Research and Practice in Technology Enhanced Learning
Vol. 4, No. 2 (2009) 133–168
© World Scientific Publishing Company &
Asia-Pacific Society for Computers in Education

KEY ASPECTS OF COMPUTER ASSISTED VOCABULARY LEARNING (CAVL): COMBINED EFFECTS OF MEDIA, SEQUENCING AND TASK TYPE

SAMUEL R. H. JOSEPH*

Department of Information and Computer Sciences
University of Hawaii at Manoa
1680 East West Road
Honolulu, HI 96822 USA
srjoseph@hawaii.edu
http://lilt.ics.hawai.edu

YUKIKO WATANABE[†], YI-JIUN SHIUNG[‡]
and BOSUN CHOI[§]

Department of Second Language Studies
University of Hawaii at Manoa
1890 East West Road
Honolulu, HI 96822, USA
[†]*yukikow@hawaii.edu*
[‡]*yijiun@hawaii.edu*
[§]*bchoi@hawaii.edu*

CODY ROBBINS

Department of Linguistics
University of Hawaii at Manoa
1890 East West Road, Honolulu, HI 96822, USA
codyr@hawaii.edu

This paper reviews software design guidelines from the field of Computer Assisted Language Learning (CALL) and empirical results from the field of vocabulary acquisition. We categorize the empirical evidence in terms of three key aspects of instructional software for vocabulary teaching: task type, media and scheduling. We describe how design of an online adaptive vocabulary teaching system incorporated each of these aspects. This paper also presents a study that investigates the effectiveness of this system in comparison with a self-paced vocabulary learning system designed without the benefit of optimal task types, media or scheduling. Twenty-six adult English as Second Language (ESL) learners were assigned to a month-long vocabulary learning study that included 250 vocabulary items from the Academic Word List (AWL). Multiple choice and recall vocabulary quizzes were administered at pre-, post-, and delayed posttests. Results showed statistically significant advantages for the participants using the system designed to optimize task-type, media and scheduling. Effect sizes above 1.0 were

*Corresponding author.

observed favoring the experimental condition for both the pre vs. post and pre vs. delayed gain scores and in both the recall and multiple choice question tests. The large effect sizes indicated that the combination of the three CAVL aspects was constructive and that they likely derive benefit from different underlying cognitive mechanisms.

Keywords: CALL; vocabulary learning; expanding rehearsal series; multi-media; scheduling; motivation.

1. Introduction

Nation (2006) suggests that understanding of at least 8,000 to 9,000 English word-families is needed for comprehension of written text and around 6,000 to 7,000 for comprehension of spoken language. Clearly a significant amount of vocabulary is essential for fluent and appropriate language use in various situations, including listening to academic lectures, communicating with others, reading books, and writing essays (Hirsh & Nation, 1992; Hu & Nation, 2000; Laufer, 1989; Nation, 2001). Although many researchers recommend explicit vocabulary instruction (Nation, 1982; Paribakht & Wesche, 1997), vocabulary building is given less priority in second language (L2) classrooms (Grabe & Stroller, 2002), on the assumption that students will learn on their own through natural exposure to language or by necessity. In order to resolve this conflict over the limited amount of class time available, there is a growing interest among second language acquisition researchers and teachers in using technology for vocabulary instruction in and outside the classroom.

Learning a word usually involves learning its spoken and written forms, various meanings, as well as its idiomatic use. For a language learner, knowing a word may initially mean that they should be able to connect a word form with its meaning (Nation, 2001, p. 47). The perceived success of vocabulary learning, therefore, may largely depend on how well the form-meaning mapping process takes place, at least at the early stage of vocabulary acquisition. In this paper we review how existing theories and empirical results inform the design of a technology to support the learning of form-meaning mappings. We also consider how to evaluate a technology designed in this fashion and present the results of a study on the efficacy of a new vocabulary learning system.

The structure of this paper is as follows. First we review various design principles that inform Computer Assisted Vocabulary Learning (CAVL), and develop our own approach to CAVL design. In Section 3, we present an adaptive multimedia CAVL software called iKnow! which follows these design guidelines, and in Section four, we present an experimental method to compare the effectiveness of the iKnow! software with a non-adaptive, non-multimedia CAVL software. In Section 5, we present the results of the experimental study and Section 6 includes the conclusions we can draw based on the results as well as discussion of future research.

2. Computer Assisted Vocabulary Learning (CAVL)

CAVL is a subfield of Computer Assisted Language Learning (CALL) which itself is a subfield of Computer Assisted Learning (CAL[1]). Beatty (2003, p. 133) suggest that defining a model of CALL is problematic because CALL in its broadest sense, is "any process in which a learner uses a computer and, as a result, improves his or her language". Beatty suggests narrowing the scope of any definition and does so by adapting Duncan and Biddle's (1974) classroom teaching model to CALL, and thus indicating the sets of variables (presage, context, process and product) that influence what happens in the "virtual classroom". Beatty does not suggest that his model covers all possible relevant variables, and presents alternative views of learning such as Bloom's (1956) taxonomy of questions (Knowledge, Comprehension, Analysis, Synthesis, Evaluation), and Chandler's (1984) locus of control model (ranging from Behaviorist to Constructivist).

Given the narrower scope of CAVL, one might be tempted to define a model for CAVL along similar lines, but we do not attempt that here. Rather we focus on trying to synthesize the different guidelines that are relevant to the design of CAVL applications. Like Beatty's CALL model, Overbaugh's (1994) guidelines for CAL are adapted from an earlier general learning model; in this case Gagne's (1974) nine events of learning. Overbaugh proposed a detailed list of criteria as to what software designers should consider when developing computer-based learning programs:

1 Instructional Set
  - 1.1 Gaining Attention
  - 1.2 Orienting Activities
  - 1.3 Stimulating Recall of Prior Learning and Supplying Missing Pre-requisites

2 Teaching Strategies
  - 2.1 Presenting Stimuli with Distinctive Features
  - 2.2 Providing Learning Guidance
  - 2.3 Enhancing Retention and Learning Transfer

3 Student Performance
  - 3.1 Eliciting and Assessing Performance
  - 3.2 Providing Feedback
  - 3.3 Other Feedback Issues

4 Other Design Issues
  - 4.1 Learner Control
  - 4.2 Teaching Tools

[1] Also called Computer Based Instruction (CBI), Computer Based Learning (CBL), Computer Aided Instruction (CAI), etc.

However before considering these individual criteria Overbaugh indicates that instructional needs must be identified, broad course objectives formulated, and conditions that may affect knowledge acquisition be considered. These key points are not dissimilar to a set of questions that Nation (2001: 60) suggested teachers should ask about any learning or teaching activity:

1. What is the learning goal of the activity?
2. What psychological conditions does the activity use to help reach the learning goal?
3. What are the observable signs that learning might occur?
4. What are the design features of the activity which set up the conditions for learning?

There questions are presented in the introduction to Nation's chapter on "Teaching and explaining vocabulary", and are answered by Nation in the context of his model of vocabulary learning. Furthermore, these questions follow a similar line to the CALL interface design approach advocated by Plass (1998) who proposed that "*Interface design is the process of selecting interface elements and features based on their ability to deliver support for the cognitive processes involved in the instructional activities facilitated by the application.*" and encouraged designers to follow a three-step process:

(1) Select the instructional activity that supports cognitive processes of the competence or skill to be developed
(2) Select the attributes of the [associated interface] feature
(3) Select the [actual] design feature [and the form of its implementation]

Examples of each of the steps in Plass' approach might be:

(1) Select annotations of vocabulary items (glosses) to support the process of building a vocabulary base from a text and organizing information in short-term memory
(2) Given the use of vocabulary annotations in a written text, make decisions about accessibility, and consider issues such as the extent to which they will obstruct main text
(3) Choose the specific audio or pictorial features for each annotation

We would argue that Plass' first step corresponds to the first two of Nation's questions, and the second and third of Plass' steps correspond to the last of Nation's questions. Nation's inclusion of a focus on the observable signs of learning is important and this also appears in Overbaugh's CAL guidelines. This is not to suggest an omission on Plass' part, more that we can see Overbaugh's finer-grained approach fitting into the latter portions of both Nation and Plass' suggested approaches. Nation and Plass' key set of focused questions/steps correspond roughly to a coarse-grained view of Overbaugh's guidelines.

In our work the Nationian learning goal, Plassian skill, or Overbaughian course objective is that the learner masters the form-meaning mapping for a set of vocabulary items. This could be considered a behaviorist objective, but we do not wish to underplay the importance of other aspects of knowing a word or the constructivist or socio-cultural theories of learning. Clearly vocabulary knowledge needs to be integrated into the learner's existing knowledge in a constructivist sense, and vocabulary is by its nature a component of the socio-cultural complex. However the instructional approach taken here focuses on the individual learner trying to achieve an initial form-meaning mapping.

In designing instructional software for vocabulary teaching we suggest there are three key aspects that need to be considered: (a) what activity should the learner perform to enhance vocabulary learning? (b) through which media should the vocabulary be presented? (c) for how long and how frequently should the learner undertake the activities? These can be viewed as questions concerning task type, media, and scheduling. We propose that they are essential to the design of a CAVL program that promotes form-meaning mapping. In the following section, we incorporate our questions into Plass' (1998) model and summarize the design implications of empirical results from the relevant literature.

2.1. *Selection of instructional activities*

2.1.1. *Task type: What activity should the learner perform to enhance vocabulary learning?*

According to Nation (2001, p. 63), the three important general processes that may lead to a word being remembered are noticing, retrieval, and generation. Noticing involves giving attention to an item, retrieval involves perceiving a word form and recalling its meaning; while generation concerns using a word in a different way from that in its initial presentation. Nation (2001, pp. 72–74) also describes activities to encourage each of these processes, e.g. highlighting or defining important words to increase the chance of them being noticed; repeating stories to promote retrieval; requiring retelling to promote generation.

