the jaltcalljournal

ISSN 1832-4215
Vol. 9, No.1 Pages 79–98

Vocabulary learning through an online computerized flashcard site

Stuart McLean
Temple University, Japan
beso293@hotmail.com

Nicholas Hogg
Osaka Yuhigaoka Gakuen High School, Japan
nicholashogg@hotmail.com

Thomas W. Rush
Kinki University, Japan
tom.rush.w@gmail.com

Regular Paper

Recent research has shown that paired-associate learning is an effective way to acquire second language vocabulary. However, much research in the field has measured small-scale vocabulary learning. This is due to the methods which learners used to conduct paired-associate learning: paper flash cards. This project sought to measure vocabulary growth over one academic year to determine what can be learned through an online flashcard website. Using three groups, one control and two two-hour treatment groups, learners completed vocabulary size pretests at the beginning of the study and posttests after one year of weekly online flashcard site use. The gains were statistically significant, and suggest that weekly flashcard site use may increase vocabulary sizes as measured using the Vocabulary Size Test (Nation & Beglar, 2007).

Introduction

Paired-associate vocabulary learning has long been considered a valuable method to quickly learn large amounts of second language meaning-form relationships (Elgort, 2011). Paired-associate vocabulary can take different forms such as word lists, paper word cards, and electronic word cards. Traditionally, learners do deliberate study of vocabulary using word cards. Word cards usually have the second language word written on one side and the first language meaning on the other. Learners use the cards to learn and review vocabulary very quickly. However, there are some limitations of paper word cards. To address these issues, educators and researchers use computers to assist learners with learning and

reviewing new words. This paper examines the learning achieved when using a computer assisted paired-associate learning program.

Learners may have many language learning goals: grammatical accuracy, clear pronunciation, pragmatic goals, as well as vocabulary goals. Vocabulary is important because it is the foundation of language competence; as Beglar and Hunt (2005) assert "...vocabulary acquisition is a crucial, and in some senses, the central component in successful foreign language acquisition" (p. 7). A large functional vocabulary enables learners to access a broad range of texts, both spoken and written. Ideally, learners become as functional as native speakers with the second language. Goulden, Nation and Read (1990) suggest that educated native speakers know about 20,000 word families, however, this is a very large number for many second language learners. Therefore, the question of what size vocabulary is necessary for successful performance of second language activity becomes important. Reading, writing, listening and speaking are the four main language performance activities which second language learners do. The vocabulary demands of each skill will depend on the purpose and topic of the language use. Therefore the following section will outline the vocabulary requirements to successfully participate in an academic environment using each skill.

Laufer (1989) found that 95% of vocabulary in a written text needs to be known for successful comprehension. Hu and Nation (2000) conclude that around 98% of the words in a text need to be known for both adequate comprehension and also to be able to guess unknown words from context successfully. This 95–98% of vocabulary coverage depends on three factors: 1. the type of text 2. the length of the text and 3. the homogeneity of the text (Nation, 2001). Defining the level of vocabulary control necessary for unassisted reading depends on the above three factors which, in turn, will be decided by the student's language use needs.

For second language learners, writing is used for a broad range of purposes, but one important goal is to communicate in academic settings. Although a range of skills should be under control to communicate effectively in writing, vocabulary mistakes are rated as being the most serious by **EFL** student's university professors (Santos, 1988). Therefore, it seems that one important goal for learners is to have productive control of a large amount of vocabulary.

When listening to colloquial spoken language, it's estimated that a 2000 word vocabulary can cover about 95% of the language (Schonell, Meddleton, and Shaw, 1956). This suggests that second language listeners need to know more than the 2000 most frequent words of English. Nation (2001) concludes that "a much smaller vocabulary is needed for speaking than for writing... We tend to write more about weighty matters than speak about them" (p. 125).

However, all vocabulary items are not equally important to all learners. Due to the large amount of vocabulary in the English language, some words are inherently more useful than other words to different groups of language users. Michael West developed perhaps the most widely used list in his General Service List (1953). West used important criteria to choose the words including frequency, ease of learning, the usefulness of the concept, and the register of the word (pp ix–x). The General Service List is one example of a way to choose vocabulary for learners using well-thought principles. However, there are many such lists for various groups of learners (Coxhead, 2000; Browne and Culligan, 2008).

As there are a variety of ways to choose vocabulary to learn, there are a broad range of learning methods which vary in their efficacy. The primary distinction between learning methods is whether the learning is incidental or deliberate. Incidental vocabulary learning

refers to any learning where the learner meets a word in communicative use without the intention of studying new vocabulary. Reading a graded reader, listening to easy stories, and having a conversation are all examples where incidental learning could happen. Deliberate learning, on the other hand, refers to an activity where the learner is aware of learning as the goal of an activity (Hulstjin, 2003). Studying from word cards, memorizing words lists, using a dictionary to look up new words in a difficult passage are examples of deliberate learning.

Both incidental and deliberate learning are important methods of learning, and it seems that both types of learning occupy different roles in vocabulary acquisition, and are therefore not mutually exclusive. In vocabulary learning, the choice of incidental or deliberate learning depends on the vocabulary learning goals. When learners want to establish an initial form-meaning representation, deliberate learning is both efficient and efficacious for future use. However, when learners want to learn more about items whose meaning they already know, it seems incidental learning is better (Waring & Takaki, 2003; Nation & Wang, 1999). One major criticism of deliberate learning is that deliberate study of language features could not lead to language acquisition (Krashen, 1981). However, Elgort (2011) found that lexical items learned with word cards, a type of intentional decontextualized study, are accessed in a similar manner as items already acquired. This suggests that deliberate decontextualized vocabulary learning is, psychologically, an efficacious learning method. Traditionally, word cards involve the learners choosing words to learn, making word cards, and studying the cards (Nation, 2001). Each of these steps is important for learning and should not be overlooked. However, computers can augment the process considerably by easing the workload for the learner at each stage. The following section will look at these three steps and consider how computers can assist learning at each stage.

