular reader level are be recycled throughout the level, even if not within a particular title. It is not necessary to ensure that all the headwords appear a minimum number of times with each title, but it seems to be necessary within each level. Not doing so will mean that there is a high chance that new words will be forgotten. At the early levels of graded reader series, this is rather easier to do because the word lists are rather small and self-contained. However, at the later levels, it becomes more difficult as the range of words an author may wish to incorporate into a graded reader will be much wider. The third choice is for the series editor to not be overly concerned with presenting *new* vocabulary but provide a rich input of *already known* vocabulary in various contexts and with a variety of collocations and colligations. The tension among these three choices will remain with us in perpetuity but each has its own implications for the syllabus design of the graded readers.

While this study has given us a few more insights into what kinds of vocabulary are learned from reading, and the rate at which words need to be met in order to learn them, there are still several unanswered questions. Firstly, we are still not clear whether increasing the number of occurrences of target items will lead to higher acquisition rates. If the subjects had met the target items say 25 or 30 times we can presume that more of them would have been learned, but this is not clear. Another unanswered question concerns whether it is due to the nature of graded reading itself (where the focus is on understanding the message rather than on the learning of new vocabulary) that certain words cannot be learned easily in this way. It may be that the type of cognitive effort expended depending on whether the subject is focused on word learning or on the message may be part of the explanation. This fruitful research area may also investigate whether certain types of words are best learned from reading than others, or whether they are best learned out of context. Other questions relate to how much vocabulary is learned by

reading, say, all the titles of one level of a graded reader series, in order to determine just how many titles need to be read to master the vocabulary at that level.

In conclusion, the results of this study seem to support Nation and Wang's (1999) research that recommends a high volume of reading (a book a week at the learner's reading level), or more. If this amount of reading were done, the rather disappointing forgetting rate evident from reading only one book would be reduced to some degree. The data also support Nation and Wang's contention that graded readers might be best used for recycling *already known* vocabulary than for introducing new words. This is because the results of this and other studies suggest that few *new* words seem to be learned from graded reading. As has been mentioned elsewhere, vocabulary growth is not the main aim of graded or extensive reading (e.g., Day and Bamford, 1998, 2002; Waring, 1997; Waring and Takahashi, 2000; Prowse, 2002). Teachers and learners alike would be best advised to be aware of this and not to expect too many new words to be learned from their graded readers. However, learners should be encouraged to read them for the other informational and enjoyable aspects of reading in a foreign language, as well as the many language learning and affective benefits they offer.

Acknowledgement

We thank Nick Bullard of Oxford University Press for his kind permission to allow us to use a digital version of *A Little Princess.*

References

Anderson, R. C. and Freebody, P. (1981). Reading comprehension and the assessment and acquisition of word knowledge. In J. Guthrie (Ed.), *Comprehension and teaching* (pp. 77-117). Newark, Delaware: International Reading Association.

Baddeley, A. (1990). *Human memory*. London: Lawrence Erlbaum and Associates.

Bensoussan, M. & Laufer, B. (1984). Lexical guessing in context in EFL reading comprehension. *Journal of Research in Reading*, 7, 15-32.

Burnett, F. H. retold by J. Bassett. (1991). *A little princess.* Oxford: Oxford University Press.

Coady, J. (1997). L2 vocabulary acquisition through extensive reading. In J. Coady and T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 225-237). Cambridge: Cambridge University Press.

Day, R. & Bamford, J. (1998). *Extensive reading in the second language classroom*. Cambridge: Cambridge University Press.

Day, R & Bamford, J. (2002). Top ten principles for teaching extensive reading. *Reading in a Foreign Language*,14(2), 136-141. http://nflrc.hawaii.edu/rfl/October2002/day/day.html

Day, R., Omura, C., & Hiramatsu, M. (1991). Incidental EFL vocabulary learning and reading. *Reading in a Foreign Language*, 7(2), 541-551.

Dupuy, B. & Krashen, S. (1993). Incidental vocabulary acquisition in French as a foreign language. *Applied Language Learning*, 4, 55-64.

Ebbinghaus, H. (1885/1913). *Über das Gedächtnis*. Leipzig: Dunker. (Translation by H. Ruyer and C. Bussenius, 1913, Memory. New York: Teacher's College, Columbia University.)

Grabe, W. & Stoller, F. (1997). Reading and vocabulary development in a second language: A case study. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 98-122). Cambridge: Cambridge University Press.

Hayashi, K. (1999). Reading strategies and extensive reading in EFL classes. *RELC Journal* 30, 114-132.

Horst, M., Cobb, T., & Meara, P. (1998). Beyond a Clockwork Orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language*, 11(2), 207-223.