Nation's three general processes are related to Laufer & Hulstijn's (2001) three levels of processing involvement, need, search and evaluation. Laufer & Hulstijn's linguistic research indicating that learning increases with involvement level has a parallel in the psychological literature: Craik & Tulving's (1975) "Levels of Processing" theory. The implication is that the more deeply a learner processes a word the more likely they are to remember it. Taken in combination with results that indicate that words learnt under more difficult conditions are more likely to be retained (Schneider *et al.*, 2001) it is tempting to think that there is increasing benefit to be gained from more challenging activities. Superficially, generative use is harder than retrieval, which in turn is harder than simply noticing a word. Taking into account that learners are likely to be de-motivated by initial encounters with overly challenging tasks[2] it follows that activities should be designed to become increasingly more difficult in order to maximize the chances of both the learner being motivated to continue and also to get the retention benefits of more complex and challenging activities. This principle of gradually increasing difficulty is exploited in effective computer game design. See for example Ducheneaut, Yee, Nickell & Moore's (2008) study of *World of Warcraft.*

Vocabulary learning activities can also be distinguished along two other orthogonal lines: receptive/productive and recognition/recall (Nation, 2001). Receptive vocabulary knowledge allows the meaning of a word to be understood based on having perceived its form, whereas productive knowledge allows an individual to produce a particular word form having first conceived of its meaning. Many experimental vocabulary learning studies have operationalized the receptive direction as linking a L2 word to its first language (L1) translation, while the productive direction is associated with a L1–L2 mapping. Several experimental studies about relative difficulty between receptive and productive learning have reported that receptive learning is easier than productive learning (Ellis & Beaton, 1993; Griffin, 1992; Stoddard, 1929; Waring, 1997). The other distinction is between recognition, where an existing form is recognized from one or more possibilities, and recall, which requires the test-taker to provide an active response such as writing or saying a word. Studies indicate that recall is generally harder than recognition (Anderson & Bower, 1972; Kintsch, 1970; Mandler *et al.*, 1969; Nist & Olejnik, 1995), however several authors have demonstrated that there are a number of ways to make recognition tests as hard as or harder than recall tests such as to increase the number of choices available (Davis *et al.*, 1961), and the similarity (Bahrick & Bahrick, 1964) or closeness in meaning (Nagy *et al.*, 1985) between the distractors and the correct answers.

In summary, there are three orthogonal dimensions that can be used to describe a vocabulary learning activity, the level of processing dimension (noticing, retrieving, generating), receptive/productive dimension, and recognition/recall dimension. Noticing must arguably be both a receptive and recognition type of activity, but retrieval and generative tasks may involve reception or production and recognition or recall. In general generative tasks will be harder than retrieval; productive harder than receptive and recall harder than recognition, with the addendum that this is not always the case. In choosing vocabulary learning activities a balance will need to be achieved in terms of moving the learner towards more and more challenging tasks while at the same time maintaining their motivation, which will all be a function of their current ability level.

[2] Dornyei (2001: 89) says "there is no better recipe for building someone's confidence than to administer regular doses of success", but "that too easy tasks beat the purpose". Wlodlowski (1999: 155) suggests "Just within reach" is a good rule of thumb to create tasks that challenge learner's capability but provide a high chance of success.

2.2. *Select the attributes of the feature*

2.2.1. *Media: Through which media should the vocabulary be presented?*

There are at least four media types that can be employed during vocabulary instruction: audio, video, still images and text. The key question is which media or combination of media should be used to best promote learning? One of the most commonly cited theories in the literature, Paivio's Dual Coding Theory (1969), has been used as a theoretical basis to explain why visual and verbal information presented together is more effective for retention of a vocabulary item than either alone. Images or videos are generally classified as visual input, while audio and textual information are generally classified as verbal input. Dual Coding theory suggests that visual and verbal information are coded differently in the mind, and that linking them creates more effective pathways to retrieval, thus aiding retention. Table 1 summarizes a number of studies that have compared the effectiveness of different media combinations.

To illustrate, Dubois and Vial (2000) looked at the effectiveness of interrelated visual and verbal information in a CALL environment and found an overall improvement in learning L2 words when the experimental conditions combined visual and verbal media. Jones (2004) used pictorial and written test items and reported positive effects of combined visual and verbal annotations on L2 vocabulary learning when words were presented in a listening comprehension passage. Thompson and Paivio (1994) found that free recall was enhanced when they combined pictures and environmental sounds, as well as seeing additional benefits from increased exposure time to multiple images. However, both Dubois and Vial (2000) and Jones (2004) noted that too much extra information may lead to cognitive overload.

With regard to the comparative effect of still images versus video, Chun & Plass (1996) concluded that images were better than videos and more conducive to incidental vocabulary learning, while Al-Seghayer (2001) showed the opposite results. The conflicting results may be due to the fact that Al-Seghayer asked his participants to define the meaning of the target words in L2, while Chun & Plass (1996)

Table 1. Summary of studies comparing the effectiveness of combined visual and verbal input for vocabulary learning.

Empirical Studies	Experimental Conditions	Comparison
Dubois & Vial, 2000	image alone; image + audio; image + text	visual vs. visual + verbal
Al-Segahyer, 2001; Chun & Plass, 1996	text alone; text + image; text + video	verbal vs. visual + verbal
Jones, 2004	image alone; text alone; image + text	visual vs. verbal vs. visual + verbal
Thompson & Paivio, 1994	image alone; audio alone; image + audio	visual vs. verbal vs. visual + verbal

had their participants provide L1 equivalents to the targeted words. Providing an L1 equivalent is likely to be a less taxing task than asking for a definition in L2.

It seems that not only media, but also timing of media presentation may produce different learning outcomes. More recently, Barcroft (2007) found that with second semester Spanish learners learning concrete nouns, staggering presentation of cue (picture) and response (target word) was beneficial compared to showing cue and response together, which may suggest that limited attention resources are best split between pictorial and textual modes and that presentation timing is also important.

There is also evidence to suggest that presenting multiple media within the verbal modality (e.g. text and audio together) may have benefits for long term retention, as we shall see in Section 2.3.2 below.

2.2.2. *Scheduling: For how long and how frequently should the learner undertake the activities?*

In creating the content of the CALL vocabulary material, it is important to assure not only effective training conditions but also optimal scheduling of the content, in order to maximize long term retention. Many studies have investigated how spacing of repetitions of target vocabulary and repeated presentation of words in an expanding schedule (described below) affect the strength of vocabulary retention. Bloom and Shuell (1981) as well as Dempster (1987) provided support for distributed practice in the language classroom. They concluded that spaced presentation of L2 words, where words are re-presented after long intervals, promotes retrieval over massed presentation, where words are repeatedly presented with only short intervals. Interestingly recent studies by Folse (2006) indicate that repeated exposure may be more important than depth of processing.

Evidence from computer-based studies offers some interesting ideas on sequencing and spacing of materials. Atkinson (1972) reported that students learning L2 words sequenced according to a probabilistic model out-performed learners in a self-directed study condition as long as individual item difficulties were taken into account. Van Bussel (1994) demonstrated something similar but only when learning styles were taken into account. He emphasized the importance of chunking, feedback, elaboration, and sequencing in CALL environment and suggested that adaptive sequencing procedures can be helpful for learners who need or prefer external regulation.

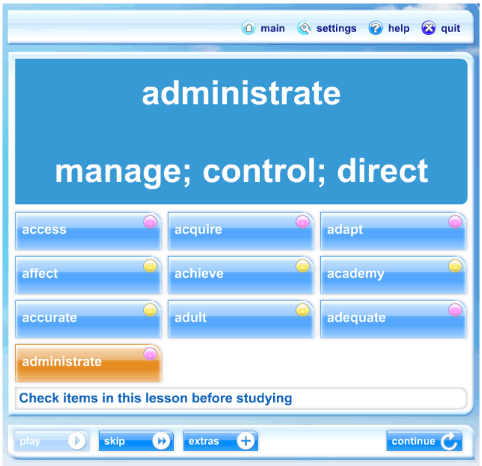
Landauer & Bjork (1978) and others (Pimsleur, 1967) advocated study based on an expanding rehearsal series (ERS) for paired associate learning of L2 vocabulary. An ERS is a pattern of increasing intervals between successive study opportunities designed to maximize long-term retention of an item. Repeated opportunities to encounter an item and retrieve it again appear to be a key factor to consider in system development (Baddeley, 1997: 112). If too much time has passed between previous meeting and the present meeting with the item, the memory of the word built before is unlikely to be sustained. This is why the repetition of new items should occur very soon after they first are studied, before too much forgetting occurs (Nation, 2001: 67).

2.3. *Select the design feature*

In designing CALL platforms and materials, feasibility and technological resource availability are some of the key considerations in creating the most manageable and optimal CALL program. From previous positive findings on the utilization of various types of testing as a learning tool, the use of multiple modes, and the incorporation of expanded rehearsal, an online vocabulary learning software program called iKnow!, which incorporated the above research findings, was developed to provide users an optimal online vocabulary learning environment. The iKnow! program used in this experimental study consists of two components: a preview component, and a study component which includes quiz functionality. Each lesson starts with the preview component and is followed by the study component, which includes different types of quizzes. At the end of each lesson, a learning progress report is presented to the learners. The details of each component are described below.

2.3.1. *Preview component*

In the preview component learners can see an overview of the items they will study in the current session, and can look through the cue/response pairs and example sentences before moving on to the study component when ready. Preview is likely to be beneficial as Bandura & Schunk (1981) showed that setting short term goals related to completing a specific set of material each study session significantly improves learning results, and is advocated by Overbaugh as part of his instructional set, i.e. orienting activities. The preview component, shown in Figure 1, was



administrate

manage; control; direct

Check items in this lesson before studying

Figure 1. An overview of the items in a given study session.

designed to take into account a number of factors, such as how best to first introduce a word, the optimal number of words to study in a session, and the importance of simple definitions and presentation timings.

Noticing new items is seen as a fundamental pre-requisite for learning (Schmidt, 1995). Providing the facility to indicate already known items encourages the learner to consider the items in the current lesson and "notice" the ones that they do not know. Clicking on each item in the presentation screen causes an audio playback of the target word, and presents the word definition.

Choosing the optimal number of words to be studied at one time is complicated as it depends on the difficulty involved in learning the word (Crothers & Suppes, 1967) which is affected by time constraints, type of study exercise, difficulty of pronouncing the target words, learners' proficiency and similarity with known words, (Higa, 1965). Psychological studies demonstrate a list-length effect, i.e. that memory performance will deteriorate as the number of items in a study list increases (Gillund & Shiffrin, 1984), and second language studies indicate a similar effect specifically for vocabulary study (Van Bussell, 1994). Lesson length in iKnow! is generally limited to ten items at a time in order to maximize study effectiveness.

There is evidence that the explanation provided during the first encounter with a word should not be too complicated (Nation, 2001), and studies have shown that a short definition of a word is often more effective than an elaborate explanation (Ellis, 1995; Chaudron, 1982; Laufer & Shmueli, 1997). During iKnow! preview presentations, the target word is presented along with a simple definition, or, alternatively, a translation in the learner's first language. For example in Figure 1, "administrate" would be in L2 and "manage; control; direct" would be in L1. Several studies show that providing first language translation speeds up learning, especially for beginning students (Lado *et al.*, 1967; Mishima, 1967; Laufer & Shmueli, 1997). However for the purposes of our experimental study only L2 was used at the request of the language school the participants were recruited from.