Choosing the words

When learners choose words to learn, they should choose words which serve some communicative purpose or which will be useful in the future. Accordingly, learners should have access to the most useful information about word usefulness. Lists such as Michael West's (1953) General Service List, and Averil Coxhead's (2000) Academic Word List are freely available. Although the Academic Word List's usefulness as a representation of truly academic vocabulary has been called into question (Neufeld, Hancioglu, & Eldridge, 2011; Cobb, 2011), it remains the most widely used, and thoughtfully constructed word list of its type. These lists attempt to rank words by their usefulness and frequency according to differing student needs. Once the relative usefulness of words is established, learners need to be careful to not choose words which may cause semantic or formal interference (Higa, 1963; Tinkham, 1993, 1997; Waring, 1997). Ease of learning may also vary depending on the learner's first language sound system. Ellis and Beaton (1993) found that the pronounceability of a word affects the learning burden of productive vocabulary. Ideally, a vocabulary teacher would provide learners with these lists and practice in recognizing potential interference and deciding the learning burden of a word based on its pronounceability. However, such training may be avoided by a well-made computer-assisted vocabulary course. It could account for these principles and could potentially save valuable time and energy for study. Furthermore, a program could sort words based on the short-term memory burden imposed by the L1 sound system, and give learners lists balanced for difficulty.

Making the cards

As in choosing words to learn, learners may benefit from using principles when making word cards. Nation (2001) outlines six guidelines for making cards: encourage recall by using small cards, use L1 translations and understand that the L1 and L2 may vary to some extent, use pictures where possible, keep the cards simple, and use appropriate pack size. Learners should use small cards with the L2 form on one side and the L1 meaning on the other. These cards are convenient and easy to study at any time. Portable devices such as smart phones are not much larger than a pack of word cards, and software is increasingly being developed for these devices.

As the goal of using word cards is to quickly and efficiently learn the form and meaning of second language words, learners should use the most efficient method to convey the meaning. Nation (2001) argues that despite the use of L2 glosses learners usually translate the word into the L1. Laufer and Shmueli (1997) found that learners retained L2 words better when provided with an L1 translation rather than an L2 gloss. Nation (2001) states that learners can be trained to look for the underlying meaning of words, and when possible to use that core meaning on their word cards. Some institutions or teaching contexts may not allow the first language to be used in the classroom. Therefore it may be difficult to use this strategy. However, a well-designed computer-assisted language program might be able to take this into account and provide the core first language translation, thus bypassing explicit use of the L1 by the teacher.

As pictures provide an instantiation of the word, pictures should be used when learning L2 words. Lado, Baldwin and Lobo (1967) found that the most effective presentation of a word involved using a picture, a spoken example and a written example. Drawing pictures takes time though, and learners may not be motivated to draw pictures. However, a well-designed vocabulary program could pre-select appropriate pictures to help learners, a process which would be even more automatic using computer-aided learning.

Nation (2001) argues that learners should keep cards simple, and not expect to learn too much about a word when studying. Learning is incremental, and including too much information may be overwhelming for learners. A paired-associate learning course will be quick and challenging enough for learners to maintain motivation to continue studying.

Finally, Crothers and Suppes (1967) found that learning is optimal when packs are not too large given the difficulty of the words which learners are studying. Difficulty refers to time constraints, whether the task is a recognition or recall task, pronunciation difficulty and the part of speech of the word. A well-made vocabulary course would ideally take these difficulty factors into account with an algorithm and optimize the content and pack size of the words.

Using the cards

Nation (2001) suggests six principles for the use of word cards: use retrieval, learn receptively then learn productively, use adapted sequencing, say the word aloud, put the word in a phrase or sentence, process the words deeply and thoughtfully.

Use retrieval. There are four forms, all of varying difficulties, of retrieval practice; receptive recognition, productive recognition, receptive recall and productive recall (Laufer, Elder, & Congdon, 2004; Nakata, 2011). Receptive recognition describes the process of selecting from

a list the L2 meanings of the target L1 word (Figure 1), while productive recognition is the process of selecting the target L2 form from a list when given the L1 meaning. Receptive recall is the retrieval of the target L1 meaning of the L2 form without a list. Finally, and most difficult is productive recall where the target L2 form is retrieved from the L1 meaning without a list. This kind of retrieval results in better learning than learning without retrieval (Baddeley, 1990; Landauer & Bjork, 1978).

Figure 1. Modes of retrieval

First learn receptively then learn productively. The question of whether productive or receptive retrieval result in greater learning is a complicated one. Substantial research has shown that receptive learning more effectively results in receptive knowledge while productive knowledge is more effectively brought by productive learning when using word pairs (Monderia & Wiersma 2004; Waring, 1997; Webb, 2009). In contrast, Webb (2005) looked at the effectiveness of productive and receptive tasks on vocabulary knowledge. He found that when time on task was the same, receptive learning resulted in both greater receptive and productive word knowledge. However, when time on task was not a factor productive tasks resulted in more productive and receptive word knowledge than receptive tasks. The issue was further complicated by more recent research by Webb (2009). Webb found that productive learning resulted in greater productive knowledge for all tested aspects (orthography, meaning, association, syntax, grammar) and some receptive aspects (orthography and meaning) than receptive learning. As a result, it may be concluded that time permitting, a combination of receptive and productive learning will result in optimal learning conditions. A computer algorithm could ensure that this happens.

In addition to Nation's (2001) suggestion about retrieval type, increasing retrieval effort may also be considered to be useful. The increasing retrieval hypothesis states that the greater the effort made to successfully retrieve information the greater memory of it is enhanced (Pyc & Rawson, 2009). As stated by Nakata (2011), this means that carrying out a number of learning tasks in increasing difficulty is preferable with a well-designed vocabulary program. Ideally, receptive and productive recognition tasks would precede receptive

and productive recall tasks. While learners could theoretically do this with paper flashcards, a lot of effort would be used on organizing such a system, a well-designed computer assisted flashcard program can reduce the amount of effort spent on organization.

