ReCALL **20**(1): 3-20. 2008 © European Association for Computer Assisted Language Learning
DOI: 10.1017/S0958344008000219 Printed in the United Kingdom

English vocabulary learning with word lists, word cards and computers: implications from cognitive psychology research for optimal spaced learning

TATSUYA NAKATA

Department of Language and Information Sciences, The University of Tokyo Graduate School of Arts and Sciences, 3-8-1 Komaba, Meguro-ku Tokyo, 153-8902, Japan
(*email: howtoeigo@gmail.com*)

Abstract

The spacing effect is known to be one of the most robust phenomena in experimental psychology, and many attempts have been made to realize effective spaced learning for L2 vocabulary learning. This study compares vocabulary learning with word lists, word cards, and computers in order to identify which material leads to the most superior spaced learning. In the experiment, 226 Japanese high school students studied ten English words with one of the three learning materials: lists, cards, and computers. One-way analysis of covariance (ANCOVA) showed that although no significant difference existed between the Card group and the other two, the PC group significantly outperformed the List group on the delayed post-test. Item analysis using Chi-squares demonstrated that on the delayed post-test, the List group's successful recall rates for four of the ten items were significantly lower than those of the Card or PC group. Correlational analysis indicated that the time invested in learning and the subsequent post-test scores did not correlate significantly for the List and Card groups. Paradoxically, a negative correlation was observed between the PC group's study time and their post-test scores. The lack of meaningful relationships between the study time and subsequent retention may be partially due to the limited ability of certain learners to learn effectively while using certain materials. A questionnaire given to the participants found that, in general, computers were evaluated more favorably than lists or cards. At the same time, however, learners exhibited large variations in their evaluation of computers, implying the importance of considering individual differences when introducing CALL to learners. In summary, the study has demonstrated the superiority of computers over lists, the limited advantage of word cards over lists, and no statistically significant difference between computers and cards. The findings are significant because although the advantages of cards or computers have been advocated, no study has ever tested such claims empirically.

Keywords: vocabulary learning, spacing effect, list learning, computer-based sequencing algorithm, word list, word card

1 Introduction

Vocabulary knowledge constitutes an integral part of learners' general proficiency in a second/foreign language (L2) and is a prerequisite for successful communication (Nation, 2001). Word lists and word cards (flashcards) are among the most frequently used tools for vocabulary building. A word list is a sheet of paper where L2 words are printed along with their L1 translations or definitions. Word cards are a set of cards where the L2 word is written on one side and its L1 translation or definition on the other. Some researchers note that cards facilitate vocabulary acquisition more effectively than lists (Mondria & Mondria-de Vries, 1994; Nation, 2001; Schmitt & Schmitt, 1995; Waring, 2004). Others argue that computers are superior to lists or cards because they offer benefits that paper-based materials do not (Ellis, 1995; Hulstijn, 2001; Nation, 2001). Despite such claims, empirical evidence is still scant, and the relative effectiveness of the three types of learning materials has yet to be investigated. The present study compares vocabulary learning that employs these three types of materials and aims to empirically test the alleged efficacy of lists, cards, and computers.

2 Previous research

2.1 The neglect and reappraisal of list learning

Vocabulary learning activities are generally categorized into two types: *intentional* and *incidental* learning of vocabulary (Nation, 2001). The former refers to activities that are conducted primarily for vocabulary development. List learning, where learners associate the L2 word form with its L1 translation equivalent, falls into this category. By contrast, when vocabulary is learned as the by-product of activities that do not have a primary focus on vocabulary development, the activities are called *incidental*. Message-focused reading or listening may allow learners to learn vocabulary incidentally. In the field of CALL, numerous studies have looked into how computers can facilitate incidental vocabulary learning. In particular, the issue of what kinds of multimedia annotations facilitate lexical acquisition has attracted the attention of researchers. Studies have shown that providing multimedia annotations such as pictures or video clips can enhance retention because such additional information gives a learner multiple access routes to the word and leaves a deeper memory trace (e.g., Al-Seghayer, 2001; Chun & Plass, 1996, Yoshii, 2006).

Although incidental vocabulary learning certainly makes a significant contribution to L2 learners' lexical development (Nation, 2001), it is not without limitations. One of its major shortcomings is that it is slow and haphazard. It is estimated that in order to learn 108 words from context, ESL/EFL learners may need to read a text of 200,000 words (Laufer, 2003), which is an unrealistic expectation in most classrooms. Research has also shown that incidental learning is not a reliable source of lexical acquisition for learners at an elementary level (see Laufer, 2003, for further limitations of contextual vocabulary learning).

Because of the recent widespread use of communicative language teaching, intentional vocabulary learning activities, especially those involving memorization of vocabulary, have been relatively unpopular. Rote learning tends to be dismissed as a relic of the old-fashioned behaviorist learning model, and applied linguists have held list learning in abhorrence (Hulstijn, 2001). Empirical studies, however, seem to suggest that list learning nonetheless should have a place in L2 vocabulary instruction. Studies comparing list learning and contextual vocabulary acquisition have consistently demonstrated that the former is far more effective and efficient than the latter (Laufer, 2003; Nakata, in press; Nation, 2001; Waring, 2004; among others). Furthermore, despite a popular belief that list learning does not lead to long-term retention, studies demonstrate that the truth is quite to the contrary: vocabulary learned in lists is found to be resistant to decay and can be retained over several years (Hulstijn, 2001; Nation, 2001).

In summary, although it is often assumed that most vocabulary is learned from natural context (Krashen, 1989), vocabulary learning from context alone is not sufficient, particularly in a context where learners have little exposure to the target language outside the classroom, and should be complemented with word-focused activities such as list learning, which is more certain and efficient (Hulstijn, 2001; Laufer, 2003; Nation, 2001; Waring, 2004). With the above discussion in mind, this study sets out to investigate how list learning can be successfully implemented with the help of computers. The remainder of this literature review, therefore, limits itself to studies on vocabulary acquisition in the form of list learning and excludes those on other computer-based vocabulary learning methods such as multimedia annotations (e.g. Al-Seghayer, 2001; Chun & Plass, 1996; Yoshii, 2006), computer-based dictionaries (e.g. Hill & Laufer, 2003), or computer-mediated communication (e.g. Smith, 2004; Stockwell, 2005).