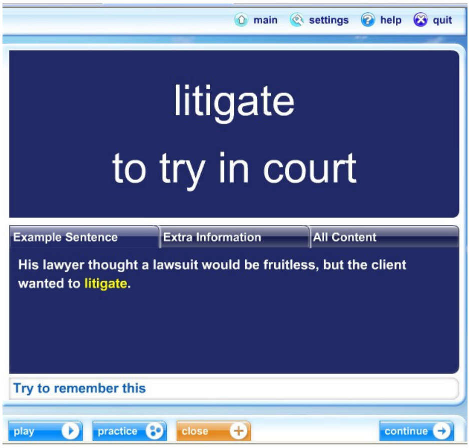
Another issue is how exactly to present a target word and its associated definition/translation. Research indicates that simultaneous presentation of word and definition leads to more effective learning when learners are encountering a word for the very first time (Forlano & Hoffman, 1937; Lado *et al.*, 1967), but that a delayed presentation is more effective for subsequent encounters (Nation, 1982). Thus in preview mode iKnow! presents cue and response together, while as we shall see in the next section, the study component staggers their presentation.

2.3.2. *Study component*

In the study component, the cue word is initially presented alone giving the learner an opportunity to perform receptive written form and meaning retrieval. Barcroft (2007) demonstrated that presenting a cue first in order to give a retrieval opportunity led to an increase in learning performance compared to a control condition where cue and response were presented together. Subsequently, the cue word disappears and the definition/translation is presented, before finally both word and definition/translation are presented together. Nation (1982) indicated this staggered or delayed presentation was most effective for encounters after the first presentation, thus the use of this technique in the study, rather than the preview component.

Word translations or definitions by themselves serve as a helpful initiating event for learning, but prevent the learner from achieving more than a shallow level of word knowledge (Nagy, 1997). A great deal of research has shown that when learners study definitions alone their ability to comprehend text containing the target words does not improve (Graves, 1986; Stahl & Fairbanks, 1986). Solutions include providing opportunities for more elaborate processing, such as example sentences (Mondria & Wit-de Boer, 1991) or multimedia content (Chun & Plass, 1996, 1997). The study component also provides the learner access to much more detailed information about the target word, such as an example sentence and additional information about the word, through the multiple tabs that can be seen at the middle of Figure 2. Providing an example sentence for each item is valuable because it allows the item to further develop associations and collocations that are necessary to understand the word and improve retention (Mondria & Wit-de Boer, 1991).

In addition, audio recordings of the sentence and spelling practice (see Figure 3) support the development of sound form mapping and example sentences support learning of receptive grammar and collocation. Providing audio content is known to help second language vocabulary acquisition (Adepoju & Elliott, 1997; Elliott & Adepoju, 1997). Audio presentation helps learners with pronunciation, and research indicates that improved retention can be achieved if learners say target words aloud (Seibert, 1927). Repeatedly pronouncing a word ensures that phonologically coded information will be retained in the phonological loop, thus increasing the chance



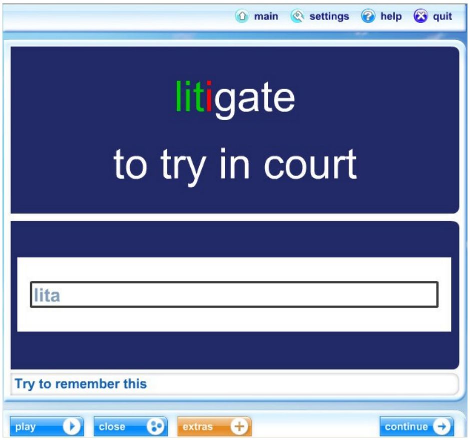
litigate

to try in court

His lawyer thought a lawsuit would be fruitless, but the client wanted to litigate.

Try to remember this

Figure 2. Word and definition together and multiple tabs that more provide detailed information.



litigate

to try in court

lita

Try to remember this

Figure 3. In spelling practice the user is shown both cue and response and is trying to type the word "litigate" and has typed in the first three letters correctly, and the first three letters of the cue word litigate are green as a result. However the fourth letter they have typed "a" is incorrect and thus the fourth letter of the cue word turns red to indicate the mistake. The main objective here is to increase the speed with which the word can be typed which will help the learner when typing is required with a time limit in the study component.

that it will pass into long-term memory (Ellis, 1995, 1997). The inclusion of pronunciation information in iKnow! encourages at the very least sub-vocal rehearsal on the part of the learner, which is likely to have similar benefits (Gathercole & Baddeley, 1993).

Furthermore, iKnow! target words and sentences are read aloud by multiple speakers, both male and female, which promotes attention by avoiding habituation to any one voice. Exposure to multiple voices is not only useful in helping learners become accustomed to a wide range of speech, but also help them remember words more easily (Goldinger *et al.*, 1991; Barcroft & Summers, 2005). Ideally images would also be presented to take advantage of the retention benefits indicated by Pavio's (1969) Dual Coding theory, however an image enabled version of iKnow! was not ready in time for the current study. Results from a study on the effectiveness of the image-enabled iKnow! will be presented in a forthcoming paper.

2.3.3. *Quizzes in the study component*

The content and order of the study component are dependent on learners' performance on different quizzes, as well as on the expanded rehearsal algorithm. In line with second language acquisition literature addressing involvement load, receptive/productive, and recognition/recall (text entry) distinction, the iKnow! system adopts eight different types of vocabulary quizzes: six multiple choice quizzes and two recall (text entry) quizzes. Multiple choice quizzes require the users to select the form of the word for the given definition (productive) and then asks them to select the meaning for the given form (receptive). Further variation on the two basic types of multiple choice quizzes is created by varying the number of choices the user must select from: specifically three, five and ten options. The expectation for the difficulty level is as follows: (1) providing form will be more difficult than providing meaning, (2) the more options the more difficult. Recall (text entry) activities are considered to be the hardest, and iKnow! adopts a strategy of starting the learner with the easiest activity and moving on to successively harder and harder activities as the learner succeeds at each stage. If the learner fails on a particular activity the next activity for the same vocabulary item will be the same or slightly easier.

Participants receive feedback on the accuracy of their responses after each quiz. When a given study session is finished an overview screen is shown that indicates progress on all items seen so far, and the learner is free to start another study session. The learner's performance on all the quizzes is also stored on a server-side component and used as input to a teaching algorithm to generate an optimized selection of items for the learner's next study session. The teaching algorithm follows an adaptive expanding rehearsal series (Mondria & Mondria, 1994). When the learner starts a subsequent study session the ten items chosen by the teaching algorithm will consist of a mixture of new items and review items. Review items will be those judged in need of additional study in order to be effectively learned by a target date.

2.4. *Related systems and their evaluation*

There are various commercial software products available that contain some of the same features as the iKnow! software described above. The Rosetta Stone software package provides various types of multi-media quizzes but does not support scheduling. The SuperMemo and Pimsleur systems both support different types of adaptive scheduling, but do not have the same range of multimedia quizzes as Rosetta Stone. In general there is not much in the way of experimental evidence for their effectiveness in comparison with each other. While the research literature contains many studies designed to examine the effect of a particular feature, e.g. the presence or absence of concordancing (Cobb, 1997). It is less common for research prototype consisting of multiple features to be compared with completely alternate approaches; the few studies that evaluated entire systems include: Groot (2000), who compared simple bilingual lists with a carefully crafted program that gradually increased the depth of vocabulary processing; Horst *et al.* (2005), who made multiple online tools available as part of a language course and assessed the retention of words selected by learners; Ma (2007), who compared the theoretically informed WUFUN software against a less structured version of itself and Yip & Kwan (2006), who showed that a selection of online digital games led to better long term retention compared to conventional activity based lessons. An ideal experimental comparison would be between multiple research prototypes and commercial systems, but there are significant challenges relating to getting the same content into multiple commercial systems, and also with obtaining permission from companies to conduct these studies.

In summary, the authors are unaware of any study that empirically tests, against a control, the effectiveness of a vocabulary teaching instructional software that has been designed with theoretically informed optimal task type, media and scheduling. Thus the question remains, will combining all the suggestions of the various theories related to vocabulary learning lead to an effective system compared to one that is theoretically uninformed? Naturally this is not the only question we would like ask. Given that all the aspects of the system do not destructively interfere with one another, the next question is which are the load bearing aspects, and how does this system compare to other systems comprising of different subsets of features. We hope to address these latter questions in future studies. In the study presented in the remainder of this paper, we address the question of whether a system informed by multiple vocabulary learning related theories produces better retention compared to a control condition.

3. Method

3.1. *Design*

The experiment employed two different software applications in order to create the different experimental conditions. The iKnow! software encouraged users to perform active recall through a series of quizzes, provided audio content and adapted to the pace of the individual user by sequencing re-presentation of material in such a fashion as to maximize subsequent retrieval. The iTango[3] software used in this study provided all the same text content as the iKnow! system, i.e. target words, definitions and example sentences, but did not structure the learning process; the material was simply available in 10 lists of 25 items each for the participants to review. The experiment was a mixed design where participants studied either exclusively on iKnow! or iTango (for between-subjects comparisons), and each participant was tested on both multiple choice and recall tests (for one within-subjects comparison) both before and after four weeks of study (for another within-subjects comparison). In addition a delayed post-test was conducted 5 weeks after the post-test. The dependent variable was the score achieved on the different tests, and the main independent variable was the study condition (iKnow! or iTango — differences summarized in Table 2 below). Participants in both iKnow! and iTango conditions were given complete freedom to study at the time of their choosing as long as they completed at least 1.5 hours a week by the end of the 4 week study period.[4]

[3] From the Japanese word "tango" meaning "word".

[4] The objective here was to allow participants to study longer if they felt so motivated, since we are also interested in whether one study condition was more intrinsically motivating than another.

Table 2. Differences between iKnow! and iTango conditions.

	iKnow!	iTango
Cue and Response	Yes	Yes
Example Sentence	Yes	Yes
Multiple Choice Tests	Yes	No
Recall Tests	Yes	No
Audio	Yes	No
Adaptive Sequencing	Yes	No

3.2. *Experimental hypotheses*

The following hypotheses were posed to test the effectiveness of the learner adaptive vocabulary learning system against traditional non-adaptive vocabulary list learning system.

Null Hypothesis[5]: Study under both conditions will not lead to an increase in post-test scores over pre-test scores.

Hypothesis 1: Study under both conditions will lead to an increase in post-test and delayed test scores over pre-test scores.