Use adaptive sequencing. Adaptive sequencing has been shown to result in improved vocabulary retention (Atkinson, 1972). Adaptive sequencing for word cards often refers to items which are not recalled correctly being practiced more frequently until they are correctly recalled. In other words, words which are not correctly retrieved return to the first stage of the learning program. However, Mondria and Mondria-De Vries (1994) showed that the use of adaptive sequencing with paper flash cards is reliant upon learners being able to judge their own knowledge of a word honestly and accurately. It may be that such a judgment adds to the perceived workload of the task. Computer assisted vocabulary study could significantly reduce this perceived workload by automating such judgments.

Use expanded rehearsal. Expanded rehearsal is a review method whereby the time between reviews increases. It has been shown to be the most effective method of sequencing items (Ellis, 1995; Hulstijin, 2001; Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008). In theory, this is possible without the use of computers, but it is highly impractical owing to the large amount of time that would be taken from learning words to managing the learning of words.

Say the words aloud to yourself. One benefit of computerized word cards is that other media may be included to add to the knowledge of a word, including an audio example sentence. Knowledge of a word can be broken into three parts; form, meaning and use. Each part contains three subparts, and productive and receptive aspects, see Table 1, adapted from Nation, 2001. Although in theory many of the aspects of knowing a word can be added to what would become a very crowded word card, many teachers and learners would agree that, in practice, it is impractical. In addition, film and audio clips cannot be added to a paper word card. Computerized word cards, on the other hand can more effectively and efficiently display the main aspect of word knowledge (Nakata, 2011; Nakata, 2008; Nation 1990) as well as other aspects of word knowledge. Through using various types of media when presenting information learners benefit from dual coding. Dual coding is the result of verbally and visually learned information being stored in different systems (Grabe & Stroller, 2011).

Put the words in a phrase or sentence. Nation (2001) suggests that providing a sentence context for a word "because more information is provided about the word; learners however have to have the ability and motivation to use this information" (p. 309). He concludes that, "the few well-conducted studies do not show a striking superiority for sentence context over isolated words" (p. 309). A well designed-program computerized flashcard program could provide ideal context sentences for learners to have access to, as well as pulling example sentences from a corpus.

Process the words deeply and thoughtfully. Ideally, learners will use strategies to strengthen the form-meaning connection of new words. These strategies include "memory tricks, thoughtful processing, deliberate analysis and elaboration, and conscious connections to previous knowledge" (Nation, 2001 p. 310). This is work that must be done by

Table 1: Types of word knowledge

(from Nation, 2001, p. 27)

Form	spoken	R What does the word sound like? P How is the word pronounced?
	written	R What does the word look like? P How is the word written or spelled?
	word parts	R What word parts are recognizable in the word? P What word parts are needed to express this meaning?
Meaning	form and meaning	R What meaning does this word form signal? P What word form can be used to express this meaning?
	concepts and referents	R What is included in this concept? P What concepts can this item refer to?
	associations	R What other words does this make us think of? P What other words could we use instead of this one?
Use	grammatical functional	R In what patterns does the word occur? P In what patters must we use with this word?
	collocations	R What words or types of words occur with this one? P What words or types of words must we use with this word?
	constraints of use	Where, when and how often would we expect to meet this word? Where, when and how often can we use this word?

Note: R = receptive knowledge, P = productive knowledge

the learner, but in an ideal vocabulary program, conditions could be set up for the initial learning of each word. For example using generative theory to establish the form-meaning connection (Meyers, 2010). A well-designed program could prepare tasks specialized for each word in the program.

Word engine

Word Engine (http://www.wordengine.jp) is an example of a well-designed vocabulary learning program. Browne and Culligan (2008) described the theory and layout of an early version of Word Engine. It operationalizes many of the principles discussed in the above section into a professional and user-friendly layout. Nakata (2011) looked at a host of CALL paired-associate programs with a variety of criteria similar to those presented above. There were two categories of criteria: flashcard creation and editing criteria, and learning criteria. As the criteria are presented, it will be indicated whether or not Word Engine includes the criteria. The flashcard creation and editing criteria were as follows:

- whether the program supports users creating their own flashcards (no)
- whether it has multilingual support (yes)
- whether it includes support for multi-word units (yes)
- what kind of information is included on the card (meaning, context and audio)
- whether learners are supported by the program when they enter their new words with information automatically from a database (no)
- whether learners can create their own sets of flashcards (yes).

Although flashcard creation is not supported by Word Engine, their extensive database includes a wide variety of words. As learners can create new sets of flashcards, it seems that most vocabulary in which learners might be interested would be in Word Engine's database.

The learning criteria were as follows:

- whether the program allows learners to become familiar with the words before the studying begins (yes)
- whether the software has retrieval mode (yes)
- whether the software has receptive recall (yes)
- whether the software has receptive recognition (yes)
- whether the software has productive recall (no)
- whether the software has productive recognition (no)
- whether the software progressively increases the retrieval effort for an item (yes)
- whether the software encourages generative use of an item (yes)
- whether the number of words studied in one session can be controlled and altered (yes)
- whether the software supports adapted sequencing (yes)
- whether the software supports expanded rehearsal (yes).

The most obvious deficiency is the omission of productive tasks in Word Engine. However as stated above Webb (2005) found that receptive tasks were superior when time on task was controlled for. As flashcards are generally only one part of a language course, it makes sense to choose the most efficient learning tasks.

In addition to Nakata's criteria, Word Engine provides a system for monitoring student use. This is important for using the program as part of the student's assessment, as well as for collecting data for the current study. Because of the above reasons, Word Engine was chosen as the intervention in the current study. It seems that Word Engine's efficacy has not been measured using an instrument which measures vocabulary size. Agawa, Black and Herriman (2011) looked at the relationship between Word Engine and **TOEIC** scores, however **TOEIC** measures many other aspects of **L2** proficiency than vocabulary. To our knowledge, since its introduction Word Engine's efficacy has not been measured using the Vocabulary Size Test (**VST**) (Nation & Beglar, 2007).

The research questions which prompted the current study are:

1. Does Word Engine lead to increased scores on the **VST**?
2. Does more Word Engine use lead to increased scores on the **VST**?
3. Do the number of "acquired" vocabulary items correlate with **VST** scores?
4. Do the number of studied vocabulary items correlate with **VST** scores?

