REHEARSING L2 ACADEMIC VOCABULARY WITH CLOZE EXERCISES: A COMPUTER-ASSISTED LANGUAGE LEARNING INTERVENTION

by

William C Price III

B. A. in Linguistics, Cornell University, 2009

Submitted to the Graduate Faculty of the Kenneth P.

Dietrich School of Arts and Sciences in

partial fulfillment of the requirements for

the degree of Master of Arts in Applied Linguistics

University of Pittsburgh

2011

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This thesis was presented

by

William C. Price III

It was defended on

December 5, 2011

and approved by

Alan Juffs, PhD, Associate Professor of Linguistics, University of Pittsburgh

Yasuhiro Shirai, PhD, Associate Professor of Linguistics, University of Pittsburgh

Philip I. Pavlik Jr., PhD, Assistant Professor of Psychology, University of Memphis

Thesis Advisor: Alan Juffs, PhD, Associate Professor of Linguistics, University of Pittsburgh

REHEARSING L2 ACADEMIC VOCABULARY WITH CLOZE EXERCISES: A COMPUTER-ASSISTED LANGUAGE LEARNING INTERVENTION

William C. Price III, M.A.

University of Pittsburgh, 2011

Choosing appropriate methods and levels of scaffolding (see, e.g., Wood, Bruner, & Ross, 1976) is a crucial skill in second language instruction. The observation that too little or too much scaffolding for a task leads to an inferior learning outcome, known as the assistance dilemma (Koedinger & Aleven, 2007), has resisted quantitative analysis. However, it is now possible to take advantage of computerized tutors' ability to precisely measure response latencies and accuracy rates to provide quantitative data to analyze the merits of different methods of scaffolding with regard to students' performance on individual tasks. The present study describes a computer-aided language learning intervention in which 46 intermediate-level adult ESL speakers used a web-based vocabulary rehearsal program several times over the course of nine weeks. The tutor led participants in completing cloze exercises of the target words, with half of the exercises being presented with a hint in the form of a short definition of the target word and half of the exercises being presented without a hint. The results of the experiment indicate that the presence of the hint significantly increased participants' accuracy on the task, but also significantly increased time on task. These results suggest that the form of support selected was an appropriate scaffold. However, L1 speakers of Arabic (N = 29) proved exceptional in a few ways: they expressed negative attitudes toward L2 writing tasks in general and did not show any

increase in accuracy in the scaffolded condition, despite the fact that speakers of other L1s showed a very large and statistically significant improvement in accuracy in that condition. These issues may relate to Arabic speakers' exceptional difficulties processing English orthography (Martin, 2011) and warrant future study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I wish to thank Dr. Alan Juffs, Dr. Philip Pavlik, and Dr. Yasuhiro Shirai for their advice, assistance, and thoughtful input throughout the duration of this project. I would especially like to thank Dr. Pavlik for lending the use of the FaCT System for this project, his tireless assistance in the software development phase of this project, and his assistance with statistical analyses. Any abuses or misconfigurations of the software—or misunderstandings of its theoretical underpinnings—reflected in this project are, of course, my own fault!

In addition, I would like to thank the administration of the English Language Institute, especially Christine O'Neill, Dawn McCormick, Dorolyn Smith, and Greg Mizera, for their cooperation and assistance in determining how best to fit this vocabulary tutor into the curriculum of the ELI. Of course, I also owe thanks to the teachers of the Writing 4 course, Carol Harmatz, Kimmy Rehak, Peter Kolenich, and Timmy Podnar, for lending me class time, relaying questions and problems to me, and reminding students of their rehearsal homework.

I would also like to thank the other graduate students of the University of Pittsburgh's Linguistics department for their comments, suggestions, and moral support. I would particularly like to thank Mary Lou Vercellotti for her suggestion that I try using Mechanical Turk to elicit possible cloze exercise sentences and Katherine Martin for her insights into the difficulties Arabic speakers experience processing English orthography.