Hypothesis 2: Changes in post-test and delayed test scores over pre-test scores will follow the pattern iKnow! > iTango

3.3. *Participants*

A total of 36 native Japanese speakers studying English in Hawaii at TransPacific Hawaii College (TPHC) were recruited for the study. Participation was voluntary and open to all 200 TPHC students. The 36 participants, with ages ranging from 18–25 ($M = 19.9$, $SD = 1.68$), were from different classes in both years of the two year college. Assignment to a particular condition was based on when participants could attend training sessions. While this was not thus completely random assignment, the language survey (see Appendix A) did not indicate any particular trends in the two groups, i.e. they were evenly distributed in terms of English language experience and ability. Nineteen (7 male and 12 female) participants joined the iKnow! study condition (experimental group) and seventeen (3 male and 14 female) participants joined the iTango condition (control group). 4 iTango participants were eliminated

[5] It is important to note that there has been criticism of null hypothesis significance testing (NHST; Gigerenzer, 2004). The problem with this "ritual" is that it is a mixture of a number of statistical theories, and that statistical analysis of the data does not support statements about the probabilities that hypotheses are true. More specific shortcomings of the NHST ritual are the failure to provide alternative hypotheses, effect size, statistical power, confidence intervals and others. It is tempting then to avoid asserting a null hypothesis, however the value of the null hypothesis is that it makes explicit the researchers' expectation about the default situation given that none of their competing hypotheses holds true. As such we consider there to be value in presenting the null hypothesis along with multiple alternatives, particularly given that we will present effect sizes and statistical power where possible.

from the study after they failed to meet their weekly study requirements, leaving a total of 13 participants in the iTango condition, and 19 in the iKnow! condition. The four individuals were not distinguished in terms of the language learning questionnaire, having mixed TOEIC scores and time spent studying in the US. Three more participants were lost from each group at the delay test stage since they had graduated and left the institution.

The experimental group and the control group had similar language learning experiences (see Table 3), including average length of stay in English speaking countries (iKnow!: $M = 19.16$ months; iTango: $M = 17.92$ months), length of stay in Hawaii (iKnow!: $M = 15.16$ months; iTango: $M = 17.46$ months), and length of taking college level non-ESL classes (iKnow!: $M = 11.63$ months; iTango: $M = 12.92$ months). The range and dispersion for the length of taking college level non-ESL classes were larger in the iKnow! condition than in the iTango condition, since one of the iKnow! group participants had three years of experience taking such classes. Most were intermediate level students and their TOEFL scores ranged between about 430 and 560 (iKnow!: $M = 488.44$, $SD = 33.89$; iTango: $M = 480.23$, $SD = 33.37$).

The background information for the groups does not change a great deal as a result of the exclusions based on failure to complete study time requirements and attendance at the delayed test as shown in Table 4. There was also no gender pattern to the exclusions. The iTango's mean TOEFL score did drop by 10 points due to the exclusion of the participant with the highest reported score (failure to attend delayed test).

Table 3. Learners' language background information.

	iKnow! (experimental group)			iTango (control group)		
	M	*SD*	Range	*M*	*SD*	Range
Length of stay in English speaking countries	19.16 months	7.60 months	7–36 months	17.92 months	4.92 months	7–21 months
Length of stay in Hawaii	15.15 months	5.40 months	7–19 months	17.46 months	4.68 months	7–21 months
TOEFL score	488.44	33.89	427–560	480.23	33.37	450–563

Table 4. Learners' language background information after exclusions.

	iKnow! (experimental group)			iTango (control group)		
	M	*SD*	Range	*M*	*SD*	Range
Length of stay in English speaking countries	19.63 months	7.60 months	7–36 months	18.50 months	4.14 months	7–21 months
Length of stay in Hawaii	15.06 months	5.22 months	7–19 months	17.90 months	3.84 months	7–20 months
TOEFL score	486.69	34.71	427–560	470.70	21.40	450–537

3.4. *Materials*

The study material was taken from the academic word list (AWL), a list of 570 word families identified by range and frequency analysis as being particularly useful in academic study (Coxhead, 2000). Results from three previous pilot studies indicated that 250 AWL words would be appropriate given the study times involved in this experiment. Study items consisted of the AWL head word itself, an associated synonym set or short definition and an example sentence, all in English. TPHC faculty members created all study materials in terms appropriate to the ability level of the participants. The example sentences were designed to make it possible to infer the meaning of the target AWL word from the context of the sentence. The original intention had been to create content for all 570 academic head words, however the number of close synonyms in the AWL made this difficult. Ultimately 350 words with sufficiently distinct meanings were selected, and each associated with a response (short definition or list of synonyms) and an example sentence.

Table 5. Example AWL words.

Freq.	Cue	Response	Example Sentence
1	approach	manner; way of doing something	My **approach** to studying is to do it as quickly as possible and then forget it.
2	achieve	accomplish; reach; do	My grandfather was able to **achieve** many things in his life. He had a happy marriage, six children, built a house by himself, and won two awards for his work with the environment.
3	alternative	other; another; different	There is a traffic jam on the highway, so people should find an **alternative** way to get to work and school.
4	access	ability to enter; right to use	The students did not have **access** to the information in the computer because only teachers were allowed to use it.
5	academy	school; college	My sister goes to Sacred Hearts **academy**. It is a very good school and she really likes studying there.
6	accurate	true; correct in every detail	The story in the newspaper is **accurate**. It is true that John Bishop won $1,000,000 in Las Vegas.
7	adapt	change to fit a new situation	When foreign students come to the United States, they have to **adapt** in many ways. For example, sometimes they change the foods they eat or the clothes they wear.
8	abandon	leave and not go back; throw away	Many people **abandon** their cars during a storm. They just leave their cars on the road and walk home.
9	anticipate	wait for; look forward to	I happily **anticipate** summer vacation every year. I really look forward to it and can't wait for it to arrive.
10	adjacent	next to; nearby	The bank is **adjacent** to the post office. The two buildings are next to each other.

Each of the AWL words has an associated frequency level from 1 to 10, with 1 representing the most frequent and 10 representing the least. Table 5 shows examples from each frequency level:

Earlier pilot studies in which participants attempted to study 150 items for a week indicated that it might be difficult for the participants in this month long study to work through all 350 items. As a result the content was reduced to 250 items by removing all the level 8, 9 and 10 frequency words. Table 6 below shows the number of words in each frequency category.

At each stage of testing all participants took exams consisting of 75 item recall and 75 item multiple choice questions (MCQs). The MCQs gave six possible responses including "none of the above", "I don't know", the correct definition, and three distractors drawn from the other 74 items. The recall tests were in the reverse direction from the MCQ tests and required the participant to type in the target word given the definition. The 75 items in the recall and MCQ exams were different and drawn randomly from the 250 items used in the study, and presented in a different random order at each test.

3.5. *Procedure*

The study employed a mixed subjects design with each participant studying under one condition (iKnow! or iTango). Participants in both conditions were given complete freedom to study at the time of their choosing as long as they completed at least 1.5 hours a week for a total of 6 hours study by the end of the four week study period. Instructions detailing the requirements were printed out in Japanese and English (see Appendix B), and handed out to all the participants at an orientation meeting. The instructions were then explained verbally to the group in both English and Japanese.

In order to assure students made progress during the study period, the researchers tracked learners' study time for all the participants and sent reminders to those who were studying less than 1.5 hours per week. In this experiment the target date in the iKnow! program was set to the day of the post-test. The only effect of the target date in the iKnow! system was to give an indication to the participant as to whether they would see all of the content given their current rate of study. The iTango system had no target date and thus there was no indication to the participant about their rate of study. All participants were given a 20 minute introductory lecture on their study condition and had all their questions answered.

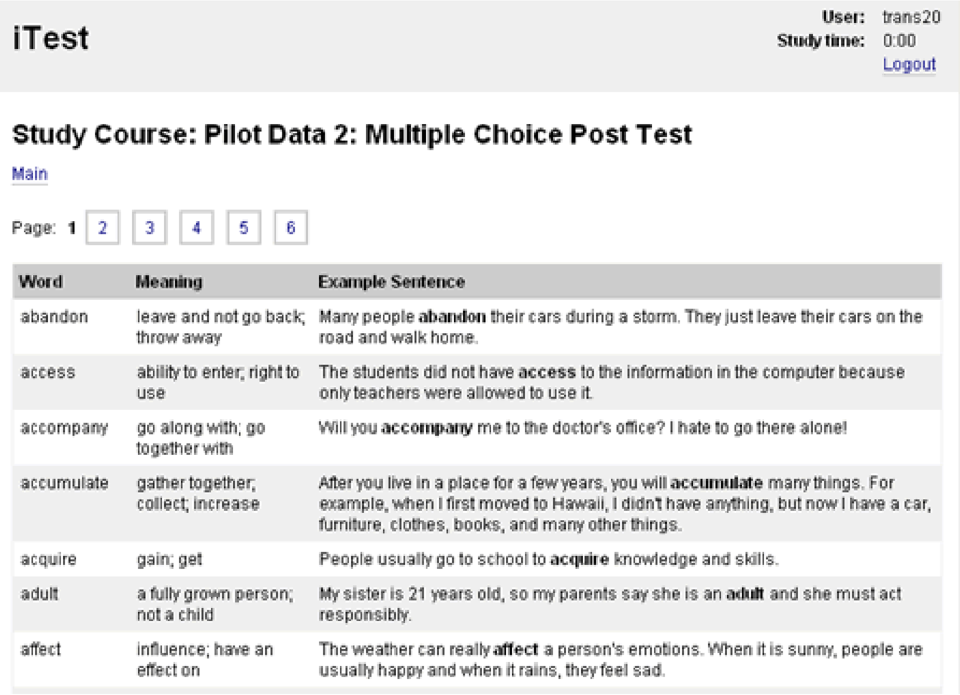
Participants were offered a monetary incentive to complete the study as long as they studied the required minimum six hours, but would not receive any additional reward if they studied longer, although they were free to do so if they felt so inclined.

Table 6. Number of words from each frequency level.

Frequency	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Level 7
#words	43	34	36	43	35	39	20

The iKnow! study condition used the iKnow! program to study the targeted words, while the iTango study condition simply provided access to the study materials, i.e. cue, response and example sentence for each of the 250 items. An example study screen from iTango is shown in Figure 4. iTango condition participants could scroll up and down to see 25 items per page, and click through to one of each of 10 pages. The amount of time studied was indicated at the top right of the screen.