Methodology

Design of the study

This study used a quasi-experimental research design that included 12 intact university classes. Four classes acted as the control group conducting only extensive reading for homework while the other eight classes were divided into two treatment groups consisting of four classes each. Vocabulary size was measured at the beginning and end of the academic year. The use of a control group and a pretest facilitated the exploration of the size and direction of selection bias. Using Word Engine's computerized flash card site (**CFCS**)

administration site, researchers continuously monitored the length of time students studied, the number of words studied, and number of words acquired. Within Word Engine the number of words students studied was defined as those words which students had guessed correctly between one and five times. While acquired words are defined as words which students correctly guessed six times consecutively over a period of 90 days.

It was hoped that all groups would conduct 2 hours of weekly homework to control for time on task. With one treatment group conducting two hours of vocabulary study a week and the other treatment group conducting one hour of vocabulary study and one hour of extensive reading a week, while the control group would conduct two hours of ER a week. The treatment groups differed in duration of Word Engine use and inclusion of ER because considering the findings of Waring and Takaki (2003), that ER results in minimal or no vocabulary acquisition, it was hoped to investigate if double the use of Word Engine would result in double the vocabulary size gains. Additionally, greater vocabulary gains experienced following greater Word Engine use while controlling for time task would help provide evidence of the efficaciousness of Word Engine and CFCs. However, unfortunately, the control group was not taught by the researchers, and while it was originally planned that the control group would conduct ER for what was believed to be two hours, this did not turn out to be the case.

Table 2: In-class and out-of-class activities of the groups

Group	In-class activity	Out-of-class activity
Control	Retelling and writing based on ER	One Penguin graded reader a week
Vocabulary and ER	Oral and written discourse activities	One graded reader a week and an hour of Word Engine a week
Vocabulary	Oral and written discourse activities	Two hours of Word Engine a week.

Participants

The 182 participants were first-year Japanese EFL learners attending one of two large, private universities in western Japan. All of the participants had studied English formally for 6 years in secondary school, and were receiving two 90 minute English-language classes a week, one class of productive instruction and one class of receptive instruction at the time of this study. The productive instruction classes in which this study was conducted met once a week for 90 minutes per session for a total of 28 weeks of classes over the two semesters (one academic year).

Control group. The control group (n = 57) consisted of compulsory English students from three social studies department (*hensachi* 50) classes and one law department (*hensachi* 47) class attending compulsory English classes. A department's *hensachi* is calculated by the average of its students' scores on a nationwide exam. Individual students' *hensachi* scores are assigned based on their scores on nationwide ability tests. A *hensachi* score of 50 represents the national average. Outside of class each week students read a Penguin graded reader. The students reading levels were established through the use of Penguin graded reader level test. In class typically students were given five minutes to make notes on the book they read and then completed book reviews on graded readers read outside of class

without looking at the graded reader. Then students orally told their partners about the book read in English. Common errors found in their written reviews were pointed out to the classes. Looking at their written book reviews, students then told a second and third student about books read. This took place in both the first and second semester.

Deliberate vocabulary learning groups. The participants in the intentional vocabulary learning groups came from two universities and were enrolled in compulsory English courses. The vocabulary and extensive reading group (n = 61; hereafter referred to as the vocabulary and ER group) was made up of 4 classes from a single university from the Economic and Business Administration departments with *hensachi* scores of 47 and 48 respectively. The vocabulary and ER group read a single extensive reader a week and completed one hour use of Word Engine a week. Students all read Oxford Bookworm graded readers at level one or the starter level in line with their Oxford graded reader level test. Students were required to complete tests related to the books read on the Moodle Graded Reader Module. If a test was not passed it was not recognized that the book was read and students were required to read a further book that week. Students were strictly monitored and given weekly encouragement to ensure that they completed one book a week or made up for missed reading the following week. Students who did not average a book a week, 12 books per term, were not included in the final data set.

The final treatment group consisted of four classes of university students (n = 64; hereafter referred to as the vocabulary group). One of the classes came from the Business Administration department (*hensachi* 48) of the same university, while the other three classes came from the Economics department (*hensachi* 52) of a second large private university in Western Japan. The vocabulary group completed two hours of Word Engine a week. Weekly monitoring of both the vocabulary group and vocabulary and ER group ensured that students were correctly using Word Engine. Word Engine's admin page shows the amount of time that the students were using the site. The number of words that had been correctly retrieved between one and six times monitored the seriousness with which students were using Word Engine. All eight treatment classes conducted exercises to improve their spoken and written discourse. All 12 classes involved in this study did not conduct deliberate vocabulary learning in class in decontextualized or contextualized form.

The participants in the present study used Word Engine's Basic or Advanced courses. The Basic course consists of the most frequent 5000 words from Lexxica's 860 million-word modern general English corpus and includes all of West's (1953) General Service List and Coxhead's (2000) AWL. The advanced course "teaches 99% of the English vocabulary words that occur in all situations and at all levels of English (from beginner L2 to advanced L1) including: conversations, emails, websites, mass media, novels, classrooms, university textbooks, academic papers, and encyclopedias" (Word Engine, http://www.wordengine.jp) and includes phrasal verbs, chunks, and idioms based on their frequency of occurrence and their relative contribution to coverage of general English.

However, participants did not learn all words in the courses nor in the same sequence after a given point because Word Engine utilizes Item Response Theory (IRT) and Computer Adaptive Tests (CAT) when establishing a user's vocabulary size, and IRT when selecting target items for learners. Word Engine presents words to students depending on individual students needs. Word Engine first establishes a learner's vocabulary size by the use of its online diagnostic tool, V-Check, which identifies the probabilities of a learner knowing a word from Lexxica's 850 million word corpora. V-Check utilizes IRT which assign a score to

the difficulty of an item regardless of the group who took the test. Then through CAT items are selected and administered based upon the response pattern of the test taker, until the desired level of accuracy has been achieved.