2.2 Implications from cognitive psychology for effective vocabulary learning

Put into a broader perspective, list learning can be categorized as a kind of rehearsal. Rehearsal is defined as an activity to encode new information into our long-term memory through overt or silent articulation. Unless they are rehearsed frequently, most new words will eventually be forgotten due to the fragile nature of human memory (Baddeley, 1997; Ellis, 1995; Hulstijn, 2001). In rehearsing lexical items, the distribution of rehearsal opportunities impinges on the effectiveness and efficiency of the learning activities. Scheduling of rehearsal opportunities is usually divided into two types: spaced learning and massed learning. In spaced learning, rehearsal activities for a given item are spread over a longer period of time, while in the latter they are condensed into a smaller number of sessions. For a given amount of study time, spaced learning yields significantly higher retention than massed learning, a phenomenon referred to as a spacing effect (Baddeley, 1997; Hulstijn, 2001). The effect is one of the most robust phenomena in experimental psychology (Ellis, 1995) and has been unanimously supported by numerous psychological experiments since its discovery in the 19th century (Baddeley, 1997; Hulstijn, 2001; Mizuno, 2003a).

Given that spaced learning is superior to massed learning, how should we space learning opportunities in order to obtain the largest spacing effect? The following two principles can be derived from cognitive psychology literature:

1. A successful recall from memory yields superior retention to mere presentation of the target item because the very act of retrieving information from memory strengthens retrieval routes to memory. Hence, testing one's memory to recall the L2 word form or its meaning (*retrieval practice*) is beneficial to long-term retention. This phenomenon is known as the *retrieval practice effect* (Baddeley, 1997; Ellis, 1995; Nation, 2001).

2. Research has shown that “the degree to which a recalled item is strengthened *increases* with the length of the interval preceding the test” (Ellis, 1995:121). In other words, successfully recalling the meaning of a lexical item after a delay is more effective than recalling it immediately after we learn it (Baddeley, 1997; Mizuno, 2003a).

The above two principles are contradictory: the first principle implies that intervals between tests should be short because if the interval is too long, the item will not be successfully recalled. This is not effective since only a successful recall strengthens memory. The second principle, on the other hand, recommends using longer intervals between tests. Based on the above two, it follows that we can obtain the biggest spacing effect when we test a given item at the longest delay compatible with correct recall (Baddeley, 1997; Ellis, 1995; Mizuno, 2003a).

The above discussion implies the following: when a new lexical item is introduced, it should be tested immediately. Otherwise, our memory for that item will decay to the extent that a successful recall is impossible because less consolidated memory decays relatively quickly (Baddeley, 1997; Ellis, 1995; Mizuno, 2003a). The interval between the first and the second test can be longer than that between the initial presentation and the first test since the successful recall at the first test allows the item to be remembered longer. Similarly, the intervals between tests can be gradually increased as learning proceeds. This type of schedule, which tests items with ever increasing intervals, is called *expanded rehearsal*. Expanded rehearsal is considered to be the optimal spaced learning schedule because when conducted systematically, it allows the learner to test a given item at the longest delay compatible with correct recall, which leads to the largest spacing effect at each retrieval opportunity (Baddeley, 1997; Ellis, 1995; Mizuno, 2003a; Siegel & Misselt, 1984).

2.3 Lists or cards vs. computers: from an applied linguistic perspective

In view of the reliability and significance of the spacing effects, many attempts have been made to implement optimal spaced learning in the field of L2 lexical acquisition. Previous research in this field has discussed the effectiveness of three types of materials, namely, word lists, word cards and computers. Cards are believed to enhance learning more than lists for four main reasons (Mondria & Mondria-de Vries, 1994; Nation, 2001; Schmitt & Schmitt, 1995; Waring, 2004). Firstly, expanded rehearsal can be implemented more easily with them. Learners can divide cards into several decks based on the learning states of each item, enabling them to review difficult items more frequently than easy items. Studying with a list does not offer such flexibility, and most words tend to be reviewed at similar intervals, violating the principle of expanded rehearsal. Secondly, as we have seen already, the retrieval practice effect implies that an active recall of the L2 word form or its meaning results in superior retention to mere presentation. Learners can easily practise retrievals with cards, where the L2 word and its meaning are presented on different sides. Lists, by contrast, normally expose learners to both the L2 word and its meaning at the same time, making them a less desirable material than cards. Thirdly, there is a danger that the position of the word in the list offers inappropriate help in remembering, a phenomenon known as the *list effect*. As a result, learners may have no problem retrieving memory of the item within the list, but have considerable difficulty in doing so when it is presented outside it. Lastly, when items are presented in series, some items in particular positions on the list tend to get less attention than others. Word cards are free from such problems because they offer more flexibility in the ordering of items.

Others maintain that computers offer several benefits lacking in lists and cards. Firstly, a computer-based sequencing algorithm helps the implementation of expanded rehearsal. It is true that by using cards, learners can divide items into several decks and implement expanded rehearsal more easily than with a list. However, a systematic use of word cards requires sophisticated strategies on the part of the learners. If they cannot monitor their learning accurately and plan their review schedule accordingly, they cannot make the most of word cards and may run the risk of inefficient learning, e.g. over-learning (devoting more time than necessary) of easy items or under-learning of hard items. A computer program, on the other hand, can easily keep a record of the learner's performance on individual words, control the sequencing of items, and guarantee a certain level of success irrespective of learners' skills (Ellis, 1995; Hulstijn, 2001; Nation, 2001). Secondly, retrieval practice can be easily implemented with simple programming on computers (Allum, 2004). Thirdly, as computers can vary the presentation order flexibly, learners do not have the problem resulting from the list effect. Other advantages offered by CALL may include enhanced presentation of materials due to its multimedia capabilities, a high degree of autonomy given to learners, introduction of new exercise types, or positive effects on students' motivation (Ellis, 1995; Nakata, 2006a, 2006b; Nation, 2001).