To the participants of this study, I wish to thank those who found the tutor useful and apologize to those who chose to stop using it. One of the inevitable issues of materials development are the “growing pains” of implementing and refining a new intervention, and I accept responsibility for the high level of attrition in this project. Hopefully future versions of this type of tutor will be more engaging and helpful!

On a more personal note, I would like to thank Gary Wolf for inspiring me to begin my journey in the field of CALL, Dr. Barbara Lust for supporting my initial foray into the field as an undergraduate, and Alex Buzick and Fred Rogers for inspiring me not to abandon my interest.

Finally, I would like to thank my wife, Sally Kim, just for being who she is and for walking alongside me on this strange journey. We are both graduate students aspiring to obtain our PhDs, but what we have already achieved transcends careers and ambitions.

1.0 INTRODUCTION

The present experiment was an in-vivo instructional intervention which employed a certain type of computer-assisted language learning software to guide adult ESL students from several language backgrounds in completing a one-step, retrieval-based rehearsal activity—namely, cloze exercise completion—to practice academic vocabulary in English. The level of support provided to participants in the completion of these cloze exercises was a factor manipulated in this experiment. The significance and justification of these elements of the study are explored in depth in the sections below.

Multiple dependent measures were analyzed, including pre-test data, post-test data, and latency and rehearsal data from the rehearsal software. Results show that providing support to participants significantly increased one measure of latency in completing the task, but also significantly increased participants' accuracy on the rehearsal task. Further analysis demonstrated that the participants who were native speakers of Arabic differed from all other participants in that the Arabic speakers demonstrated a significantly smaller increase in accuracy when provided support compared to participants from other language backgrounds. These findings have implications for computer-assisted language learning in general as well as for the teaching of Arabic speakers in particular.

1.1 ENCODING- VS. RETRIEVAL-BASED REHEARSAL

Traditionally, the processes of human memory have been divided into *encoding* (sometimes referred to as *storage*) processes and *retrieval* processes (Tulving, 1991). Information enters associative memory via encoding processes and is transferred from there into working memory via retrieval processes. The present experiment relies upon a comprehensive retrieval-based studying methodology which only secondarily prompts students to use encoding processes in the course of their learning. Justification for this design decision hinges upon recent work conducted by Karpicke & Roediger (2008) which is explored and critiqued below.

Karpicke and Roediger (2008) observe that studies of memory have classically followed a split-mechanism paradigm, in which participants first undergo an encoding-based "instruction" phase and then undergo a retrieval-based "testing" phase. They outline a "standard assumption" that learning primarily occurs "while people study and encode material" (p. 966), and that the process of information retrieval—the transfer of information from associative memory into working memory—is "a relatively neutral event that measures the learning that occurred during study but does not by itself produce learning" (p. 966). Contrary to this traditional dichotomy between "productive" encoding and "passive" retrieval, Karpicke and Roediger claim to demonstrate a powerful learning effect correlated with repetitions of information retrieval tasks and not with repetitions of information encoding tasks.

The claim that repeated information retrieval tasks yield a greater learning effect than information encoding tasks has far-reaching implications in pedagogy, autodidactics, and the study of the mind. Simply put, if repeated retrieval-based studying methodologies consistently yield superior results to encoding-based studying methodologies, then certain current approaches to instruction and rehearsal will be made obsolete. In particular, Karpicke & Roediger (2008)

note that these findings would refute the advice to students to "study something until it is learned…and then drop it from further practice" (p. 967), a method "endorsed by contemporary theories of study-time allocation…[and] explicitly encouraged in many popular study guides" (p. 966). In addition, students may be provided with powerful new learning tools designed to harness the power of information retrieval tasks to strengthen learning—a goal which the present experiment attempts to work toward. To this end, Karpicke and Roediger lament that "students exhibited no awareness of the mnemonic effects of retrieval practice" and that their surveys of college-aged students indicate that "self-testing is a seldom-used strategy [for studying]" (p. 968).