Participants underwent pre and post and delayed-post-tests, each consisting of 75 productive MCQ questions and 75 receptive recall questions, covering a subset of the 250 academic word list items that were to be studied. There was a degree of mortality or attrition such that the number of participants in successive tests was reduced as shown in Table 7 below.



iTest

User: trans20
Study time: 0:00
Logout

Study Course: Pilot Data 2: Multiple Choice Post Test

Main

Page: 1 2 3 4 5 6

Word	Meaning	Example Sentence
abandon	leave and not go back; throw away	Many people **abandon** their cars during a storm. They just leave their cars on the road and walk home.
access	ability to enter; right to use	The students did not have **access** to the information in the computer because only teachers were allowed to use it.
accompany	go along with; go together with	Will you **accompany** me to the doctor's office? I hate to go there alone!
accumulate	gather together; collect; increase	After you live in a place for a few years, you will **accumulate** many things. For example, when I first moved to Hawaii, I didn't have anything, but now I have a car, furniture, clothes, books, and many other things.
acquire	gain; get	People usually go to school to **acquire** knowledge and skills.
adult	a fully grown person; not a child	My sister is 21 years old, so my parents say she is an **adult** and she must act responsibly.
affect	influence; have an effect on	The weather can really **affect** a person's emotions. When it is sunny, people are usually happy and when it rains, they feel sad.

Figure 4. iTango study interface (iTest is the name of the quiz software used in pre and post tests, and the same framework supports the iTango interface, a term created to avoid confusion between test and study modes).

Table 7. Number of participants attending each test.

Condition/Test	PreTest	PostTest	DelayedTest
iKnow	19	19	16
iTango	17	13	10

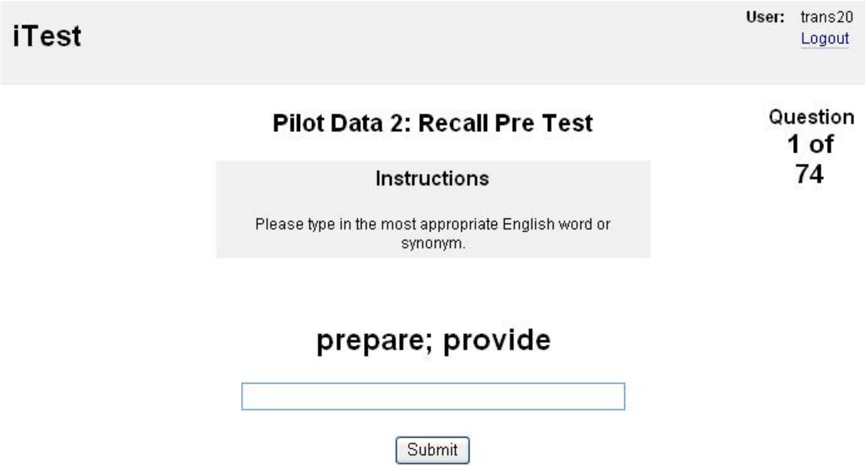
The delayed post-test was not announced until the week before the test itself in order to avoid providing any incentive to the participants to keep studying after the post-test. For both types of test 75 items were selected at random, ensuring that either 10 or 11 items came from each of the 7 available frequency categories. The order of the items was then randomized. The item pool for both types of test was maintained across the three stages of the test, i.e. the 75 items presented in the MCQ pre-test were then represented in the MCQ post-test and then presented again in the MCQ delayed post-test, although in a different randomized order each time. The recall test interface is shown in Figure 5 below.

No feedback was given as to whether an answer was correct or otherwise for either the recall test above, or the multiple choice test, shown in Figure 6.

Participants were given 25 minutes to complete the recall test and 20 minutes to complete the MCQ test.

4. Results

On average, iKnow! participants studied slightly less than those in the iTango group (iKnow!: $M = 6$ hrs 19 mins, $SD = 2$ hrs 40 mins; iTango: $M = 6$ hrs 40 mins, $SD = 2$ hr 6 mins). The iTango group had unrestricted access to all study materials from the start of the study period, but the iKnow! group did not necessarily see all study items due to the step by step progress enforced by the iKnow! (iKnow!: $M = 186$ items, $SD = 61$). For iKnow! the average number of days between study sessions was 2.9 ($SD = 1.3$), while for iTango it was 2.8 ($SD = 1.7$). The average number of days on which studying took place for iKnow! was 9 ($SD = 3.9$ days) and the average study time per day when studying took place was 42 minutes ($SD = 18$ min). The average number of study days for iTango was also 9 ($SD = 4.7$



iTest

User: trans20

Pilot Data 2: Recall Pre Test

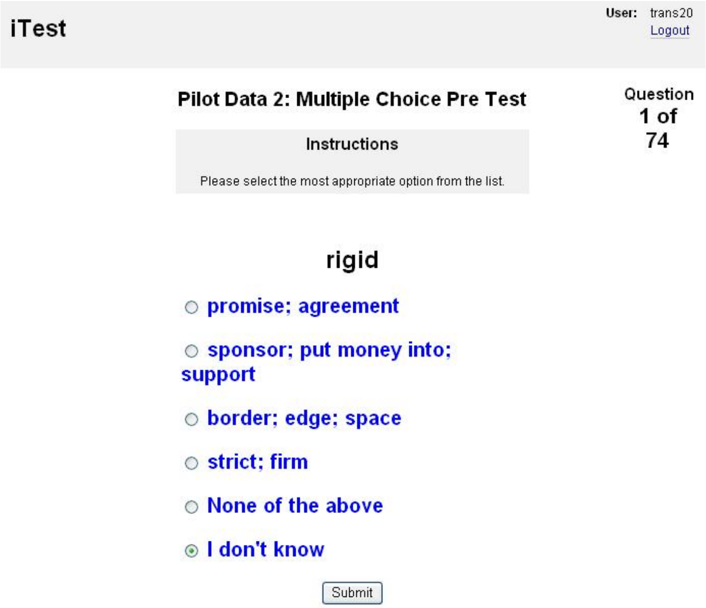
Question 1 of 74

Instructions

Please type in the most appropriate English word or synonym.

prepare; provide

Figure 5. The recall test interface. This sort of retrieval is very challenging even for native speakers, unless they have recently reviewed a set of words in which a possible response is present.



iTest

User: trans20
Logout

Pilot Data 2: Multiple Choice Pre Test

Question
1 of
74

Instructions

Please select the most appropriate option from the list.

rigid

○ promise; agreement

○ sponsor; put money into; support

○ border; edge; space

○ strict; firm

○ None of the above

● I don't know

Submit

Figure 6. The MCQ test interface.

days) and the average study time per day when studying took place was 43 minutes (*SD* = 28 min). Participants of both conditions appeared to study according to similar patterns, but the iTango condition did not allow a detailed tracking of items seen.

The imposition of time limits on the pre-, post- and delayed-tests meant that not every participant was able to answer all of the 75 questions in each test. This problem was compounded by a technical issue during the pre-test which slowed the rate at which the participants could progress through the test, particularly in the iTango condition. In the MCQ pre-test, the participants in the iKnow! condition saw an average of 84% (*SD* = 15%) of the test items, whereas participants in the iTango condition only saw an average of 40% (*SD* = 11%). As for the recall pre-test, on average, the iKnow! group went through 79% (*SD* = 26%) of the items, while the iTango group only managed to get through 49% (*SD* = 11%) of the items. These problems were resolved before the post and delayed tests, where all participants were able to go on to respond to all 150 questions within the time limits. Thus, in order to compare the test results across three trials, all test results were converted into percentage scores (i.e. correct answers divided by the number of items seen within the time limit).

Both groups made improvements from pre to post recall tests; however, in the multiple choice tests the iKnow! group showed a marginal pre to post improvement while the iTango group actually deteriorated. Table 8 summarizes the mean, standard deviation, and the difference scores for each group for each measure. In recall

Table 8. Descriptive statistics of the pre-, post- and delayed-tests (means are percentage correct).

Group	Test	*N*	Pre-Test		Post-Test		Delayed Post-Test		Pre-post difference	Pre-delayed difference
			M	*SD*	*M*	*SD*	*M*	*SD*		
iKnow	MCQ	16[6]	73%	13%	83%	15%	80%	15%	+10%	+7%
	Recall	16	8%	5%	39%	26%	23%	16%	+31%	+15%
iTango	MCQ	10	70%	14%	66%	11%	61%	13%	−4%	−9%
	Recall	10	4%	4%	8%	4%	7%	4%	+4%	+3%

tests, improvement by the iKnow! group was greater than that of the iTango group; i.e. the iKnow! group showed a 31% improvement compared to a 4% improvement by the iTango group. In the MCQ test, the iTango group showed a slight pre to post vocabulary loss, while the iKnow! group improved by 10%. All the delayed test scores were lower than the immediate post-test scores, but the pattern of the test performance remained constant with the iKnow! group staying well ahead of the iTango group for both recall and MCQ tests.

Figure 7 shows boxplots for each of the datapoints in Table 8, clustered first by group (iKnow!/iTango) and then by test type (MCQ/Recall). Figure 7 highlights the increases achieved by the iKnow! group versus the much smaller increases or even

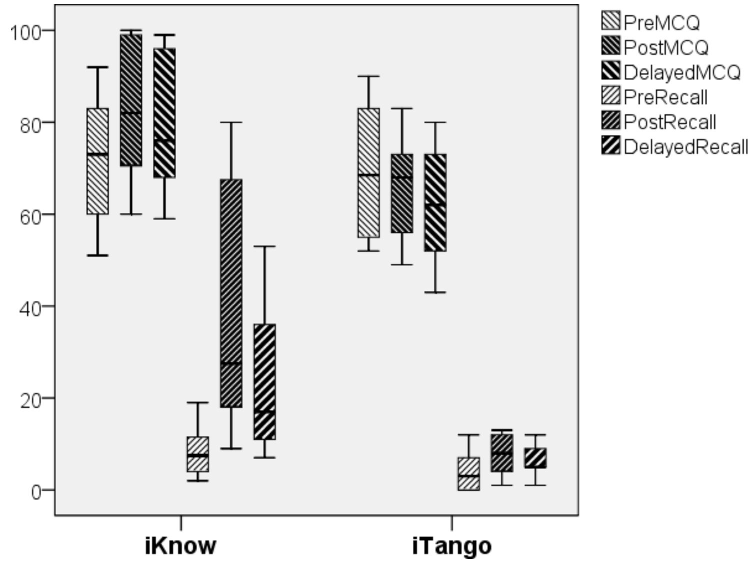


Figure 7. Pre, post and delayed post-test score box plots on MCQ and recall tests.

[6] Those participants that did not participate in the delayed post test were excluded. This meant 3 participants from both iKnow! and iTango groups were excluded, on top of the 4 participants from the iTango group that did not keep up their study hours.