Instrument

The VST was developed by Paul Nation to provide a reliable, accurate, and comprehensive measure of second language English learners' written receptive vocabulary size from the first 1000 to the fourteenth 1000-word families of English (Nation and Beglar, 2007). However, in the present study a shorter 8000-word version tests was used. The words included on the Vocabulary Size Test are based on fourteen 1000 BNC word lists developed by Nation (2006) (available at http://www.vuw.ac.nz/lals/staff/paul-nation/nation.aspx). The 80-item version used in the present study simply used the first 80 items which are used to measure vocabulary knowledge of the first eight 1000 BNC word lists. The 14 lists are organized using the notion of word family units.

The word lists used in the selection and sequencing of the test items are not based on the complete 100,000,000 token British National Corpus owing to its formal written nature. For example "items such as *cat* and *hello* occur in the fourth 1000-word list, while formal words such as *civil* and *commission* occur in the first 1000-word list" (Beglar, 2010 p 103). As a result, the first eight 1000-word lists, used in the construction of the instrument, are based on the 10 million token spoken section of the British National Corpus (The spoken corpus lists are available from http://www.vuw.ac.nz/lals/staff/paul-nation/nation.aspx; Nation and Beglar, 2007). The multiple-choice format of the VST allows for a wide range of content to be sampled efficiently, indicate answer selection as efficiently and reliably as possible and demonstrate knowledge of each item. Each test item is first read in context in a short non-defining sentence as is shown by the following example item from the fourth 1000-word level.

vocabulary:

You will need more <vocabulary>.
a. words
b. skill
c. money
d. guns

The four options are written using a restricted vocabulary. Items at the first and second 1000-word frequency levels were written only using words from the first 1000 of West's (1953) General Service List (GSL). The words used to write the four definition options were written using words of higher frequency than the target item being defined. However, for the highest frequency items this was not always possible (e.g., *time* could not be defined except with words of a slightly lower frequency such as hours). For items at the third 1000-word level and above, the defining vocabulary was drawn from the first 2000 words of West's GSL. In the case that a word was not present in the GSL, the frequency of the defining words and the test item were checked using the British National Corpus, and a defining word that was significantly more frequent than the item being defined was used (Nation and Beglar, 2007).

All four defining options are substitutable in the context sentence, and the context

sentences were chosen as they represent the most frequent use of the target item. For instance, the word *instance* most frequently occurs in the phrase *for instance,* so this was used as the sentence context for this target word. Likewise the most commonly used part of speech was selected when presenting target items. Beglar (2010, p 104) believes that "test-takers must have a fairly well-developed idea of the meaning of the word to correctly answer the items because the correct answer and the distractors frequently share elements of meaning." Considering the anticipated vocabulary sizes of the target students an 80-item test designed to measure second language English learners' written receptive vocabulary size from the first 1000 to the eighth 1000-word families of English was administered.

The purpose of administering the **VST** in the study was to measure the current written receptive vocabulary sizes of a sample of Japanese university students. The **VST** is not the most sensitive vocabulary-measuring instrument considering that ten items represent the vocabulary knowledge of 1000 words. However, considering the pedagogically advantageous nature of Word Engine providing individuals with personalized learning programs which contents were unknown by the researchers it was not possible to test for increased word knowledge of the items covered by each individual participant. As a result, the **VST**, which attempts to provide a broad estimate of a test taker's vocabulary knowledge in a relatively short period of time, was utilized. Additionally, while the **VST** may not measure the increasing word knowledge of individual items learned during the use of Word Engine or the control treatment of **ER**, the **VST** does measure vocabulary size in a way that would not systematically affect the treatment groups more so than the control groups or vice versa.

Procedures

During the first week of class in April the pretest **VST** and questionnaire were administered, and students registered with the Moodle Reader Module and Word Engine. The group-specific intervention started from the second week of term. The posttest **VST** was administered in the last week of December. Students who did not complete both the pre and post tests were not included in the present study. Finally, because of the importance of adaptive sequencing (Atkinson, 1972) and expanded rehearsal (Ellis, 1995; Hulstijin, 2001; Cepeda *et al.,* 2008) data for students who did not complete Word Engine for three weeks or more consecutively were not included in the study.

Results

As shown in Table 3, the three groups began the study with different vocabulary sizes, in particular the vocabulary group. A one-way analysis of covariance (**ANCOVA**) was conducted to evaluate the hypothesis that use of the online word card program led to increased **VST** scores. Due to the non-random sampling method, an **ANCOVA** was chosen to try to ameliorate the effects of confounding variables. The independent variable online word-card use had three levels: no use, one hour of use and two hours of use. The dependent variable was the posttest **VST** scores. The covariate was the pretest **VST** scores.

Table 3: Descriptive statistics mean pre and post vocabulary size test scores

Group	*n*	*M* Pretest	*M* Posttest	Prepost Diff.
Control	57	2570.18	2645.61	75.44
Vocabulary + ER	61	2360.66	3508.20	1147.54
Vocabulary	64	3214.06	4321.88	1107.81

As a preliminary, an interaction effect between the covariate (pretest) and the factor (the group) was tested. There was no significant interaction F (2, 176) = 2.18, ***MSE*** = 248973.11, p = .12, η^2 .02. Using these results we assumed homogeneity of slopes and continued with the main **ANCOVA**. The **ANCOVA** was significant F(2, 178) = 108.77, ***MSE*** = 252270.28, p = < .001, η^2= .55. The strength of the relationship between the word-card variable and the dependent variable was very strong, accounting for 55% of the variance of the dependent variable, holding constant the vocabulary size before the word card treatment. Table 4 details these results.