Karpicke and Roediger's experiment centered upon a foreign language vocabulary acquisition, retention, and retrieval task. They divided participants into four treatment groups. Each group underwent several cycles of two phases: an encoding phase and a testing phase. The critical manipulations among the four groups were permutations of the binary levels of two factors: (1) whether successfully-recalled terms were included on subsequent word pair lists during the encoding phase, and (2) whether successfully-recalled terms were included on subsequent word pair retrieval tests during the testing phase. Karpicke and Roediger assigned the labels ST, S_NT, ST_N, and S_NT_N to these four groups, in which "S" signifies "encoding task"; "T" signifies "retrieval task"; "$_N$" signifies "only nonrecalled word pairs"; and the lack of "$_N$" signifies "all word pairs, regardless of whether they have been recalled successfully." These conditions are summarized by Table 1 below.

Table 1. Conditions of Karpicke and Roediger (2008)

		Encoding Task (S)	
		Learned items retained	**Learned items dropped**
Retrieval Task (T)	**Learned items retained**	ST	S_NT
	Learned items dropped	ST_N	S_NT_N[1]

All participants were given a list of 40 Swahili-English word pairs to study. Then, all participants were tested on the contents of the entire list. Participants were shown the Swahili word and asked to provide the English translation. Following the testing period, participants were given a 30-second distracter task "that involved verifying multiplication problems" (966). After this first learning cycle, the critical manipulations came into effect: the participants engaged in three more cycles of encoding-based studying and retrieval-based testing in which the ST group always studied all words and was tested upon all words, the ST_N group always studied all words but was only tested upon nonrecalled words, the S_NT group only studied nonrecalled words but was tested on all words, and the S_NT_N group only studied and was tested upon nonrecalled words. Following the four rounds of learning, the participants were dismissed, and were given a post-examination one week later upon the full word pair list.

[1] Karpicke and Roediger note that the S_NT_N condition "represents what conventional wisdom and many educators instruct students to do: Study something until it is learned (i.e., can be recalled) and then drop it from further practice." (p. 967)

Over the course of the learning trials, four groups demonstrated almost identical learning curves, as seen below in Karpicke and Roediger's Figure 1, reproduced below as Figure 1:

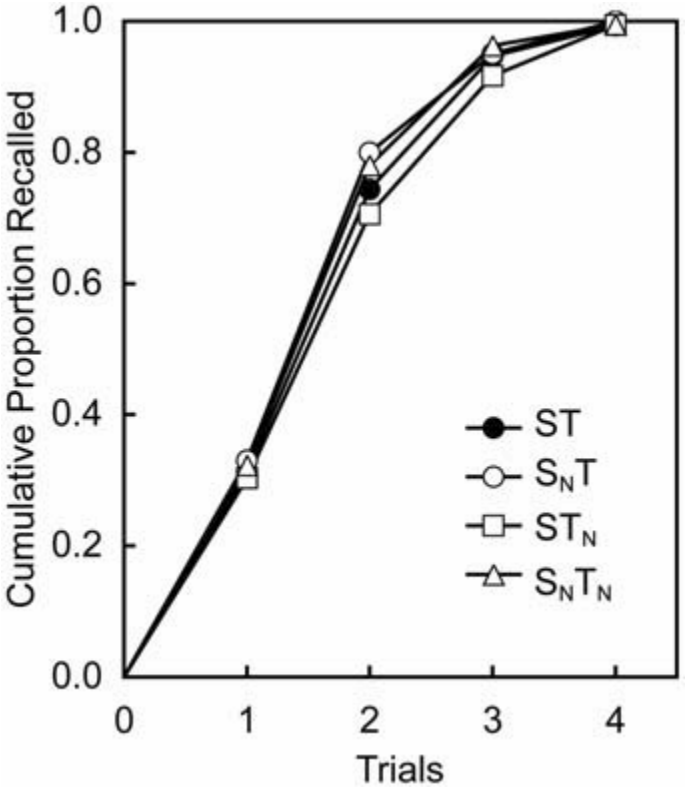


Figure 1. Cumulative performance during the learning phase. (Karpicke & Roediger, 2008, p. 967)