2.4 Implementation of optimal spaced learning with the help of computers

Expectations for computers to realize optimal spaced learning have led cognitive psychologists to develop a computer-controlled sequencing algorithm for effective spaced learning. Studies on an algorithm called the Low-First Spaced Learning Method, which was developed by Dr. Rika Mizuno, are among such attempts (e.g., Mizuno, 2000, 2003b, 2004). The algorithm is derived from the reactivation theory of spacing effects proposed in the realm of cognitive psychology (Mizuno, 2003a). The basic tenet of the theory is that in spaced learning, memory reactivation during the subsequent retrieval trial is larger than in massed learning and that it is these changes in memory reactivation that cause the spacing effects. Based on the predictions made from the theory and experimental results, Mizuno formulated equations to express changes in memory activity and a relationship between memory reactivation and the probability of recall. Based on such models, the Low-First Method generates a review schedule so that the largest spacing effect can be obtained (see Mizuno, 2003b, for further details about the Low-First Method).

Although some sequencing algorithms other than the Low-First Method have been proposed (e.g. Atkinson, 1972; Siegel & Misselt, 1984; van Bussel, 1994), the Low-First Method appears to be more promising than some earlier attempts for several reasons. Firstly, the method is based on sound principles for optimal spaced learning derived from the reactivation theory, the first theory to offer a coherent explanation for the cause of the spacing effect (Mizuno, 2003a). Secondly, the effectiveness of the method has been proved consistently by a number of experiments (Mizuno, 2000, 2003b, 2004; Nakata, in press), although studies on other algorithms have produced mixed results. Lastly, continual attempts have been made to improve the effectiveness of the algorithm in the last decade (Mizuno, 2000, 2003b, 2004; Nakata, 2006a, 2006b) while most studies on other algorithms were conducted at least a decade ago without replication studies in recent years.

2.5 Limitations of the previous studies

In summary, the question of how to implement optimal spaced learning has been attracting attention from many researchers, presumably due to the robustness of the spacing effects. From a pedagogical perspective, some of the advantages of cards and computers have been highlighted. Attempts have also been made to model changes in memory activity and develop a computer-based sequencing algorithm to realize effective spaced learning. One limitation of the previous research may be that no study has empirically investigated the relative effectiveness of lists, cards, and computers. Although researchers have frequently claimed the superiority of cards over lists or the advantage of computers over the paper-and-pencil approach, such arguments lack empirical support and remain speculative. With the limitation of the previous research in mind, the present study compares L2 vocabulary learning with lists, cards, and computers to expose the alleged efficacy of the three types of materials to empirical testing.

In light of the growing body of research on vocabulary acquisition with multimedia annotations (e.g. Al-Seghayer, 2001; Chun & Plass, 1996, Yoshii, 2006), computer-based dictionaries (e.g. Hill & Laufer, 2003), or computer-mediated communication (e.g. Smith, 2004; Stockwell, 2005), it might be of interest to compare the effectiveness of computer-controlled list learning with that of other kinds of computer-based vocabulary learning methods. However, the author chooses not to make such a comparison for two reasons. First, as has already been discussed, studies comparing list learning and incidental learning from context have consistently demonstrated that the former is far more effective than the latter (Laufer, 2003; Nakata, in press; Nation, 2001; Waring, 2004). Even with the help of multimedia annotations, it does not seem very likely that incidental vocabulary learning through computers can outperform computer-based list learning. Second, vocabulary learning using lists or cards is a very common strategy among L2 learners (Schmitt, 1997). The issue of whether computers can better assist learners in implementing effective list learning than the paper-and-pencil approach is therefore expected to offer valuable pedagogical implications for learners.

3 Hypotheses

Based on arguments gleaned from earlier research, the following three hypotheses are formed concerning the relative effectiveness of lists, cards, and computers for L2 vocabulary learning:

Hypothesis I: The use of computers will result in superior retention to lists.
Hypothesis II: The use of cards will result in superior retention to lists.
Hypothesis III: The use of computers will result in superior retention to cards.

4 Method

4.1 Participants

Participants were 226 first and second-year male students, aged 15-17, at a senior high school in Tokyo, Japan. Their L1 was Japanese and L2 was English. The subjects were assigned to one of the following three conditions: (1) the List group, (2) the Card group, and (3) the PC group, taking care that there would be no significant difference in scores on the GTEC for Students, a nationwide English proficiency test, among the three groups (see section 5 Results).

4.2 Materials

The List group was provided with a word list (30 cm × 21 cm) that had ten English low-frequency nouns along with their Japanese translations. The items were listed in the order of *grig, toil, saliva, loach, antic, sentry, cavity, mane, dike,* and *debris*. These target items were chosen from several word lists such as Nation's (2001). The Card group studied with ten individual flash cards (6 cm × 10.5 cm), each of which was printed with one of the ten target words on one side and its Japanese translation on the other.

The subjects in the PC group studied with computer software programmed by the author to replicate the Low-First Method. The participants were presented a Japanese translation as a cue and requested to type the corresponding English word in the answer box (Figure 1). When a correct response was typed, the program displayed the message "Good!" and proceeded to the next item. To a blank or incorrect response, the correct English word was provided as corrective feedback. Sequencing of items was controlled by the Low-First Method (Mizuno, 2003b). The study session ended when all the ten items reached their pre-determined retirement criterion determined by the algorithm.