Hu, M. & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1), 403-430.

Hulstijn, J. (1992). Retention of inferred and given word meanings: Experiments in incidental vocabulary learning. In P. Arnaud & H. Bejoint (Eds.), *Vocabulary and applied linguistics* (pp. 113-125). London: Macmillan.

Krashen, S. (1993). *The power of reading: Insights from the research*. Englewood, Co.: Libraries Unlimited.

Laufer, B. & Sim, D. (1985). An attempt to measure the threshold of competence for reading comprehension. *Foreign Language Annals*, 18(5), 405-411.

Liu Na & Nation, P. (1985). Factors affecting guessing vocabulary in context. *RELC Journal*, 16, 33-42.

Mason, B. & Krashen, S. (1997). Extensive reading in English as a foreign language. *System*, 25, 91-102.

Meara, P. & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, 4, 142-151.

Nagy, W., Herman, P., & Anderson, R. (1985). Learning words from context. *Reading Research Quarterly*, 20, 233-253.

Nagy, W., Anderson, R., & Herman, P. (1987). Learning word meanings from context during normal reading. *American Educational Research Journal*, 24 263-282.

Nation, P. (1999). *Learning vocabulary in another language*. English Language Institute Occasional Publication 19. Victoria University Wellington, New Zealand.

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

Nation, P. & Wang, M. (1999). Graded readers and vocabulary. *Reading in a Foreign Language*, 12, 355-380.

Pitts, M., White, H., & Krashen, S. (1989). Acquiring second language vocabulary through reading: A replication of the Clockwork Orange study using second language acquirers. *Reading in a Foreign Language*, 5(2), 271-275.

Prowse, P. (2002). Top ten principles for teaching extensive reading -- a response. *Reading in a Foreign Language*, 14(2), 142-145. http://nflrc.hawaii.edu/rfl/October2002/discussion/prowse.html

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11, 129-158.

Shu, H., Anderson, R. & Zhang, Z. (1995). Incidental learning of word meanings while reading: A Chinese and American cross-cultural study. *Reading Research Quarterly*, 30, 76-95.

Sternberg, R. J. (1987). Most vocabulary is learned from context. In M. G. McKeown & M. E. Curtis (Eds.), *The nature of vocabulary acquisition* (pp. 89-105). Hillsdale, NJ: Lawrence Erlbaum Associates.

Waring, R. (Ed.). (1997). Special edition on extensive reading. *The Language Teacher*, 21.

Waring, R. (1999). *Tasks for assessing receptive and productive second language vocabulary*. Ph.D. Thesis. University of Wales, Swansea. http://www1.harenet.ne.jp/~waring/papers/phd/title.html

Waring, R. & Takahashi, S. (2000). *The Oxford University Press guide to the 'why and 'how' of using graded readers*. Tokyo: Oxford University Press.

Waring, R. (in preparation). *Extensive reading in second languages: A critique of the research*. Draft available at http://www1.harenet.ne.jp/~waring/papers/assesser.html

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

Yamazaki, A. (1996). *Vocabulary acquisition through extensive reading*. Unpublished Ph.D. Dissertation, Temple University, Japan.

Appendix A

Word-form recognition test

Circle the words you met in the story.
(物語の中に出てきたと思う単語に丸をして下さい。)

bandle	bettle	tantic
bing	windle	sind
borch	tance	vack
clath	parrow	jurgs
crasty	greal	blund
doce	mear	mork
diggle	brench	yelt
fale	bick	prink
flart	yoot	mand
mave	tring	toker
nutious	cadle	palk
quent	smorty	stoll
sheddle	molden	rimple
smick	nase	speat

Appendix B

Multiple-choice recognition test

Circle the word with the nearest meaning.

（１～２５の単語の意味に近いと思う単語を４つの中からそれぞれ選んで下さい。分からない場合はI do not know を選んで下さい。）

blund	sun	mountain	photo	flower	I do not know
palk	happy	doubtful	special	easy	I do not know
tance	air	moment	love	respect	I do not know
vack	hard	busy	free	wrong	I do not know
rimple	world	mouth	music	club	I do not know
parrow	letter	piano	hand	name	I do not know
jurg/s	year/s	sea/s	bird/s	song/s	I do not know
molden	peaceful	hot	clean	dead	I do not know
tring	rich	dark	pretty	interesting	I do not know
toker	shoe	bread	car	stair	I do not know
mork	red	clever	mad	good	I do not know
cadle	tree	night	college	glass	I do not know
smorty	dry	crazy	beautiful	dirty	I do not know
tantic	new	intelligent	cold	active	I do not know
bettle	cow	window	mud	station	I do not know
nase	bag	head	paper	desk	I do not know
bick	late	ugly	wet	exact	I do not know
prink	box	bike	week	hat	I do not know
sind	snow	pepper	chair	eye	I do not know
greal	paper	tape	game	winter	I do not know
windle	bread	elephant	house	book	I do not know
yoot	yes	oh	why	OK	I do not know
mand	dog	room	face	sky	I do not know
brench	water	mine	help	cake	I do not know
mear	money	pen	cat	file	I do not know

Appendix C

Meaning (translation) test

What do these words mean? Write the meaning in Japanese.