However, the groups demonstrated highly differentiated levels of performance on the post-assessment administered one week after the learning phase of the experiment, with strong separation occurring over the two levels of the retrieval-based learning factor. Karpicke and Roediger's Figure 2, which reflects this discrepancy in post-assessment scores, is reproduced below as Figure 2:

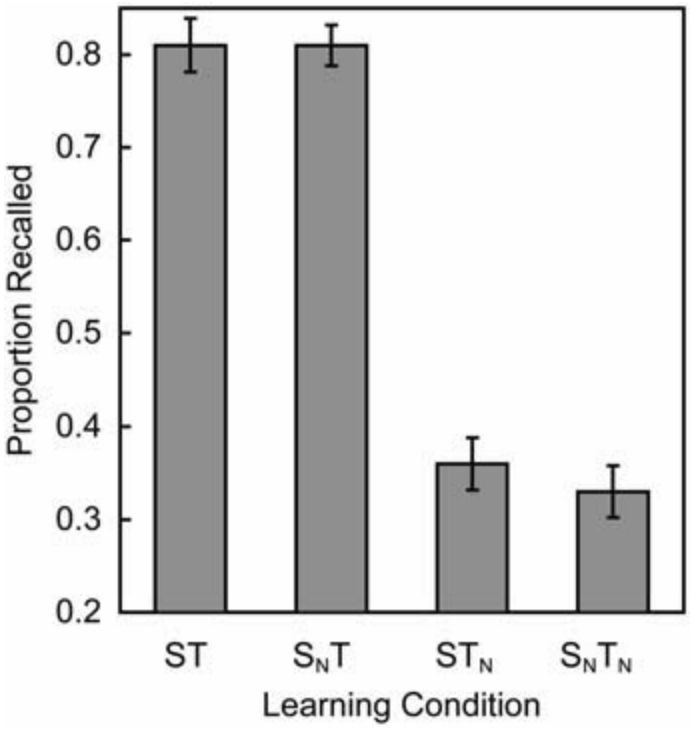


Figure 2. Proportion recalled on the final test 1 week after learning. Error bars represent standard errors of the mean. (Karpicke & Roediger, 2008, p. 967)

The effect is clear: the participants who were consistently tested on the entire range of words, not just those which had not yet been learned, performed exceptionally well on the post-assessment compared to the participants who had dropped word-pairs from testing upon the first successful recollection. Comparing the "T" groups with the "T_N" groups reveals that the former groups scored four standard deviations ($d = 4.03$) above the latter groups (p. 967). The other factor under examination did not yield such results: repeated encoding-based studying "produced virtually no effect on long-term retention" (p. 967).

Karpicke and Roediger's results strongly support the claims that (1) tasks of encoding and retrieval have different effects upon learning and memory retention, and that (2) retrieval-based learning tasks yield enhanced long-term memory retrieval success for items studied in that fashion compared to items studied via encoding-based learning tasks. An alternate hypothesis

would be that the ST group performed so much better than the S_NT_N group on the post-assessment because the former group underwent so many more trials[2] than the latter group: 320 trials vs. 154.8 trials per subject. However, when the two "mixed conditions"—S_NT and ST_N—are compared, the different number of trials encountered by the different treatment conditions was wholly unrelated to the results obtained. The S_NT group experienced an average of 236.8 trials (76.8 encoding, 160 retrieval) while the ST_N group experienced an average of 242.7 trials (160 encoding, 82.7 retrieval), and yet the S_NT group performed at a level about four standard deviations above the ST_N group. This demonstrates that the results were not a trivial consequence of the time spent on the learning cycles, but were truly a consequence of the different types of rehearsal tasks employed, i.e., encoding practice vs. retrieval practice.