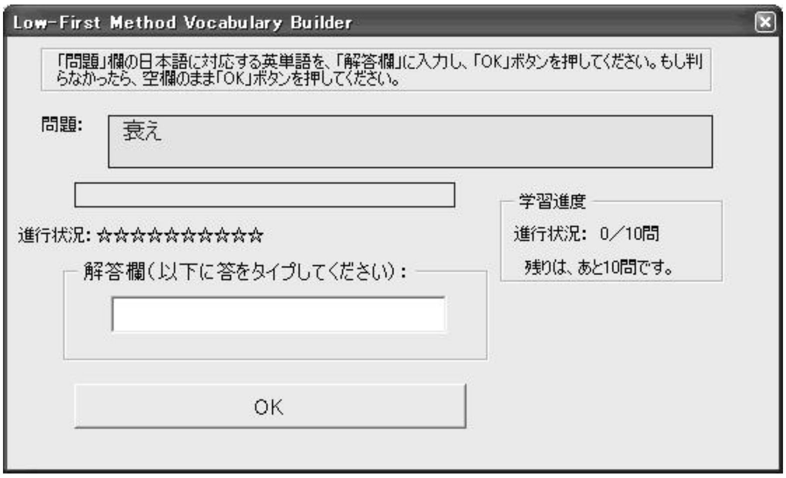

Fig. 1. A screenshot of the computer program used by the PC group.

It should be noted here that the program used in the present experiment exploited only one of the apparent advantages of computer-based instruction, i.e. a computer-controlled sequencing algorithm. Other potential benefits of CALL such as multimedia capabilities or new exercise types were not introduced. Since the purpose of this study was to explore the effectiveness of lists, cards, and computers for list learning, it was necessary to make the PC group's learning condition as similar as possible to that of the List and Card groups and assure comparability between the three groups.

4.3 Procedure

The procedure for the List and Card groups is given below.

(i) Instruction
All the details of the experiment were explained with the exception of the delayed post-test, which was to be administered without prior notice in order to reduce the likelihood of intervening study.

(ii) Pre-test
The subjects took a pre-test that requested them to translate the ten target words into Japanese.

(iii) Study session
The students were provided with their learning materials: a word list or ten individual flash cards. The learners in both the groups were instructed to remember the target English words on their list or cards so that they could recall the English words from Japanese translations without any misspelling. They were also given a scratch paper (30 cm $\times$ 42 cm) and allowed to write freely on it. When they felt that they had mastered all the ten words, they recorded their study time on their scratch paper and proceeded to the immediate post-test individually.

(iv) Immediate post-test
The subjects were requested to recall the ten English target words from Japanese translations.

(v) Questionnaire
The learners were requested to fill in a questionnaire which included questions regarding the experiment and their own background.

(vi) Delayed post-test
Four days after the study session, the participants took an unannounced delayed post-test. The test was the same as the immediate post-test except for the order of the items. In scoring the immediate and delayed post-tests, any misspelling was regarded as an incorrect answer, and no partial credit was given.

The procedure for the PC group was the same as that for the other two groups, with three exceptions. Firstly, all the experimental procedure was conducted with a computer. Secondly, the PC group was not provided with any scratch paper, unlike the List or Card groups. This was done to isolate the learning effects of the computer program from those of the paper-and-pencil approach. Lastly, while the subjects in the List and Card groups were instructed to continue studying until they felt that they had mastered all the target words, the computer program automatically terminated the study session for the PC group when all the ten items reached the retirement criterion determined by the Low-First Method.

5 Results

5.1 Effectiveness of lists, cards, and computers

Nine out of the 226 participants were excluded from the analysis because they exhibited prior knowledge of target words on the pre-test. For the remaining 217 students, an analysis of variance (ANOVA) test was conducted to examine possible differences in the three groups' English proficiency. The test indicated that differences in GTEC scores among the three groups were not statistically significant, $F(2,214) = 0.03$, *n.s*. Since their English proficiency and prior knowledge of target words were controlled, it will be assumed that the different outcomes obtained for the three groups resulted solely from differences in their materials. A close examination of the data revealed that there were five outliers whose study time was greater than two standard deviations above the mean. These five were excluded from the following statistical analysis because their presence invalidates the use of a parametric ANOVA test. The following table summarizes the average study time and post-test scores for the three groups.

A brief look at the table tells us that the PC group yielded the highest scores among the three, but at the same time, their study time was the longest. ANOVA revealed a significant difference among the three groups in their average study time, $F(2,195) = 12.31$, $p<.001$. Since the PC group's highest post-test scores might have resulted from the fact that their study time was the longest, one-way analysis of covariance (ANCOVA) will be used instead of ANOVA for the following analyses, thus controlling for the differences in the three groups' study time. ANCOVA with study time as a covariate found that while there was no significant difference in their immediate post-test scores among the three groups, $F(2, 194) = 0.39$, n.s., differences in their delayed post-test scores were statistically significant, $F(2, 180) = 3.23$, $p<.05$. Bonferroni post

Table 1 *Means and standard deviations by group (the study time and post-test scores)*[1]

	List			Card			PC		
	n	*M*	*SD*	*n*	*M*	*SD*	n	M	SD
Study time	61	366.05[a]	157.63	70	426.77[ab]	146.62	67	509.75[b]	187.59
Immediate	71	8.20[a]	2.19	74	8.31[a]	2.17	67	8.43[a]	1.84
Delayed	63	2.61[a]	2.21	71	3.20[ab]	2.24	62	3.82[b]	2.47

Note. The possible maximum score is 10 for the immediate and delayed post-tests. The figures marked with different letters are significantly different according to a Bonferroni post hoc test ($p<.05$).

1. The *ns* for the study time and post-tests do not match because (1) ten learners in the List group and four learners in the Card group failed to record their study time, (2) six participants did not take the delayed post-test, and (3) ten learners reported that they had studied some of the target words between the study session and the delayed post-test, and they were excluded from the analysis of the delayed post-test.

hoc tests confirmed that the PC group scored significantly higher on the delayed post-test than the List group (p<.05). No significant difference was detected between the Card and the other two groups. Since the effectiveness of vocabulary learning activities should be measured by long-term retention rather than short-term retention (Ellis, 1995), we will use the delayed post-test scores as an index of learning outcomes throughout the following discussion. The results of the experiment demonstrated that, as predicted by the hypotheses, the PC group led to the highest retention, followed by the Card and List groups. However, we need to treat the results with caution due to the lack of statistical significant difference between the Card group and the other two.