(以下の単語の意味は何だと思いますか？日本語で答えて下さい。もし考えが複数ある場合は、自信のある順に書いて下さい。)

windle	1……	2……	3……
yoot	1……	2……	3……
mand	1……	2……	3……
brench	1……	2……	3……
mear	1……	2……	3……
mork	1……	2……	3……
cadle	1……	2……	3……
smorty	1……	2……	3……
tantic	1……	2……	3……
bettle	1……	2……	3……
parrow	1……	2……	3……
jurgs	1……	2……	3……
molden	1……	2……	3……
tring	1……	2……	3……
toker	1……	2……	3……
nase	1……	2……	3……
bick	1……	2……	3……
prink	1……	2……	3……
sind	1……	2……	3……
greal	1……	2……	3……
blund	1……	2……	3……
palk	1……	2……	3……
tance	1……	2……	3……
vack	1……	2……	3……
rimple	1……	2……	3……

Appendix D

Data by subject for the word-form recognition test (n=5 for each occurrence rate of the 5 bands, making a total of n=25).

	Immediate post-test							0ne week delay							Three months delay						
Subj.	15-18	13-14	8-10	4-5	1	**TOT**	miss	15-18	13-14	8-10	4-5	1	**TOT**	miss	15-18	13-14	8-10	4-5	1	**TOT**	miss
M.I	4	4	4	3	0	15	0	1	1	2	0	0	4	1	0	2	3	0	1	6	2
N.T	5	5	4	3	0	17	0	4	3	1	2	2	12	5	4	1	1	2	2	10	5
C.K	3	5	5	4	0	17	2	3	5	4	2	1	15	2	-	-	-	-	-	-	-
T.O	4	5	4	1	1	15	0	2	3	3	1	0	9	2	3	4	3	0	1	11	2
Y.O	5	5	5	2	3	20	1	5	4	3	3	0	15	3	4	4	4	4	1	17	3
M.T	4	1	3	0	1	9	1	2	2	2	0	0	6	2	2	1	2	0	0	5	4
A.T	5	4	4	1	1	15	1	3	4	2	0	0	9	0	3	4	2	1	0	10	0
M.T	5	5	4	3	0	17	1	3	4	3	0	0	10	1	1	2	0	0	0	3	2
Y.N	5	5	5	3	0	18	0	4	4	5	4	2	19	1	2	3	4	2	1	12	2
R.A	5	4	4	2	1	16	1	2	2	2	1	0	7	3	1	1	1	1	0	4	1
T.S	3	3	2	3	0	11	3	4	4	4	5	4	21	14	4	2	0	0	1	7	3
T.O	5	5	5	5	0	20	2	4	5	4	3	3	19	1	4	4	3	3	2	16	2
K.K	1	5	3	1	0	10	7	2	2	2	2	1	9	6	1	2	3	2	3	11	6
S.A	4	4	4	2	1	15	1	1	0	2	2	0	5	1	2	2	2	2	1	9	3
S.S	5	3	5	1	0	14	0	2	1	2	0	1	6	0	1	1	2	0	1	5	1
mean	4.2	4.2	4.1	2.3	0.5	15.3	1.3	2.8	2.9	2.7	1.7	0.9	11.1	2.8	2.3	2.4	2.1	1.2	1.0	8.4	2.4
s.d.	*1.1*	*1.1*	*0.9*	*1.3*	*0.8*	*3.3*	*1.8*	*1.2*	*1.5*	*1.1*	*1.6*	*1.3*	*5.5*	*3.5*	*1.4*	*1.2*	*1.3*	*1.3*	*0.9*	*4.3*	*1.6*

The *miss* column refers to items selected in error.

Appendix E

Data by subject for the multiple-choice recognition test. (n=5 for each occurrence rate of the 5 bands, making a total of n=25).