Karpicke and Roediger's (2008) experimental results justify the employment of retrieval-based studying methods in classroom interventions. However, the modality of the task is only one facet of a much larger picture. The scheduling and administration of the task must also be considered. The following section on computer-assisted rehearsal scheduling addresses these issues.

1.2 THE SCHEDULING PROBLEM

Karpicke and Roediger's (2008) findings are of crucial importance in designing a studying methodology meant to promote long-term memory retrieval success in vocabulary acquisition. Where possible, students ought to practice retrieval-based studying methods—i.e.,

[2] i.e., the sum of all encoding opportunities added to the sum of all retrieval opportunities across the four learning cycles

some form of self-testing—to strengthen their ability to retrieve pieces of information. However, it is difficult to overstate the extent to which this basic advice falls short of actionable application when we examine the problem of long-term memory retention and retrieval in a greater sense. Karpicke and Roediger's subjects were tasked with learning a list of 40 foreign word pairs and then recalling their definitions after only one week. A student studying a language with the intention of speaking it fluently, on the other hand, will need to learn between two to four orders of magnitude more words in the target language, and will ideally retain that knowledge for life. In an article in *Wired Magazine* profiling Piotr Wozniak, a memory researcher who developed the spaced repetition rehearsal software SuperMemo, Wolf (2008) illustrates this problem by recounting Wozniak's experience as a university student in Poland attempting to learn English:

> How long would it take [Wozniak] to master the things he wanted to know? The answer: too long. In fact, the answer was worse than too long. According to Wozniak's first calculations, success was impossible. The problem wasn't learning the material; it was retaining it. He found that 40 percent of his English vocabulary vanished over time…Using some simple calculations, he figured out that with his normal method of study, it would require two hours of practice every day to learn and retain a modest English vocabulary of 15,000 words. For 30,000 words, Wozniak would need twice that time. This was impractical.

The magnitude of words suggested in Wolf's (2008) account is pessimistic; following the corpus work of Carroll, Davies, & Richman (1971), Nation (2001) noted that a vocabulary size of 5,000 words was sufficient to cover 89.4% of text in a text consisting of five million running words, while 12,448 words would be sufficient to cover fully 95% of the text (p. 15). Thus, a vocabulary of 15,000 words is certainly more than "modest" in size. However, the larger point stands. As the amount of material one wishes to master increases, the task of scheduling and

structuring comprehensive study sessions becomes prohibitively difficult. Per Karpicke and Roediger (2008), to ensure truly comprehensive retention of target items, equally comprehensive self-testing would need to be employed. In the case of a small corpus of association pairs, such as the symbols of each of the chemical elements, a massed method of comprehensive self-testing would be manageable for a learner to employ. However, for a language-learning task, this approach is untenable. To routinely test oneself on hundreds or thousands of foreign words would require both careful planning and monumental effort.

Imagine a student who wishes to use self-testing to study a corpus of 10,000 words in a target language. This task would become theoretically manageable if the 10,000 self-testing exercises were divided into 100 sets of 100 target words each, with one set to be studied per day in repeating 100-day cycles. However, the time spent in that pursuit would still be enormous in aggregate, and by the time the learner finished the first cycle of 30,000 words, he or she likely would have forgotten many of words learned in the first set due to a lack of practice in the interceding months.[3] In addition, as the years crawled by, familiar words would continue to be repeated in rehearsal as often as unfamiliar words, leading to wasted and misdirected effort.