Table 4: Results of the ANCOVA

Source	Type III *SoS*	df	*M2*	*F*	*p*	η^2
Corrected Model	103500000	3	34500000	136.77	.00	.70
Intercept	52850000	1	52850000	209.50	.00	.54
Pretest	18770000	1	18770000	74.42	.00	.30
group	54880000	2	27440000	108.77	.00	.55
Error	44900000	178	252270.28			
Total	2409000000	182				
Corrected Total	148400000	181				

a. R Squared = .697 (Adjusted R Squared = .692)

Table 5: Adjusted posttest vocabulary size test means

			95% CI	
Group	*M*	*SE*	Lower	Upper
Control	2719.00	67.07	2586.84	2851.56
Vocabulary + ER	3680.00	67.34	3547.61	3813.38
Vocabulary	4092.00	68.20	3957.54	4226.70

a. Covariates appearing in the model are evaluated at the following values: 1st = 2726.37.

The means of the posttest **VST** score adjusted for initial differences were in the order of the time spent using the word card program (Table 5). The control group had the smallest adjusted mean (M = 2719.00), vocabulary and **ER** group had a larger adjusted mean (M = 3680.00), and the vocabulary group had the largest adjusted mean (M = 4092.00).

Discussion

Research question 1 asked if the use of Word Engine led to gains in VST scores. The ANCOVA adjusted VST gain scores, displayed in Table 5, show that the vocabulary and ER group and the vocabulary group improved significantly more than the control group. The design of this study permitted a risk of confounding factors, which may have led to misleading ANCOVA results, see Limitations. However, the unadjusted mean VST gain scores of the three groups, shown in Table 3, also support this finding. While potential confounding factors are also a risk when interpreting the unadjusted mean gain scores of the three groups, it seems highly unlikely that those factors would cause such a large gap between the groups using Word Engine and the control group. Thus, it is plausible that a significant portion of the increased gains in VST scores were due to the use of Word Engine. This is also supported by previous research that shows that decontextualized study is an efficient way to increase vocabulary (Fitzpatrick, Al-Qarni, & Meara, 2008; Webb, 2007), and decontextualized vocabulary learning is best achieved through spaced repetition (Ellis, 1995; Kornell, 2009; Mondria & Mondria-De Vries, 1994; Nation, 2001; Cepeda *et al.,* 2008). The relatively small gains in VST scores achieved by the control group is supported by research by Waring and Takaki (2003) that shows that learning vocabulary through extensive reading is a very gradual process.

Research question 2 asked if two 2 hours of weekly use of Word Engine led to greater vocabulary growth than 1 hour of weekly use. The ANCOVA results, shown in Table 5, support this hypothesis, however as mentioned above and explained in the limitation section, these results may be influenced by confounding variables. Further clouding the issue, the raw mean VST gain scores failed to provide support for the claim that two hours of Word Engine led to greater VST score increases than one hour, since the vocabulary and ER group actually achieved slightly higher VST score gains, as shown in Table 3. There are a few possible factors that may explain why this occurred. First, the confounding factors could have influenced the results in either direction, in that they could have accounted for gains or lack of gains for either treatment group. Also, the vocabulary group had a higher mean VST pretest score. That was considered accounted for by using ANCOVA, but if additional support is sought from the unadjusted mean VST gain scores, then the potential influence of the higher starting point of the vocabulary group must also be considered. It seems plausible that since the mean VST pretest score was higher for the vocabulary group, those learners would, on average, have studied lower frequency vocabulary than the vocabulary and ER group when they used Word Engine. It is reasonable to expect that the lower frequency vocabulary studied by the vocabulary group would have been less represented in their other coursework than the words studied by the vocabulary and ER group. With less overlap between Word Engine vocabulary and in-class vocabulary, the vocabulary group could have been afforded fewer encounters with the words they studied than the vocabulary and ER group. A third potential influence is the role that extensive reading played in the VST score gains of the vocabulary and ER group. Yet, considering the small VST score gains of the control group and the relatively slow nature of vocabulary learning from extensive reading (Waring & Takaki, 2003), it seems safe to assume that extensive reading contributed little to the VST score increases of the vocabulary and ER group. A final plausible factor explaining for the lack of greater VST score gains for the vocabulary group, is a decrease in motivation using Word Engine. In a study of second year students at a private Japanese university, Agawa, Black and Herriman (2011), found that almost 40% of the students disliked using

Word Engine. If a similar feeling existed among some of the learners in this study than a second hour of weekly Word Engine use could have decreased their motivation, and led to them going through the motions with the program as opposed to actively engaging with it.

Research question 3 asked if the number of correctly retrieved paired-associates in Word Engine correlated with the increases in **VST** scores. As shown in Table 6, the calculated coefficient was .32, and indicated a medium-strength relationship between these two variables. Highly similar results were found for research question 4.

Table 6: Correctly retrieved paired associate correlation

		Correctly retrieved
Gains	Pearson	.311**
	Sig. (2-tailed)	.00
	n	125

**. Correlation is significant at the 0.01 level (2-tailed).

Research question 4 asked if the number of acquired words in Word Engine correlated with increases in **VST** scores. As shown in Table 7, the calculated coefficient was .31, and indicated a medium-strength relationship between these two variables.

Table 7: Acquired words correlation

		Acquired
Gains	Pearson	.32**
	Sig. (2-tailed)	.00
	N	125

**. Correlation is significant at the 0.01 level (2-tailed).

The similarity of the strength of relationships between the number of correctly retrieved paired-associates and increases in **VST** scores, and the number of acquired words and increases in **VST** scores seems a function of how Word Engine is designed. The words categorized as correctly retrieved pair-associates are any words that have been correctly answered between one and five times without making a mistake. Likewise, acquired words are those words that have been answered six times in a row without making a mistake. There were more words in the correctly retrieved category but the learners' knowledge of these words could be expected to be, on average, less than their knowledge of words in the acquired category. Thus the learners had a higher chance of correctly answering **VST** items if they had acquired the target words in Word Engine but would likely have encountered more target words they had correctly retrieved in the program.

Why greater correlations were not found between these variables was likely the result of discrepancies between the words the learners studied and the 80 target words in the **VST**. Nonetheless, the correlations found add a modicum of support to the claim that increases in the learners' **VST** scores were due to using Word Engine. A more detailed version of the **VST** would likely show a higher correlation with words correctly achieved and acquired in Word Engine and **VST** score gains but would be more time consuming to administer.