	Immediate post-test						0ne week delay						Three months delay					
Subj.	15-18	13-14	8-10	4-5	1	**TOT**	15-18	13-14	8-10	4-5	1	**TOT**	15-18	13-14	8-10	4-5	1	**TOT**
M.I	3	1	2	0	0	6	1	1	3	0	0	5	1	1	1	0	1	4
N.T	1	2	1	1	1	6	1	0	1	0	1	3	2	1	1	1	1	6
C.K	5	4	4	3	1	17	4	4	3	4	2	17	-	-	-	-	-	-
T.O	5	3	4	1	0	13	4	3	5	1	0	13	4	3	5	1	1	14
Y.O	3	3	1	1	1	9	3	2	1	3	0	9	3	1	3	0	1	8
M.T	3	1	4	3	2	13	1	1	2	1	0	5	0	0	2	1	0	3
A.T	4	2	2	0	1	9	4	2	1	1	0	8	3	1	2	1	1	8
M.T	4	2	2	0	0	8	2	1	1	0	0	4	1	0	1	0	0	2
Y.N	4	4	4	3	2	17	5	4	4	1	3	17	3	2	3	2	1	11
R.A	4	3	3	2	2	14	4	1	3	0	0	8	3	2	4	0	0	9
T.S	4	0	2	1	1	8	4	0	1	0	2	7	1	0	2	0	0	3
T.O	4	4	5	3	0	16	4	3	3	4	2	16	2	3	4	4	0	13
K.K	3	0	2	2	1	8	0	2	0	0	0	2	2	0	1	0	1	4
S.A	4	1	2	2	0	9	1	0	0	0	0	1	2	0	0	1	0	3
S.S	3	0	3	0	0	6	2	0	2	0	0	4	2	0	0	1	0	3
mean	3.6	2.0	2.7	1.5	0.8	10.6	2.7	1.6	2.0	1.0	0.7	7.9	1.9	0.9	1.9	0.8	0.5	6.1
s.d.	*1.0*	*1.5*	*1.2*	*1.2*	*0.8*	*4.0*	*1.6*	*1.4*	*1.5*	*1.5*	*1.0*	*5.4*	*1.1*	*1.1*	*1.5*	*1.1*	*0.5*	*4.2*

Appendix F

Data by subject for the meaning (translation) test data. (n=5 for each occurrence rate of the 5 bands, making a total of n=25).

	Immediate post-test						0ne week delay						Three months delay					
Subj.	15-18	13-14	8-10	4-5	1	TOT	15-18	13-14	8-10	4-5	1	TOT	15-18	13-14	8-10	4-5	1	TOT
M.I	2.5	0.5	1	0	0	4	0	0	1	0	0	1	0	0	1	0	0	1
N.T	0.5	0	1	0	0	1.5	0	0	1	0	0	1	0	0	0	0	0	0
C.K	5	3	2.5	1	0	11.5	1	1.5	1	1	0	4.5	-	-	-	-	-	-
T.O	3	1.5	4	0	0	8.5	0.5	2	3	0	0	5.5	1	0	2	0	0	3
Y.O	3.5	0.5	1	0	0	5	0.5	0	1	0	0	1.5	1.5	0	1	0	0	2.5
M.T	0.5	0	1	0	0	1.5	0	0	0	0	0	0	0	0	0	0	0	0
A.T	4	0	1	0	0	5	1	1	0	0	0	2	1	0	0	0	0	1
M.T	1	1	1	0	0	3	0	0	1	0	0	1	0	0	0	0	0	0
Y.N	2	1.5	3	1	0	7.5	2	0	1	0	1	4	0	0	2	0	0	2
R.A	2	0	1	0	0	3	1	0	0	0	0	1	0	0	0	0	0	0
T.S	0	0	0.5	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0
T.O	3	3	3.5	1.5	0	11	0	0.5	3	0	0	3.5	0	0.5	2	0	0	2.5
K.K	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
S.A	3	1	0	0	0	4	1	0	0	0	0	1	1	0	0	0	0	1
S.S	1	0	1	0	0	2	0	0	2	0	0	2	0	0	1	0	0	1
mean	2.1	0.8	1.4	0.2	0.0	4.6	0.5	0.3	0.9	0.1	0.1	1.9	0.3	0.0	0.6	0.0	0.0	0.9
s.d.	*1.5*	*1.0*	*1.2*	*0.5*	*0.0*	*3.5*	*0.6*	*0.6*	*1.0*	*0.3*	*0.3*	*1.7*	*0.5*	*0.1*	*0.8*	*0.0*	*0.0*	*1.1*

About the Authors

Rob Waring is an Associate Professor at Notre Dame Seishin Women's University in Okayama, Japan. His research interests include graded reading and graded listening and second language vocabulary acquisition. On his homepage, he maintains a bibliography of over 30,000 references to work in second language acquisition and various other extensive reading and vocabulary resources.

Misako Takaki teaches English for the Board of Education of Okayama Prefecture, Japan.