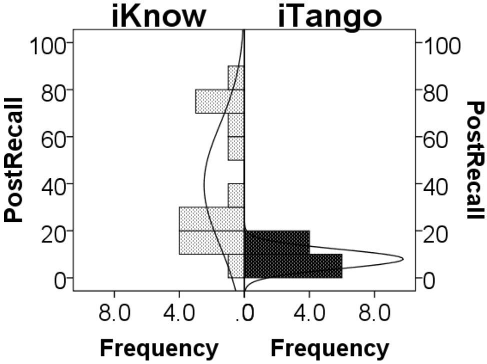


Figure 8. Post test score distribution on recall tests.

decreases of the iTango group. We can also see the very large standard deviation in the recall post-test for the iKnow! group which warrants further investigation.

The distributions of scores from the pre, post and delayed multiple choice tests were similar and roughly normal although there was evidence of a ceiling effect as scores reach the maximum 100% leading to somewhat skewed iKnow! distributions. Conversely the distributions of the recall test scores suffered from a floor effect since many participants started off scoring close to zero. Nonetheless, the initial pre-test distributions were roughly normal. However the post-test scores showed a pronounced bimodal distribution (see Fig. 8), which was still apparent although less pronounced in the delayed test score distribution.

Before analyzing main outcomes, a repeated measures analysis of variance (ANOVA) with *group* (iKnow! and iTango) as between subject variable, and *test* (MCQ and recall) was conducted for the pre-test scores to examine the comparability between the iKnow! and the iTango group. As we can see from Table 9 the analysis revealed that there was no statistically significant difference between the two groups at $p < 0.025$[7] level ($F(1, 24) = 1.043, p = 0.371$, *partial eta squared* $= 0.042$).

The parametric assumptions of normality and homogeneity of variance were not met in several of the post and delayed recall tests. These issues could not be resolved with outlier removal, or with any standard data transformations indicating the need for more robust non-parametric tests. As a result we employed Wilcox's (2008) robust three-way ANOVA with dependent variables (between groups, within time, and within test), which gave a significant effect of group ($Q = 19.11$, $p < 0.0001$), a significant effect of time ($Q = 6.99$, $p < 0.001$), and a significant effect of test ($Q = 335.46$, $p < 0.0001$). In addition there were significant effects of Group * Test ($Q = 7.25$, $p < 0.01$) and Group * Time ($Q = 3.75$, $p = 0.024$), but not Test *

[7] When performing multiple statistical tests, a family-wise error rate of $p < 0.05$ is maintained within a study to reduce false positives (i.e. Type I error) among a class (family) of tests (Frey *et al.*, 2000: 333). The current study involves two statistical tests, so statistical significance is only reported for $p < 0.025$ ($p < 0.05$ divided by 2).

Table 9. Repeated measure ANOVA results for initial differences between the iKnow and iTango.

Source	F	*df*	p	*Partial Eta Squared*	*Observed Power*
Between subject effect					
Group	1.043*	1	0.317	0.042	0.102
Error		24			
Within subject effect					
Test	901.353*	1	0.001	0.974	1.000
Error		24			
Test * Group	0.159*	1	0.694	0.007	0.036
Error		24			

$^*p < 0.025$

Time ($Q = 3.50$, $p = 0.03$) and not Group * Test * Time ($Q = 2.35$, $p = 0.099$) interactions. According to Wilcox (2008) "*There is no good way of estimating power for complex designs*" such as these, which encourages us to construct experiments that collect data that will avoid breaking parametric assumptions in future.

The effect size for the benefit of using iKnow! over iTango from pre to post-test on both recall and MCQ tests were on the large side, being 1.67 and 1.29 respectively. The effect sizes for the benefit of using iKnow! over iTango from pre to delayed test on both recall and MCQ tests were also in the large range, being 1.37 and 1.35 respectively. Measuring effect size provides a method to compare the magnitude of an effect with that found in other studies. In general an effect size of 0.2 is considered small, 0.5 is medium and 0.8, large (Cohen, 1988). In the vocabulary study literature it is relatively rare to see an effect size greater than 1.0, making this a noteworthy result.

5. Discussion

Analysis showed no significant difference between the pre-test scores of the two groups. In combination with the statistically significant effect of *time* in the three way ANOVA this suggests that the changing test scores were due to the experimental manipulation. This discredits the null hypothesis and leads us to retain $H1$ that *study under both conditions will lead to an increase in post-test and delayed test scores over pre-test scores*, for the recall tests at least, since the iTango MCQ post-test scores were actually lower than those at pre-test. This appears to be a consequence of iTango users making more frequent use of the 'I don't know' option in the post-test compared to the pre-test, when they were perhaps more inclined to make a guess. This suggests the need for a careful review of the use of "I don't know" options in future MCQ tests, and of the instructions given to participants about how to answer multiple choice quizzes.

Nonetheless the statistically significant interaction effect between time and group implies that the two groups performed significantly differently in the post and delayed tests. Thus, the larger improvement in the post-test for participants in the iKnow! condition compared to the iTango condition gives support for $H2$, that the *changes in post-test and delayed test scores over pre-test scores will follow the pattern iKnow! > iTango.*

5.1. *Methodological issues*

While the current experimental results support the first hypothesis for the recall component and the second hypothesis in general there are a number of confounding factors that must be taken into account when interpreting the results. These are:

(1) variations in time spent studying in each condition
(2) ceiling/floor effects and test type
(3) variation in the amount of material presented during study
(4) variation in the amount of material tested
(5) use of pseudo-random assignment
(6) bimodal recall distributions

5.1.1. *Variations in time spent studying in each condition*

The first concern is that due to the open ended nature of the experiment participants in one condition may have studied longer than the other. This design was intended to allow the participants to study longer if they felt so motivated, given that motivation may also be influenced by the study condition. It was assumed that participants would find the experimental condition more motivating than the control condition, and that given a significant difference in study times, learning rate could be compared. However the study time difference was in the opposite direction from that expected with participants in the control condition studying slightly longer on average than those in the experimental condition. Since this only reinforces the existing iKnow! > iTango result a learning rate analysis was deemed unnecessary.

It is important to note that we cannot rule out a degree of imprecision caused by the way study time was measured by the software in the two conditions. Careful attention to how study time is measured, as well as improving mechanisms for recording what is being studied in the control condition are indicated in future studies, although clearly it is difficult to perfectly monitor participants' study behavior. An associated concern is that participants may have been performing other tasks while studying, in order to "cheat" the system. We set up the software to automatically stop recording time after five minutes of inactivity, and at present we have no strong reason to suspect that this sort of behavior was more prevalent in one condition than the other.

5.1.2. *Ceiling/floor effects and test type*

A second concern is that the recall pre-test likely suffered from a floor effect, while the MCQ post-test suffered from a ceiling effect. This appears difficult to avoid when employing both types of test on the same content; although the MCQ test could be made harder by adjusting the number and type of distractors, while the recall test could be made easier by allowing the participants to answer in their native language, or employ a marking system that allowed synonyms. Given the high scores in MCQ it appears the participants either knew, or could effectively guess the words, and thus one might argue that the recall test is really measuring the spelling ability of the participants, rather then their recall ability as such. However, it makes little difference even if we do calculate the recall results with an algorithm that can cope with minor spelling mistakes (e.g. Soundex adjusted Recall post-test: iKnow!: $M = 39\%$, $SD = 25\%$; iTango: $M = 12\%$, $SD = 6\%$, i.e. a boost of about 2% to both conditions).

There is also a concern about whether MCQ/recall tests are really measuring something relevant to learner's objectives (which are presumably to be able to read/write and speak/understand English) although we are at least measuring something related to form-meaning mapping. We can presumably be confident that if a learner establishes the form/meaning mapping then the tests should become easier to pass – which does not rule out the learner using other approaches to passing the test such as using mnemonics, alliteration and so forth.

5.1.3. *Variation in amount of material presented*

One serious issue is that the iKnow! students did not see all of the study material. At first glance it seems that iKnow! participants would thus be at a competitive disadvantage since they might have been tested on words they have never encountered. However one might also argue that they were able to study a smaller set of words more intensely and that an iTango group studying 186 items on average might do better than one studying 250. Subsequent studies should ensure that iKnow! introduces all the items available to the iTango group.

5.1.4. *Variation in amount of material tested*

Another important issue is that not all participants were able to see every question on the pre-tests. This was partly due to time limits placed on completion of the tests for logistical reasons (25 minutes to complete 75 items in recall test and 20 minutes to complete 75 items in MCQ test), but mainly due to an unanticipated computer problem that slowed participant progress through the pre-test, particularly for the control condition. The result of this was the number of pre-test questions answered by the control group was approximately half that of the experimental group (iKnow!: $M = 61.6$, iTango: $M = 32.4$). The computer problem was fixed for the post and delayed tests, during which all participants saw all the questions. As a result all test scores were converted to percentages to allow comparison between the two groups, and tests at different times.

One might argue that converting to percentages does not fix this problem, however the items appearing in the pre- and post-tests were evenly distributed according to the frequency (and hence inferred difficulty) level, implying that the percentage score on a small proportion of questions should not be significantly different from a score obtained on larger proportion of questions. This is further supported by the fact that the groups average scores were not significantly different and that the ability levels of the two groups was similar according to TOEFL/TOEIC tests.

A more general question concerns the use of time limits on the tests, which itself conceals a deeper question about what it means to "know" a word's form-meaning mapping. If it takes a learner half an hour to answer a question, we would not presume that the learner "knows" a word's form-meaning mapping to the same extent as if the learner were able to answer the same question in a matter of seconds. Time limits are common in real-world exams where they make sense in terms of both fairness and logistics. In order for us to more precisely measure the participants knowledge relating to each word a time limit placed on each individual item might make more sense. However removing time limits and recording the time taken to answer each question on the test would be a more sensitive metric.

5.1.5. *Use of pseudo-random assignment*

Another issue is the pseudo-random assignment that was employed for logistical reasons. This should be replaced with random assignment or even better, stratified assignment based on pre-test scores. Another possibility would be to eliminate items from the analysis that the participants already knew, although it is not clear what overall effect this would have.

5.1.6. *Bimodal recall distributions*

The bimodal distribution of the recall post-test and delayed test recall results for the iKnow! group was unexpected. Detailed analysis revealted that the higher of the two modes consisted of a group of male students in the iKnow! condition that had entered into a competitive frame of mind and were challenging each other regarding who could make the most progress on the iKnow! system. This might be seen as a confounding factor, but the results have the same significance level if this group is removed (although naturally the effect size drops). One might argue that this is a fundamental difference between the two groups which threatens the validity of the experiment, but another way of looking at it is to consider that no such competitive behavior emerged in the control condition, and arguably the competitive approach may be encouraged by the incremental personalized progress provided by the iKnow!, i.e. the only avenue for competition in the iTango condition was for students to compare study times, whereas the iKnow! provides the learner with various progress metrics. It is interesting to note that none of the "competitors" in the higher mode were distinguished from the other participants in other conditions, either by amount of study time put in, or any background language ability measures, or indeed in the pre-test scores. This is perhaps indicative of the potential power of the competitive spirit to increase motivation and perhaps attention, which can then lead to learning performance gains.