5.2 Item analysis

The three groups' successful recall rates on individual items were compared with Chi-squares to ascertain whether all the ten items were learned equally successfully by the three groups. Only the results of the delayed post-test will be analyzed because a significant difference between the groups existed only in the average scores of the delayed post-test.

In Table 2, the items were sorted in the order in which they were placed in the list given to the List group (see section 4.2 Materials). Chi-squares detected significant differences in the recall rates for the second, third, fourth, and eighth items on the list. Ryan's method for multiple comparisons has revealed that a significantly higher proportion of PC learners correctly answered three items (*toil, loach, mane*) compared with the List group. The Card group significantly outperformed the List group on two items (*saliva, mane*) and the PC group on one item (*saliva*). In short, the List group's successful recall rates were significantly lower for four of the 10 items than the PC or Card group. The results appear to corroborate the observation that due to the fixed presentation order, some items on the list tend to get less attention than others (Baddeley, 1997; Mondria & Mondria-de Vries, 1994; Nation, 2001). One might reasonably hypothesize that cards and computers, by contrast, allowed learners to

Table 2. *Comparisons of successful recall rates by item between groups (the delayed post-test)*

	List	Card	PC	x2
grig	57.1%	69.0%	69.4%	2.73
toil	15.9%[a]	25.4%[ab]	40.3%[b]	9.63**
saliva	63.5%[a]	81.7%[b]	54.8%[a]	11.48**
loach	36.5%[a]	50.7%[ab]	58.1%[b]	6.04*
antic	22.2%	35.2%	33.9%	3.11
sentry	30.2%	25.4%	35.5%	1.62
cavity	14.3%	19.7%	16.1%	0.74
mane	31.8%[a]	56.3%[b]	58.1%[b]	11.07*
dike	25.4%	31.0%	41.9%	4.03
debris	15.9%	16.9%	14.5%	0.14

Note. **p<.01, *p<.05. df = 2 for all the items. The figures marked with different letters are significantly different according to Ryan's method for multiple comparisons (p<.05).

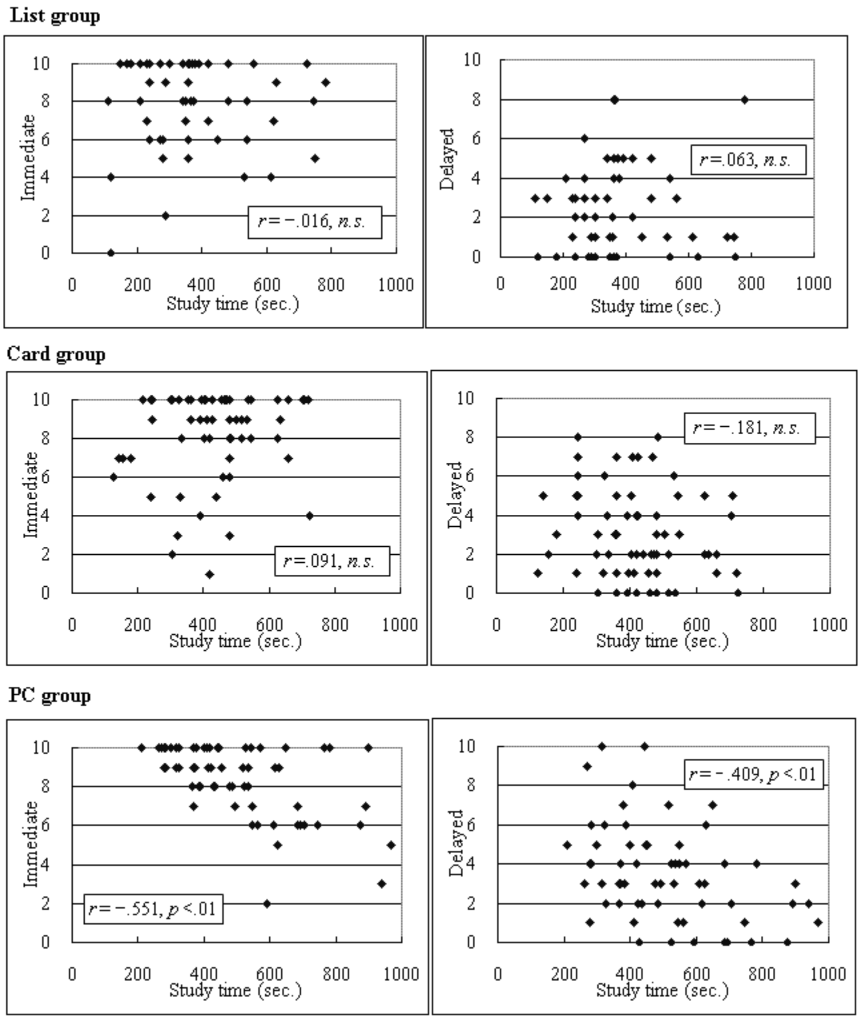

Fig. 2. Relationships between the study time and post-test scores.

acquire all the items equally well because these two materials are not vulnerable to the effects of a fixed presentation order.

5.3 Relationships between the study time and post-test scores

In order to investigate possible relationships between the participants' study time and their subsequent retention, the study time and post-test scores were plotted (Figure 2) and the correlation coefficients between them were calculated. Analysis showed that the study time and subsequent post-test scores did not correlate significantly either for the List (immediate: -.016, delayed: .063) or Card group (immediate: .091, delayed: -.181). Unexpectedly, a moderate, negative correlation was detected for the PC group (immediate: -.551, delayed: -.409, both $p<.01$). In other words, the longer study time was not necessarily associated with higher post-test scores for the List and Card groups.

Paradoxically, the longer the PC group studied, the lower the score obtained. The implications of these finding will be discussed in section 6 Discussion.