In short, this study scheduling problem lacks an *intuitive* answer beyond "study something until it is learned…and then drop it from further practice" (Karpicke and Roediger, 2008, p. 967), a strategy which we have already rejected above. For comprehensive knowledge retention over time, comprehensive self-testing must be employed, as Karpicke and Roediger (2008) have demonstrated. However, we seem to lack any intuitive mechanism by which to accomplish this fine-grained level of continuous self-testing, and the success rate of our long-

[3] If our hypothetical student were to routinely use the target language for communication alongside this studying regimen, many of the higher-frequency words would continue to be practiced in daily speech, helping to keep those words fresh in memory. Many students lack the opportunity for such routine and extended immersion, however.

term memory retrieval suffers because of it. As Wolf (2008) laments, "When it comes to language, the received wisdom is that immersion—usually amounting to actual immigration—is necessary to achieve fluency…it's an awful commentary on the value of countless classroom hours. Learning things is easy. But remembering them—this is where a certain hopelessness sets in."

1.2.1 Forgetting Curves and the Spacing Effect

In 1885, Hermann Ebbinghaus published *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie*,[4] in which he recounted a series of self-experiments pertaining to memory and knowledge retrieval. In analyzing his own performance on years' worth of nonsense syllable memorization and recitation tasks, Ebbinghaus precisely charted the sloped course of memory attrition over time, now known as the forgetting curve (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). He also demonstrated that the correct spacing of practice sessions could dramatically improve memory retrieval success on his recitation tasks, a phenomenon now called the "spacing effect." Although these findings may seem trivial in a sense—that the probability of successfully recalling a piece of information drops over time in a predictable way, and that periodic rehearsal of material spaced out over time has a strong positive effect upon memory retrieval success—regularities which Ebbinghaus charted provided an empirical basis for the murky study of memory, and his collected data and analyses became the basis of most future research in practice and forgetting (as asserted by, e.g., Bahrick & Hall, 2005, p. 566; Pavlik & Anderson, 2008, p. 111).

[4] This title is rendered in translation as, e.g., *Memory: a contribution to experimental psychology* (Ebbinghaus, 1964).

In the century since the original research conducted by Ebbinghaus, studies have generally shown a consistent and strong positive effect of spaced repetition rehearsal (versus massed, "cramming" rehearsal) upon the long-term memory retrieval success of the material so rehearsed (see, e.g., Cull, Shaughnessy, & Zechmeister, 1996; Dempster, 1988; Landauer & Bjork, 1978; Rea & Modigliani, 1985). In short: learners who review material periodically in an expanding schedule over a period of multiple days or weeks are better able to retrieve the information when tested compared to individuals who study intensely for a brief period and then neglect to refresh their knowledge. Wolf (2008) provides an idealization of the forgetting curve which illustrates the effect of spaced repetition rehearsal, reproduced below as Figure 3:[5]