Overall these various issues are confounding factors on the validity of the study indicating the necessity of further studies to verify these results.

5.2. *Theoretical implications and limitations*

Earlier in this paper we described how a variety of work in the fields of CAL, CALL, CAVL and second language acquisition informed the design of the iKnow! software. The main theoretical basis for the iKnow! design is summarized as follows:

(1) CAVL design guidelines (Overbaugh, 1994; Nation, 2001; Plass, 1998)
(2) Three key CAVL aspects: task-type, media & scheduling. Defined in this paper on the basis of the following:
    (a) Value of learner performing complex tasks comes indicated by theory of deep processing (Craik & Tulving, 1975)
    (b) Value of multiple media informed by dual coding theory (Paivio, 1969) and phonological loop (Ellis, 1995, 1997)
    (c) Value of adaptive scheduling comes from theory of active retrieval (Baddeley, 1997) and adaptive expanding rehearsal series (Mondria & Mondria, 1994)
(3) Motivation (Dornyei, 2001; Wlodlowski, 1999) is another important factor which affects the three key CAVL aspects and includes
    (a) setting short term goals (Bandura & Schunk, 1981)
    (b) providing orienting activities (Overbaugh, 1994)
    (c) optimal number words to study[8] (Gillund & Shiffrin, 1984)

The results of the study presented in this paper could be taken as a validation of our synthesis of CAVL design guidelines and focus on the three key aspects of CAVL, however it would be unwise to assume this is a robust finding. A more effective assessment as to the value of different design guidelines would require a meta-analysis of experiments on multiple systems each informed by different design guidelines. As a consequence the results of the current study can only tentatively confirm the value of the CAVL design approach we advocate.

What we can say is that the three key CAVL aspects appear to have had a constructive combination, suggesting that each has an important, independent role in promoting long term vocabulary retention. For example, it might have been the case that each aspect operated through the same underlying mechanism, e.g. the benefit of dual coding through multiple media types coming as a result of stimulating subsequent retrieval at expanding intervals, or expanding retrieval operating by promoting deep processing. Our results suggest that all three aspects are not providing benefit through the same mechanism, since the effect sizes we obtain are larger than those found in experiments when only individual aspects were manipulated (Shiung & Joseph, forthcoming). However each of the aspects is modulated

[8] Likely to have connections with cognitive factors beyond motivation.

by motivational factors, and so we cannot definitively separate each as operating through a different underlying cognitive mechanism. That would require more complex and subtle experiments. Thus the implications of our findings for the field of CAVL research and development are that more experiments should be performed on the influence of combined factors to demonstrate their independence. Also of benefit would be a more detailed analysis of what users are doing within the software, in order to determine where the maximum benefit is being derived from.

6. Conclusion

This paper focused on the techniques by which the form-meaning mappings of vocabulary items can be effectively studied with computer assistance. We explored some of the key recommendations from the literature and explained how these had been incorporated into the design of the iKnow! learning application. This paper also reported on a study of the effectiveness of designing an application in this way by comparing the iKnow! software with a traditionally oriented application called iTango. Although there were a number of confounding factors affecting the experiment, the results were strongly suggestive of large boosts to the learning outcomes experienced by participants in the experimental condition.

There were many differences between the iTango and iKnow! applications, but the central ones followed from three key aspects of CAVL: task type, media and scheduling. The implementation of each was modulated by motivational concerns leading to a number of distinguishing features including an adaptive expanded rehearsal series, staggered presentation modes, gradual transition from easier to more complex quizzes, audio content, the presence of detailed feedback and motivational support. At the very least the results presented in this paper suggest that the combination of these features does not lead to destructive interference, i.e. the different features reducing the effectiveness of the others. Furthermore the large effect sizes suggest a constructive interference, indicating the likely independence of at least some of the cognitive pathways to boosting long term retention. However it is not clear from this experiment which, if any, of the features is mainly responsible for the improvement over a more traditional study approach. This would require a more complex experiment with more conditions; one that we hope to conduct soon.

In the form that it was evaluated, a system like iKnow! could be described as inheriting from a largely behaviorist tradition of instruction theory, although with a number of concessions to constructivist principles. While behaviorist theories of learning and teaching have long been out of fashion the recent trend appears to go beyond simple constructivism to a socio-cultural position. Clearly learning second language vocabulary is part of a broader language learning activity, which itself is wedded to a larger socio-cultural enterprise. If the objective of the learner is to be able to communicate with native speakers of another language, socio-cultural approaches to learning will have to play a critical role in the learner's developing language abilities. The interesting, and as yet unresolved question, is whether there is benefit to be derived from socio-cultural approaches to learning during the initial phases of learning a language such as basic vocabulary acquisition. From the first author's experience of learning a foreign language it seems clear that vocabulary acquisition is most effective when it takes place in situations of emotional importance in the context of a learner's existing goals, e.g. negotiating with a friend over what to eat when you are both hungry. However one must acknowledge that at least some learners will lack the confidence to dive into to such language situations until they have at least confident that they know a smattering of vocabulary.

Thus one point of view would be to say that independent vocabulary study is a useful complement to other approaches to learning languages. However it remains to be shown that the time investment for a given level of vocabulary retention through independent study could not perhaps be more effectively spent interacting with other learners at the same level, e.g. discussing the meanings of vocabulary items with peers and native ability teachers. Given a lack of access to peer students and particularly teachers, it is extremely tempting for learners to invest time in solitary computer based study. However this is starting to change with the advent of online communities with simple and cheap support for audio communication, such as Second Life (Silva, 2008), an area that we are also investigating.

Perhaps the critical question is whether form-meaning mapping is something learner's must memorize, or something they must understand? Discussing all the intertwined cultural and semantic nuances in the mapping might help the learner have a more fully rounded appreciation of the word, but perhaps they are better off cramming a set of mappings and then trying to get by in the language itself, interacting with other speakers and then learning those nuances? We hope that our future experiments will help answer these questions.

Acknowledgments

The authors would like to thank the staff and students of Transpacific Hawaii College for their co-operation and patience in making this study possible, to Professor Daniel Suthers and the anonymous reviewers for their constructive comments and to Cerego Japan for providing research funding to the University of Hawaii in order to evaluate their iKnow! software. The latest version of iKnow! is available on the SmartFM website http://smart.fm

Appendix A. Language Survey

Name (   )
Language Background

1. What is your native (first) language? [   ]
2. How long have you been in Hawaii? [   ] years [   ] months

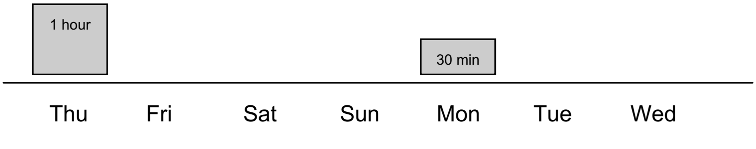
3. How long have you lived in English-speaking countries in total (not including traveling)?
   [ ] years [ ] months
4. How long have you been studying academic content (not ESL) at an English language institution (e.g. TransPacific)?
   [ ] years [ ] months
5. What is your highest education level? [High school / 2 year college / 4 year college]
6. Have you taken a proficiency test (e.g., TOEFL, TOEIC, Eiken)? If so, please write the name of the test, your score, and when you took the test. If you have taken several tests, please list all of them.
   Test name [ ] Your score [ ] When? [ ] years / [ ] months ago
   Test name [ ] Your score [ ] When? [ ] years / [ ] months ago
   Test name [ ] Your score [ ] When? [ ] years / [ ] months ago
   7 What are your areas of academic interest? [Circle all that apply]
   [Art/Business/Communication/Computer/Economics/Education/Geology/
   History/Linguistics/Psychology/Science/Sociology/Travel industry/Other]

Appendix B. Study Requirements

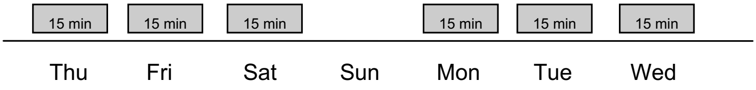
Cerego Vocabulary Study
Pre-test: 8th November, 9am or 10:30am, Computer Lab A & B
Post-test: 6th December, 9am or 10:30am, Computer Lab A & B
You must study a total of at least 1.5 hours each week:

8th November - 14th November:	1.5 hours
15th November - 21st November:	1.5 hours
22nd November - 29th November:	1.5 hours
30th November - 6th December:	1.5 hours
Total:	6 hours

You are free to study in whichever pattern you choose as long as you study 1.5 hours per week, so for example in any given week you could study like this:



or this:



You may study for more than 1.5 hours a week, but no less.

References

Adepoju, A. A., & Elliott, R. T. (1997). Comparison of different feedback procedures in second language vocabulary learning. *Journal of Behavioral Education*, *7*(4), 477–495.

Al-Seghayer, K. (2001). The effect of multimedia annotation modes on L2 vocabulary acquisition: A comparative study. *Language Learning & Technology*, *5*(1), 202–232. Retrieved April 15, 2007, from http://llt.msu.edu/vol5num1/alseghayer/default.html

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review*, *79*(2), 97–123.

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

Baddeley, A. D. (1997). *Human memory: Theory and practice.* Hove: Psychology Press.

Bahrick, H. P., & Bahrick, P. O. (1964). A re-examination of the interrelations among measures of retention. *Quarterly Journal of Experimental Psychology*, *18*, 318–324.

Bandura, A., & Schunk, D. (1981). Cultivating competence, self-efficacy and intrinsic interest through proximal self-motivation. *Journal of Personality and Social Psychology*, *41*, 586–598.

Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning*, *57*(1), 35–56.

Barcroft, J., & Sommers, M. S. (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition*, *27*(3), 387–414.

Beatty, K. (2003). *Teaching and researching computer-assisted language learning.* Essex, England: Pearson Education Limited.

Bloom, B. S. (1956). *Taxonomy of educational objectives.* Handbook I: Cognitive Domain. New York: McKay.

Bloom, K. C., & Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *Journal of Educational Research*, *74*(4), 245–248.

Brett, P. (1998). Using multimedia: A descriptive investigation of incidental language learning. *Computer Assisted Language Learning*, *11*(2), 179–200.

Chandler, D. (1984). *Young learners and the microcomputer.* Milton Keynes, Open University Press.