6 Discussion

6.1 Effectiveness of lists, cards and computers

The results of the experiment seemed to constitute support for Hypothesis I. However, Hypothesis II was supported only partially and Hypothesis III was rejected. This section will discuss the effectiveness of the three learning materials based on the results of the experiment. Firstly, the PC group scored significantly higher than the List group on the delayed post-test, supporting Hypothesis I. The result seems consistent with the argument put forward by earlier research that computers are more conducive to learning than lists. It is speculated that the PC group attained a relatively high word gain because the computer algorithm organized the learning session for them and guaranteed a certain level of success. The List group might have yielded a significantly lower score than the PC group as it is difficult to implement expanded rehearsal or retrieval practice with lists. Item analysis further confirmed that the computer led to a significantly higher score than the list for three items on the delayed post-test. It can be inferred that the List group was affected by the fixed presentation order, which apparently led to under-learning of some items at particular positions on the list.

The results of the study offered only partial support to Hypothesis II. The cards brought about a significantly higher score on two items in the delayed post-test. Yet, no significant difference emerged in the average total scores either on the immediate or delayed post-test, suggesting that the advantage of cards over lists is limited. The results seem inconsistent with the observation that cards lead to more effective learning than lists. There appear to be two reasons for this discrepancy. Firstly, the results might be in part attributed to limitations regarding the learners' abilities to use cards effectively. For example, in order to implement expanded rehearsal with cards, the learners have to be able to evaluate the learning states of each word accurately and plan their review schedule so that the interval between tests increases as learning proceeds. This will require extensive meta-cognitive abilities on the part of the learners, and if they lack such abilities, cards may prove no more effective than lists. This explanation seems plausible if we take account of the fact that the subjects in this study were high school students, who had probably not developed such abilities. In fact, Schmitt's study (1997) on vocabulary learning strategies revealed that only 29% of Japanese high school students used word cards while lists were used by 67% of them. It is likely that the subjects of the study, who were also Japanese high school students, could not take full advantage of benefits offered by cards due to lack of experience. Secondly, the duration of this experiment may have been too short for cards to outperform lists. It has been asserted that cards are more desirable than lists in that the former can be easily divided into several decks, thus enabling learners to concentrate on difficult items (Mondria & Mondria-de Vries, 1994; Schmitt & Schmitt, 1995; Waring, 2004). Such an advantage of cards may turn out to be significant when students learn a large number of items over a long period of time. In the present study, where the subjects learned only ten items for no more than ten minutes, it is possible that such differences did not play a crucial role.

Further research may be necessary to ascertain whether cards rather than lists can better assist learners in managing retrieval opportunities distributed over a longer period of time.

The PC group obtained a slightly higher score than the Card group on both the immediate and delayed post-tests. However, the differences did not reach statistical significance, and Hypothesis III was rejected. The item analysis demonstrated that the Card group scored significantly higher than the PC group on one item in the delayed post-test, which is contrary to our prediction. Therefore, the results of the study were inconsistent with the hypothesis. The discrepancy might be partially due to differences in the learners' familiarity with learning materials. The questionnaire given after the immediate post-test showed that none of the 67 learners in the PC group studied English vocabulary with computers regularly: when asked whether they studied English vocabulary using computers, 49 chose "Never," 13 replied "Rarely," and 5 chose "Occasionally." If learners had been as accustomed to CALL as paper-and-pencil learning, the PC group might have attained better retention. In particular, as the software used in the present experiment required extensive use of the keyboard, the study session might have been too demanding or demotivating for high school students, most of whom were probably not very accustomed to typing in English. Readers should also be reminded that the program used in the present study did not take advantage of several potential benefits offered by computers in order to approximate the PC group's learning environment to that of the List and Card groups. If we add features to the computer program used in the study so as to exploit the potential of computers to the fullest (e.g. multimedia capabilities or new types of exercises), the differences between computers and cards might be large enough to be statistically significant. The potential of word card use, on the other hand, seems fairly limited in that it does not offer as much room for improvement as the use of computers.

Table 3. *Learners' evaluation of lists, cards, and computers*

		List		Card		PC	
Positive	Enjoyed learning	1	3.0%	3	7.5%	14	29.8%
	Found the learning tool effective	0	0.0%	3	7.5%	14	29.8%
	Had no difficulty in learning	6	18.2%	11	27.5%	5	10.6%
	Other reasons	3	9.1%	0	0.0%	0	0.0%
	Total	**10**	**30.3%**	**17**	**42.5%**	**33**	**70.2%**
Negative	Had difficulty in learning	8	24.2%	9	22.5%	8	17.0%
	Learning was boring	8	24.2%	9	22.5%	4	8.5%
	Learning was tiring	3	9.1%	2	5.0%	4	8.5%
	Skeptical about the effectiveness of the material	0	0.0%	0	0.0%	3	6.4%
	Other reasons	1	3.0%	0	0.0%	0	0.0%
	Total	**20**	**60.6%**	**20**	**50.0%**	**19**	**40.4%**
Etc.	Others	4	12.1%	3	7.5%	1	2.1%

Note. The sum of percentages may be larger than 100 since some learners gave multiple comments.

6.2 Relationships between the study time and post-test scores

The List and Card groups' study time and their subsequent post-test scores were found not to correlate significantly. The lack of meaningful relationships between the time invested in learning and the post-test scores may imply that most List and PC learners could not monitor their learning properly and devoted more time than necessary to the study session. The presence of such learners who over-learned some items might explain why longer study time did not necessarily bring about better test scores for these learners.

Unexpectedly, the PC group's study time correlated negatively with their immediate and delayed post-test scores. One explanation for this paradox may be possible variations in the learners' familiarity with computers. It is probable that learners who lacked computer skills required more time to complete the study session than average students. They also might have ended up with relatively low scores since their lack of familiarity with computers was costly in terms of their cognitive resources and, therefore, inhibited successful learning. Computer-literate students, by contrast, might have finished the study session more quickly and attained a higher word gain because extensive typing with the keyboard did not present particular difficulty to them, allowing them to allocate more cognitive resources to memorizing. The negative correlation between the PC group's study time and their post-test scores might be in part ascribed to the presence of learners with different degrees of PC skills. Of course, this analysis is only speculative, and further research needs to investigate this issue further.