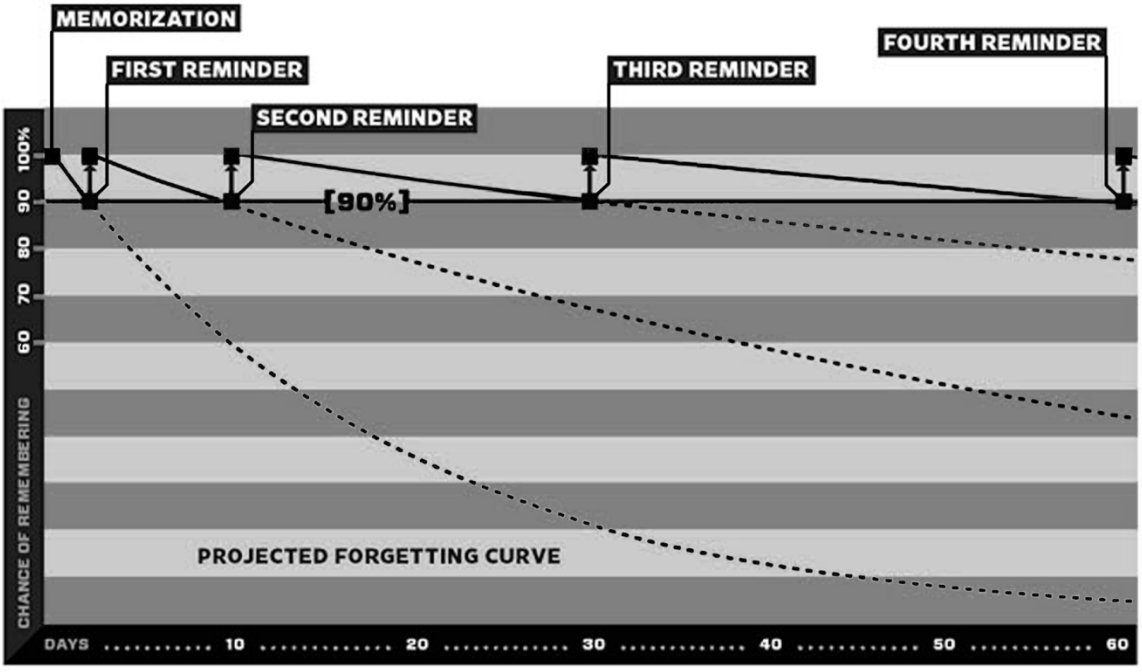


Figure 3. Idealization of the forgetting curve and spaced repetition rehearsal for a single item (Wolf 2008)

[5] Although Wolf (2008) is an article in the popular press written by a journalist and not by a researcher, this figure accurately represents the concept of a forgetting curve and qualitatively mirrors that of Pimsleur (1967) which is offered below as Figure 4.

As the figure illustrates, learning a piece of information once and then failing to rehearse it following acquisition leads to poor long-term retention. However, each subsequent rehearsal of the information strengthens its storage strength in associative memory, enabling easier recollection in the future. In Wolf's idealization, these rehearsals are calibrated to occur when the learner is predicted to have a 90% chance of correctly recalling the item from memory—a time interval which expands with every repetition of rehearsal, as each trial enhances the storage strength of the item in question. The figure of 90% is an arbitrary choice, however. Different researchers have proposed different optimal levels of error in rehearsal tasks. For example, in a recent learning experiment, Pavlik and Anderson (2008) utilized a model which predicted that a successful recollection rate of 99.2% would be optimal for a certain spaced repetition learning task (p. 111), mirroring Skinner's (1968) assertion that errorless learning is optimal, whereas Pimsleur (1967) offered a target accuracy of 60% when describing L2 vocabulary rehearsal, stating that this cutoff point represents a "good chance" of remembering the target item (p. 74-75; see Figure 4 below for Pimsleur's idealization of the forgetting curve).

1.2.2 Different conceptions of "spaced repetition rehearsal" in research

Other researchers do not define their conception of spaced repetition based upon adaptive methods designed to reach target recollection rates at all, choosing instead to employ pre-determined repetition spacings and then examining the resulting effect of these spacings upon recollection rates. Karpicke and Roediger (2007) conducted an experiment to compare the effects of a number of rehearsal schedules for a learning task, including a spaced-repetition schedule; their formulation of spaced repetition, which they termed an "expanding retrieval practice" schedule, was a hard-coded, one-size-fits-all practice schedule implemented over the course of a

single study period in a single day. Their entire experiment, including a final post-test for some participants, lasted only three days.