Chaudron, C. (1982). Vocabulary elaboration in teachers' speech to L2 learners. *Studies in Second Language Acquisition*, *4*(2), 170–180.

Chun, D. M., & Plass, J. L. (1996). Effects of multimedia annotations on vocabulary acquisition. *The Modern Language Journal*, *80*(2), 183–198.

Chun, D. M., & Plass, J. L. (1997). Research on text comprehension in multimedia environments. *Language Learning & Technology*, *1*(1) 60–81.

Cobb, T. (1997). Is there any measurable learning from hands-on concordancing? *System*, *25*(3), 301–315.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

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

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology*, *104*, 268–294.

Crothers, E., & Suppes, S. (1967). *Experiments in second-language learning.* New York, NY: Academic Press.

Davis, R., Sutherland, N. S., & Judd, B. R. (1961). Information content in recognition and recall. *Journal of Experimental Psychology*, *61*, 422–428.

Dempster, F. N. (1987). Effects of variable encoding and spaced presentations on vocabulary learning. *Journal of Educational Psychology*, *79*(2), 162–170.

Dornyei, Z. (2001). *Teaching and researching motivation.* England: Pearson Education Limited.

Dubois, M., & Vial, I. (2000). Multimedia design: The effects of relating multimodal information. *Journal of Computer Assisted Learning*, *16*(2), 157–165.

Ducheneaut, N., Yee, N., Nickell, E., & Moore, R. J. (2006). "Alone together?": Exploring the social dynamics of massively multiplayer online games. In *Proc. SIGCHI Conference on Human Factors in Computing Systems* (pp. 407–416). New York, NY: ACM Press.

Duncan, M. J., & Biddle, B. J. (1974). *The study of teaching.* New York: Holt.

Elliott, R. T., & Adepoju, A. A. (1997). First language words as extra-stimulus prompts in learning second language vocabulary. *IRAL*, *35*(4), 237–250.

Ellis, N. C. (1995). Vocabulary acquisition: Psychological perspectives and pedagogical implications. *The Language Teacher*, *19*(2), 12–16.

Ellis, N. C. (1997). Vocabulary acquisition, word structure, collocation, word-class, and meaning. In N. Schmitt and M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 122–139). Cambridge: Cambridge University Press.

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

Ellis, R. (1995). Modified oral input and the acquisition of word meanings. *Applied Linguistics*, *16*(4), 409–441.

Folse, K. S. (2006). The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly*, *40*, 273–293.

Forlano, G., & Hoffman, M. (1937). Guessing and telling methods in learning words in a foreign language. *Journal of Educational Psychology*, *28*, 632–636.

Frey, L. R., Botan, C. H., & Kreps, G. L. (2000). *Investigating communication: An introduction to research methods* (2nd ed.). Needham Heights, MA: Allyn & Bacon.

Gagne, R. M. (1974). *Essentials of learning for instruction Hinsdale.* Illinois: The Dryden Press.

Gathercole, S. E., & Baddeley, A. D. (1989). Evaluation of the role of phonological STM in the development of vocabulary in children: A longitudinal study. *Journal of Memory and Language*, *28*, 200–213.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*, 1–67.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*, 587–606.

Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(1), 152–162.

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

Grabe, W., & Stoller, F. L. (2002). *Teaching and researching reading.* Harlow, UK: Longman.

Graves, M. F. (1986). Vocabulary learning and instruction. *Review of Research in Education*, *13*, 49–89.

Griffin, G. F. (1992). Aspects of the psychology of second language vocabulary list learning. Unpublished PhD thesis, Dept of Psychology, University of Warwick.

Groot, P. J. M. (2000). Computer assisted second language vocabulary acquisition. *Language Learning & Technology*, *4*(1), 60–81.

Higa, M. (1965). The psycholinguistic concept of "difficulty" and the teaching of foreign language vocabulary. *Language Learning*, *15*(3&4), 167–179.

Hirsh, D., & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, *8*(2), 689–696.

Horst, M., Cobb, T., & Nicolae, I. (2005). Expanding academic vocabulary with an interactive on-line database. *Language, Learning & Technology*, *9*(2), 90–111.

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

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

Hulstijn, J. H. (1993). When do foreign-language readers look up the meaning of unfamiliar words? The influence of task and learner variables. *Modern Language Journal*, *77*(2), 139–147.

Hulstijn, J. H., Hollander, M., & Greidanu, S. T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *Modern Language Journal*, *80*(3), 327–339.

Jones, L. (2004). Testing L2 vocabulary recognition and recall using pictorial and written test items. *Language, Learning & Technology*, *8*(3), 122–144.

Kintsch, W. (1970). *Learning, memory, and conceptual processes.* Wiley, New York.

Knight, S. (1994). Dictionary: The tool of last resort in foreign language reading? A new perspective. *Modern Language Journal*, *78*, 285–299.

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.* U.S. Department of Health, Education, and Welfare, Washington, D.C.: 5-l095.

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

Laufer, B. (1989). What percentage of lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking machine* (pp. 69–75). Clevedon, England: Multilingual Matters.

Laufer, B., & Shmueli, K. (1997). Memorizing new words: Does teaching have anything to do with it? *RELC Journal*, *28*(1), 89–108.

Laufer, B. & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, *22*(1), 1–26.

Lomicka, L. (1998). "To gloss or not to gloss": An investigation of reading comprehension online. *Language Learning & Technology*, *1*(2), 41–50. Retrieved April 16, 2007, from http://llt.msu.edu/vol1num2/article2/default.html

Lyman-Hager, M., Davis, J. N., Burnett, J., & Chennault, R. (1993). Une Vie de Boy: Interactive reading in French. In F. L. Borchardt & E. M. T. Johnson (Eds.), *Proceedings of the CALICO 1993 Annual Symposium on "Assessment"* (pp. 93–97). Durham, NC: Duke University.

Ma, Q. (2007). From monitoring users to controlling user actions: A new perspective on the user-centred approach to CALL. *Computer Assisted Language Learning*, *20*(4), 297–321.

Mandler, G., Pearlstone, Z., & Koopmans, H. S. (1969). Effects of organization and semantic similarity on recall and recognition. *Journal of Verbal Learning and Verbal Behavior*, *8*(3), 410–423.

Mondria, J.-A., & Wit-de Boer, M. (1991). The effects of contextual richness on the guessability and the retention of words in a foreign language. *Applied Linguistics*, *12*(3), 249–267.

Mondria, J.-A., & 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.

Mishima, T. (1967). An experiment comparing five modalities of conveying meaning for the teaching of foreign language vocabulary. Dissertation Abstracts, *27*, 3030–3031A.

Nagata, N. (1999). The effectiveness of computer-assisted interactive glosses. *Foreign Language Annals*, *32*(4), 469–479.

Nagy, W. E. (1997). On the role of context in first- and second-language learning. In N. Schmitt and M. McCarthy (Eds.) *Vocabulary: Description, acquisition and pedagogy*, pp. 64–83. Cambridge: Cambridge University Press.

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

Nation, I. S. P. (1982). Beginning to learn foreign vocabulary: A review of the research. *RELC Journal*, *13*(1), 14–36.

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.

Nist, S. L., & Olejnik, S. (1995). The role of context and dictionary definitions on varying levels of word knowledge. *Reading Research Quarterly*, *30*(2), 172–193.

Overbaugh, R. C. (1994). Research based guidelines for computer based instruction development. *Journal of Research on Computing in Education*, *27*(1), 29–47.

Paivio, A. (1969). Mental imagery in associative learning and memory. *Psychological Review*, *76*, 241–263.

Paivio, A. (1994). Memory for pictures and sounds: Independence of auditory and visual codes. *Canadian Journal of Experimental Psychology*, *48*(3), 380–398.

Paribakht, T. S. & Wesche, M. (1997) *Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition.* In J. Coady and T. Huckin (Eds.) *Second language vocabulary acquisition: A rationale for pedagogy.* Cambridge University Press.

Pimsleur, P. (1967). A memory schedule. *The Modern Language Journal*, *51*(2), 73–75.

Plass, J. L. (1998). Design and Evaluation of the User Interface of Foreign Language Multimedia Software: A Cognitive Approach. *Language Learning and Technology*, *2*(1), 35–45. http://llt.msu.edu/vol2num1/article2/index.html

Plass, J. L., Chun, D. M., Mayer, R. E., & Leutner, D. (1998). Supporting visual and verbal learning preferences in a second-language multimedia learning environment. *Journal of Educational Psychology*, *90*(1), 25–36.

Schneider, V. I., Healy, A. F. & Bourne Jr., L. E. (2001). What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language*, *46*, 419–440.

Schmidt, R. (1995). Consciousness and foreign language learning: A tutorial on the role of attention and awareness. In R. Schmidt (Ed.), *Attention and awareness in foreign language teaching and learning* (Technical Report No. 9) (1–64). Honolulu: University of Hawaii at Manoa.

Seibert, L. C. (1927). An experiment in learning French vocabulary. *Journal of Educational Psychology*, *18*, 294–309.

Shiung, Y.-J., & Joseph, S. R. H. (Forthcoming) Effectiveness of L2 vocabulary instruction for intentional learning: A research synthesis and meta-analysis.

Silva, K. (2008). Second Life. *TESL-EJ* 12.1 June 2008.

Stahl, S. A., & Fairbanks, M. M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research*, *56*(1), 72–110.

Stoddard, G. D. (1929). An experiment in verbal learning. *Journal of Educational Psychology*, *20*, 452–457.

Thompson, V. A. & Paivio, A. (1994). Memory for pictures and sounds: Independence of auditory and visual codes. *Canadian Journal of Experimental Psychology*, *48*(3), 380–398.

van Bussel, F. J. J. (1994). Design rules for computer-aided learning of vocabulary items in a second language. *Computers in Human Behavior*, *10*, 63–76.

Waring, R. (1997). A comparison of the receptive and productive vocabulary sizes of some second language learners. *Immaculata* (Notre Dame Seishin University, Okayama), *1*, 53–68.

Wilcox, R. (2008). Personal Communication.

Wlodlowski, R. (1999). *Enhancing adult motivation to learn*. San Francisco: Jossey-Bass.

Yeh, Y., & Wang, C. (2003). Effects of multimedia vocabulary annotations and learning styles on vocabulary learning. *CALICO Journal*, *21*(1), 131–144.

Yip, F., & Kwan, A. (2006). Online vocabulary games as a tool for teaching and learning English vocabulary. *Educational Media International*, *43*(3), 233–249.

Yoshii, M., & Flaitz, J. (2002). Second language incidental vocabulary retention: The effect of picture and annotation types. *CALICO Journal*, *20*(1), 33–58.