7 Qualitative analysis

In an attempt to triangulate analysis of post-test scores, which are objective and quantitative, learners' evaluation of the lists, cards, and computers were elicited with a questionnaire administered after the immediate post-test. The learners were requested to give comments on the learning session, and 33 List, 40 Card, and 47 PC learners responded. The following table presents analysis of their responses.

The analysis showed that while nearly 70% of the PC learners expressed satisfaction with the study session, no more than 30% of the List and 40% of the Card learners did so. The learners' impressions generally seem to be in line with the quantitative analysis of the post-test scores, where the PC group yielded the highest scores and the List group the lowest, with the Card group between the two.

Close analysis of the learners' comments reveals a number of interesting observations. Firstly, it was found that PC learners (29.8%) were more inclined to find their learning experience enjoyable than the List (3.0%) or Card group (7.5%). A lower proportion of PC learners (8.5%) also rated the study session as boring compared with the List (24.2%) or Card learners (22.5%). These results may corroborate the view that use of computers has positive effects on students' motivation (Nakata, 2006a, 2006b). Although research has demonstrated that they are very effective and efficient, word-focused activities such as list learning are often perceived as "at best boring, and … at worst painful" (Krashen, 1989:450). The following comments from the PC group are encouraging as they might imply that boring and painful rote-learning could be transformed into an enjoyable experience with the help of computers (learners' comments were originally written in Japanese and translated into English by the author).

"I enjoyed it because it was like playing a video game."
"I enjoyed it very much. If this program is released as free software, I'd like to download it!"
"The program was very helpful. I want to use this at home!"

One caveat to be considered, though, is the novelty effect. It is possible that the PC group evaluated computers favorably just because they were new to them. Whether computers will have a positive effect on learners' motivation in the long term requires further research.

Secondly, although computers were evaluated more favorably than paper-based media in general, learners' responses seem to warn us against total reliance on computers. Eight PC learners (17.0%) wrote that they had experienced difficulty in learning. Three of them mentioned that their lack of familiarity with PCs caused them trouble. The other five noted that learning would have become more efficient if they had been allowed to write. In this study, the PC group was prohibited from writing on paper so that we could isolate the learning effects of the computer program from those of paper-based materials. From a pedagogical point of view, however, giving access to both paper and computers might make learners feel more comfortable and result in improved retention (Allum, 2004).

Thirdly, while none of the List or Card learners did so, three PC learners (6.4%) expressed skepticism about the effectiveness of their learning medium as follows:

"I could memorize words quickly, but may forget them soon."
"The program helped me to memorize, but I doubt long-term retention."
"I learned the words quickly, but I'm afraid I will forget them quickly too."

Interestingly, these learners expressed reservations about the efficacy of the computer program while at the same time admitting that they had no difficulty in learning. The quantitative analysis of the post-test scores did not necessarily verify their impression as the PC group was not inferior to either the List or Card group; they even outperformed the List group on the delayed post-test four days after the study session, implying that computers may be actually more beneficial to long-term retention than the traditional paper-and-pencil approach. Nonetheless, we may need to take these learners' concerns seriously since forcing learners to study with a material that they feel uncomfortable with might exert a negative influence on their motivation. One way to help ease such skepticism would be to explain the theoretical justification for the program such as the retrieval practice effect or expanded rehearsal. Sharing with learners an understanding of how the program is designed to contribute to long-term retention may help gain their trust with respect to the material and eventually lead to enhanced learning. Considering that the computer program was a completely new learning medium for most learners, it is understandable that some learners were reluctant to accept the software and showed an inclination toward more familiar paper-and-pencil materials.

The above three findings seem to indicate large individual differences in learners' perceptions of the computer program. Some participants apparently found the software attractive and wanted to use it at home. Others were demotivated due to their lack of experience with computers. Still others raised doubts about its effectiveness. The observed variations in PC learners' perceptions of the computer program seem to underscore the importance of considering individual differences in introducing CALL into classrooms.

In stark contrast to the PC group, the List and Card groups rarely expressed satisfaction or dissatisfaction with their learning materials. For example, 18.2% of the List and 27.5% of the Card learners noted that they were satisfied with the study session because they successfully learned the target items without difficulty. However, instead of giving credit to lists or cards, all of them attributed their success to factors that are not directly related to the effectiveness of their learning media such as the easiness of the words (e.g. "The words were easy to learn because they were short") or their effort ("I think I could learn words in a short time because I concentrated"). Similarly, the List and Card learners who had some difficulty in learning (24.2% in the List group and 22.5% in the Card group) did not put blame on their materials; they tended to cite factors such as difficulty of target words as the main source of their trouble (e.g. "Words were difficult to memorize because they were totally unfamiliar to me" or "It was difficult because I didn't know how to pronounce words"). The differences between the PC group and these two groups probably arose from differences in their familiarity with the learning media. One might speculate that most List and Card learners did not dare to mention their materials because paper-based materials were already common to them, and they had taken it for granted that they would be used for learning English words. By contrast, it seems natural for the PC group to evaluate the effectiveness of the software, a completely new tool, in comparison to the more familiar paper-based materials.