Limitations

The results of this study should be carefully considered along with its limitations, which were due to the educational context, the research design and instrumentation. The participants were all native Japanese speakers taking first-year compulsory English courses, and were non-English majors. They were enrolled in programs with entrance *hensachi* ranging from 47 to 60, at private Japanese universities. If similar learning expectations are placed on learners in other contexts then there is a risk of over-generalizing the results. This risk increases with the degree of dissimilarity between other learners and the participants in this study. Learners with different native languages or those acquiring a second language other than English, may be expected to achieve different vocabulary gains using **CFCS** depending on the phonological similarity of their **L1** and the target language. As Ellis and Beaton (1993) showed, the less pronounceable an **L2** word, the greater the learning burden.

A second limitation to the generalizability of the results is the quasi-experimental design of this study. When comparing the control group and vocabulary and **ER** group, this design permits a potential selection threat. Meaning, while their **VST** pretest scores were statistically equivalent, their **VST** posttest differences may have been due to pre-existing differences between these two groups, not the treatment itself. When comparing the nonequivalent-groups of either the control group or the vocabulary and **ER** group with the vocabulary group, the selection threat is made worse because selection bias is also possible; the factors that contributed to the pretest difference may also account for some or all of the posttest gains. Green and Salkind (2008, p. 212) caution that, "…the results of an **ANCOVA** can be misleading for studies with this design." Though it wasn't possible in this study for administrative reasons, random assignment of participants to treatment groups in future studies would greatly improve on the current design.

The students enrolled in the 12 classes examined in this study were also each enrolled in another English course at their universities. It is feasible that the **VST** score gains experienced by the experimental groups were influenced by something that transpired in the context of their other English courses. However, it seems highly unlikely that such an effect would only occur with the eight treatment classes and not at all with the four control classes. More likely, the additional English classes would have contributed a roughly equal amount to the **VST** score gains of all the participants.

An additional design limitation is that the control group was assigned one hour of out-of-class activities, while both treatment groups were assigned two hours. This lack of control for time on task permits the criticism that the **VST** score gains of the vocabulary group and the vocabulary and **ER** group may be due to the second hour of out-of-class activities, not the type of activities. This seems unlikely because the mean **VST** score gains of the vocabulary and **ER** group and the vocabulary group were both more than 14 times those of the control group.

A final design limitation is due to differences in the way extensive reading was implemented in the control group and the vocabulary and **ER** group. In the control group the learners were required to review and discuss the books they read, while in the vocabulary and **ER** group, the learners were required to pass a Moodle Graded Reader Module test each week. The vocabulary and **ER** group students who read books for which they did not pass the Moodle test, were required to read additional books and pass the test. This could have led to unaccounted-for time on task.

A limitation due to instrumentation is a potential ceiling effect in data collection. This

is not a limit of the VST itself, but rather a limitation imposed by time available to administer the test. As described earlier, an 80-item version of the VST was used, thus only testing up to the 8,000-word level. Previous knowledge of vocabulary and gains in vocabulary beyond this level would have escaped notice. The vocabulary group spent the most time using Word Engine, which gave them a greater chance of being exposed to lower frequency vocabulary. The vocabulary group also had the highest mean VST pre and posttest scores. Thus the chance that their vocabulary knowledge was underestimated was greater than the chance of underestimating the vocabulary knowledge of the other two groups. This could be prevented in the future in two ways. One way is using the entire 14,000-word level VST, however this would require more class time and may lead to exaggerated VST scores through more answers being correctly guessed. Another option would be using CFCS that allow the instructors or researchers to limit the word-level to 8,000.

Conclusion

This study looked at the efficacy of an online vocabulary learning program over the course of one academic school year. 12 intact university classes were divided into three groups that were assigned different weekly homework: the control group completed one hour of extensive reading, one treatment group completed one hour of extensive reading and used Word Engine for one hour, and the other treatment group used Word Engine for two hours. The learners also completed the Vocabulary Size Test (Nation & Beglar, 2007) as both a pre and post test. As the groups' mean VST scores on the pre-test scores differed significantly, an ANCOVA was used to examine the vocabulary gains of the three groups. The ANCOVA showed the greatest improvement on VST scores for the two-hour Word Engine group, closely followed by the vocabulary and ER group, with the ER group making relatively little improvement. The difference in VST improvement between the two treatment groups was not as large as expected and there are potential confounding variables described above. Nonetheless, the two treatment groups differed greatly with the control group, suggesting that assigning Word Engine use for out-of-class work contributes relatively quickly to learners' receptive vocabulary knowledge. Future research should use random sampling combined with more carefully controlled intervention to explore the most efficient duration of weekly Word Engine use.

Acknowledgments

We would sincerely like to thank the learners and instructors who assisted us in this project. We would also like to thank Dr. David Beglar for his kind advice and encouragement, Dr. Steven Ross for his concise input which we tried our best to follow, as well as Dr. Paul Nation for his encouragement. Any shortcomings in the methodology or writing are purely our own.

References

Agawa, G., Black, G., & Herriman, M. (2011). Effects of web-based vocabulary training for TOEIC. In A. Stewart (Ed.), *JALT2010 Conference Proceedings*. JALT: Tokyo.

Atkinson, R. C. (1972). Optimizing the learning of a second-language vocabulary. *Journal of Experimental Psychology, 96*(1), 124–129.

Beglar, D., & Hunt, A. (2005). Six principles for teaching foreign language vocabulary: A commentary on Laufer, Meara, and Nation's 'Ten Best ideas'. *The Language Teacher, 29*(07), 7–10.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27*(1), 101–118.

Baddeley, A. (1990). *Human Memory*. London: Laurence Erlbaum Associates.

Browne, C., & Culligan, B. (2008). Combining technology and **IRT** testing to build student knowledge of high frequency vocabulary. *The **JALT CALL** Journal, 4*(2), 3–16

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing Effects in Learning: A Temporal Ridgeline of Optimal Retention. *Psychological Science, 19*(11), 1095–1102.

Cobb, T. (2011). Learning about language and learners through computer programs. *Reading in a Foreign Language, 22*(1), 181–200.