8 Conclusion

The present study has demonstrated the superiority of a computer-based sequencing algorithm over lists, the limited advantage of word cards over lists, and no statistically significant difference between the computers and cards in implementing vocabulary learning. Being the first attempt to evaluate the efficacy of the three types of materials, this study offers valuable implications for effective vocabulary learning. Most importantly, the present experiment lends empirical support to the observation that studying with a list entails a number of problems and should be discouraged. It would be advisable to use computers or cards whenever resources permit. This implication is particularly significant considering that studying with lists is more than twice as common as studying with cards among L2 learners (Schmitt, 1997). A computer-sequencing algorithm such as the Low-First Method is recommended because it yields significantly higher retention than lists and may also exert a positive effect on learners' motivation. One possible drawback of CALL, however, is that not all learners are equally fond of or good at it. Word cards may serve as a convenient alternative to CALL, especially for those who do not feel comfortable with computers. Unfortunately, inconsistent with the claim made by previous researchers, the present study found that the advantage of cards over lists was limited. Yet, it may be too early to conclude that cards are no better than lists as instructing learners on strategies for effective spaced learning might help them to make the most of cards and eventually lead to significantly improved retention compared with lists. Even if learners mastered such strategies, putting optimal spaced learning into practice would be difficult with word lists, considering the lack of flexibility in learning that they offer.

The findings of this study are significant because although many researchers have advocated the relative effectiveness of lists, cards, and computers for effective spaced learning, no study has ever empirically investigated the relative efficacy of these three materials. At the same time, more research will be needed to address some limitations of the present study. Firstly, the present study has addressed only limited aspects of vocabulary knowledge, namely, the written form and the form-meaning connection. It might be valuable to look into how other aspects of word knowledge such as collocations or grammatical functions can be learned with lists, cards, and computers. The second limitation would be the small number of target items used in the study. Future research may need to test a larger number of vocabulary items for statistical reliability. Lastly, further research may need to explore what kinds of influence learners' individual differences, such as their proficiency levels, computer skills, or learning styles, have on the effectiveness of the three types of learning materials. Considering that the spacing effect is too powerful to be ignored by any instructional program (Ellis, 1995), these attempts to realize effective spaced learning will offer valuable implications for L2 vocabulary teaching and learning.

Acknowledgements

My special acknowledgment goes to my supervisor, Professor Hideo Oka, who has continuously offered me invaluable advice and warm encouragement throughout this research project. Special thanks are also due to two anonymous reviewers, also to Dr. Rika Mizuno, Dr. Hywel Evans, and Mr. Masashi Hasegawa.

References

Al-Seghayer, K. (2001) The effect of multimedia annotation modes on L2 vocabulary acquisition: a comparative study. *Language Learning & Technology*, **5**(1): 202–232.

Allum, P. (2004) Evaluation of CALL: initial vocabulary learning. *ReCALL*, **16**(2): 488–501.

Atkinson, R. C. (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* (revised ed.). East Sussex: Psychology Press.

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

Ellis, N. C. (1995) The psychology of foreign language vocabulary acquisition: implications for CALL. *Computer Assisted Language Learning*, **8**(2): 103–128.

Hill, M. and Laufer, B. (2003) Type of task, time-on-task and electronic dictionaries in incidental vocabulary acquisition. *International Review of Applied Linguistics in Language Teaching,* **41**(2): 87–106.

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

Krashen, S. (1989) We acquire vocabulary and spelling by reading: additional evidence for the Input Hypothesis. *The Modern Language Journal,* **73**(4): 440–464.

Laufer, B. (2003) Vocabulary acquisition in a second language: do learners really acquire most vocabulary by reading? some empirical evidence. *Canadian Modern Language Review,* **59**(4): 567–587.

Mizuno, R. (2000) A test of the effectiveness of the Low-First Spaced Learning Method applied to CAI. *Educational Technology Research,* **23**(1/2): 35–44.

Mizuno, R. (2003a) *Gakusyukoukano ninchishinrigaku* [The cognitive psychology of learning effects]. Kyoto: Nakanishiya Publishing.

Mizuno, R. (2003b) The application of cognitive scientific findings related to spacing effects to web-based CAI systems. In: Hesse, F. W. and Tamura, Y. (eds.) *Proceedings of the Joint Workshop of Cognition and Learning through Media-Communication for Advanced e-Learning*. Berlin: Japanese-German Center Berlin, 128–133.

Mizuno, R. (2004) The effect of mastery feedback appended to a CAI system for spaced learning. *Educational Technology Research,* **27**(1/2): 31-38.

Mondria, J.-A. and Mondria-de Vries, S. (1994) Efficiently memorizing words with the help of word cards and 'hand computer': theory and applications. *System*, **22**(1): 47–57.

Nakata, T. (2006a) English vocabulary learning based on the reactivation theory of spacing effects. *Language and Information Sciences*, **4**: 229–242.

Nakata, T. (2006b) Implementing optimal spaced learning for English vocabulary learning: towards improvement of the Low-First Method derived from the reactivation theory. *The JALT CALL Journal,* **2**(2): 19–36.

Nakata, T. (in press) English collocation learning through meaning-focused and form-focused tasks. *Bulletin of Foreign Language Teaching Association,* **11**.

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

Schmitt, N. (1997) Vocabulary learning strategies. In: Schmitt, N. and McCarthy, M. (eds.) *Vocabulary: description, acquisition and pedagogy.* Cambridge: Cambridge University Press, 199–227.

Schmitt, N. and Schmitt, D. (1995) Vocabulary notebooks: theoretical underpinnings and practical suggestions. *ELT Journal,* **49**(2): 133–143.

Siegel, M. A. and Misselt, A. L. (1984) Adaptive feedback and review paradigm for computer-based drills. *Journal of Educational Psychology*, **76**(2): 310–317.

Smith, B. (2004) Computer-mediated negotiated interaction and lexical acquisition. *Studies in Second Language Acquisition,* **26**(3): 365–398.

Stockwell, G. (2005) Syntactical and lexical development in NNS-NNS asynchronous CMC. The *JALT CALL Journal*, **1**(3): 33–49.

van Bussel, F. J. J. (1994) Design rules for computer-aided learning of vocabulary items in a second language. *Computers in Human Behavior,* **10**(1): 63–76.

Waring, R. (2004) In defence of learning words in word pairs: but only when doing it the 'right' way! (http://www1.harenet.ne.jp/~waring/vocab/principles/systematic_learning.htm)

Yoshii, M. (2006) L1 and L2 glosses: their effects on incidental vocabulary learning. *Language Learning and Technology,* **10**(3): 85–101.