Coxhead, A. (2000). A New Academic Word List. ***TESOL** Quarterly, 34*(2), 213–238.

Crothers. E., & Suppes, P. (1967) *Experiments in Second-Language Learning.* New York: Academic Press.

Ellis, N. C. (1995). Psychology of Foreign Language Vocabulary Acquisition: Implications for **CALL**. *Computer Assisted Language Learning, 8*(2-3), 103–28.

Ellis, N. C., & Beaton, A. (1993). Psycholinguistic determinants of foreign language vocabulary learning. *Language Learning, 43*(4), 559–617.

Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language. *Language Learning 61*(2), pp. 367–413.

Fitzpatrick, T., Al-Qarni, I., & Meara, P. (2008). Intensive vocabulary learning: A case study. *Language Learning Journal, 36*(2), 239–248.

Goulden, R., Nation, I.S.P., & Read, J. (1990) How large can a receptive vocabulary be?, *Applied Linguistics, 11*(4), 341–363.

Grabe, W., & Stoller, F.L. (2011) *Teaching and researching reading.* Harlow: Pearson.

Green, S., & Salkind, N. (2008). *Using **SPSS** for Windows and Macintosh* (5th ed.). Upper Saddle River, **NJ**: Pearson Prentice Hall.

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

Hulstijn, J.H. (2001). Intentional and incidental second language vocabulary learning: Appraisal of elaboration, rehearsal, and automaticity. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 258–286). Cambridge: Cambridge University Press.

Hulstijn, J.H. (2003). Incidental and intentional learning. In C. J. Doughty & M. H. Long (eds.), *The handbook of second language acquisition* (pp. 349–381).

Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology, 23*(9), 1297–1317.

Krashen, (1981). *Second Language Acquisition and Second Language Learning.* Oxford: Pergamon.

Lado, R., Baldwin, B., & Lobo, F. (1967) *Massive vocabulary expansion in a foreign language beyond the basic course: The effects of stimuli, timing and order of presentation.* Washington, **DC: US** Department of Health, Education and Welfare.

Landauer, T.K., & Bjork, R.A. (1978) Optimum rehearsal patterns and name learning. In M.M. Gruneberg, P.E. Morris, & R.N. Sykes (eds.), *Practical Aspects of Memory* (pp. 625–632), London: Academic Press.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension?. In C. Lauren and M. Nordman (eds.), *Special Language: From Humans Thinking to Thinking Machines* (pp. 126–132). Clevedon: Multilingual Matters.

Laufer, B., & Shmueli, K. (1997). Memorizing New Words: Does Teaching Have Anything to Do with It?. ***RELC*** *Journal: A Journal of Language Teaching and Research in Southeast Asia, 28*(1), 89–108.

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: do we need both to measure vocabulary knowledge? *Language Testing, 21*(2), 202–226.

Meyers, P. (2010). *Incidental foreign language vocabulary learning from generative tasks.* Doctoral dissertation: Temple University Japan.

Moodle Graded Reader Module. (n.d.). Moodle Graded Reader Module [Website]. Retrieved January 20 2012, from moodlereader.org/index.html/.

Mondria, J. A., & Wiersma, B. (2004). Receptive, productive, and receptive + productive **L2** vocabulary learning: What difference does it make? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: selection, acquisition, and testing* (pp. 79–100). Amsterdam: John Benjamins.

Mondria, J., & Mondria-De Vries, S. (1994). Efficiently Memorizing Words with the Help of Word Cards and "Hand Computer": Theory and Applications. *System, 22*(1), 47–57.

Nakata, T. (2011). Computer-assisted second language vocabulary learning in a paired associate paradigm: A critical investigation of flashcard software. *Computer Assisted Language Learning, 24*(1), 17–38.

Nakata, T. (2008). English vocabulary learning with word lists, word cards, and computers: Implications from cognitive psychology research for optimal spaced learning. *Re****CALL*** *Journal, 20*(1), 3–20.

Nation, I.S.P. (1990). *Teaching and Learning Vocabulary.* New York: Newbury House.

Nation, I.S.P. (2001). *Learning Vocabulary in Another Language.* Cambridge: Cambridge University Press.

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

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13.

Nation, I.S.P., & Wang, K. (1999). Graded readers and vocabulary. *Reading in a Foreign Language, 12*(2), 355–380.

Neufeld, S., Hancioglu, H., & Eldridge, J. (2011). Beware the Range in **RANGE**, and the Academic in **AWL**. *System, 39*(4), 533–538.

Pyc, M., & Rawson, K. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60*(4), 437–447.

Santos, T. (1988). Professors reactions to the academic writing of nonnative speaking students. ***TESOL*** *Quarterly, 22*(1), 69–90.

Schonell, F., Meddleton, I., & Shaw, B. (1956). *A study of the oral vocabulary of adults.* Brisbane: University of Queensland Press.

Tinkham, T. (1997). The effects of semantic and thematic clustering on the learning of second language vocabulary. *Second Language Research, 13*(2), 138–163.

Waring, R. (1997). A study of receptive and productive learning from word cards. *Studies in Foreign Languages and Literature, 21*(1), 94–114.

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language, 15*(2), 130–163.

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition, 27*(1), 33–52.

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics, 28*(1), 46–65.

Webb, S. (2009). The effects of receptive and productive learning of word pairs on vocabulary knowledge. ***RELC*** *Journal, 40*(3), 360–376.

West, M. (1953). *A General Service List of English Words.* London: Longman, Green & Co.

Word Engine (n.d.). Word Engine [Website]. Retrieved January 17 2012, from http://www.wordengine.jp/.

Author biodata

Stuart McLean is teaching in the Kansai area of Japan while completing his master's degree at Temple. He is interested in vocabulary and extensive reading research, and research design and statistics.

Nicholas Hogg is teaching in the Kansai area of Japan while completing his master's degree at Temple University Japan. His primary professional interests include vocabulary acquisition and language assessment.

Tom Rush is teaching in the Kansai area of Japan. His main research interests include vocabulary learning, reading comprehension, and cognition.