This conception of the spacing effect as a phenomenon to be studied on the timescale of minutes and seconds (in addition to days and weeks) is also illustrated by Pimsleur (1967). He provides an idealization similar to that of Wolf (2008) given above, but with a radically smaller timescale, reproduced below as Figure 4:

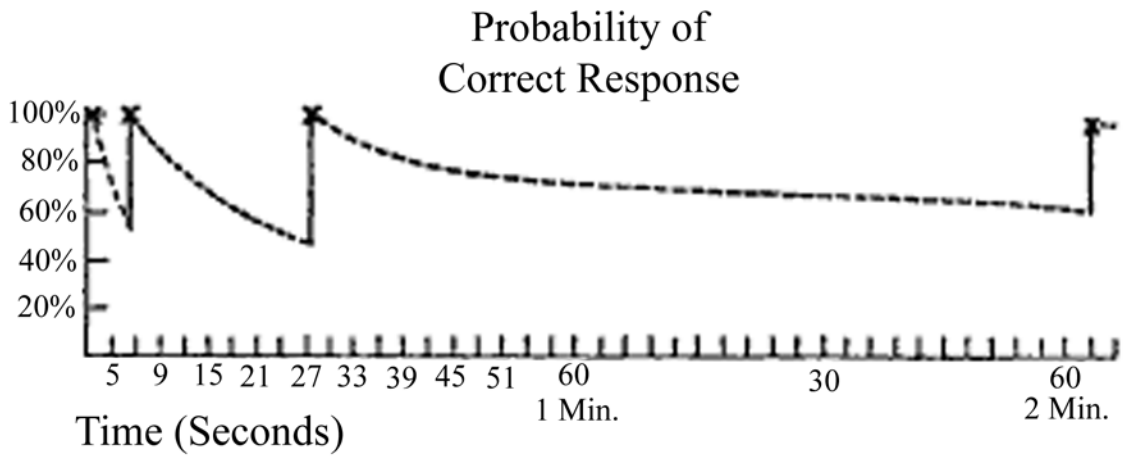


Figure 4. Idealization of the forgetting curve and spaced repetition rehearsal for a single item (Pimsleur, 1967, p. 75)

The present experiment utilizes software[6] which is appropriate for use on both timescales. It tracks the user's performance on drill trials to infer the user's mastery of each rehearsal item. Using those estimations of mastery, the program manipulates the order of future trials to ensure that items which most urgently require rehearsal are reviewed sooner than items which less urgently require rehearsal. Thus, a single difficult item (as determined by the accuracy and latency of the user's responses) might be drilled several times within a span of minutes to

[6] See section 2.2.8: FaCT (Fact and Concept Training) System

enhance mastery. An easier item, on the other hand, might not be drilled for days or weeks at a time, so long as the scheduling algorithm estimates that the user is still very likely to remember the correct answer to that item.

1.2.3 Current applications of spaced repetition learning methods

Despite being "one of the most remarkable phenomena to emerge from laboratory research on learning," spaced repetition learning methods failed to find significant classroom application in the century following the pioneering work of Ebbinghaus (Dempster, 1988). Scattered exceptions do exist, including the "Pimsleur Method" of language instruction. The details of this system are not available for scrutiny, as it is a commercial product; however, Pimsleur's own writing on the topic (1967) appears to give an adequate account of the concepts behind the methodology, though not the specific details of its implementation. The Pimsleur Method employs some of the most fundamental concepts of retrieval-based spaced repetition rehearsal, such as the concepts of the forgetting curve, memory retrieval strength, and memory storage strength, as well as the basic idea of spreading rehearsal over a long period of time rather than massing it. Material is learned and rehearsed in a framework of expanding interval practice (termed "graduated interval recall" by Pimlseur), beginning with multiple repetitions over the course of a minute (review Figure 4 above) and ultimately progressing to multi-month intervals between rehearsals on an exponentially-expanding timescale (Pimsleur, 1967, p. 75).

While these methodologies still have little presence in the classroom, the advent of low-cost personal computers has enabled individual learners to utilize computer-aided spaced repetition rehearsal software in learning and memorizing material. In his MA thesis, Wozniak (1990) recounted his experience creating and calibrating a computerized implementation of a

flash card-based spaced repetition review regimen which would schedule every flash card for future review based upon an expanding-interval schedule modified by the learner's self-scoring of memory retrieval success during rehearsal. In other words, through this method of adaptive scheduling, newer cards are reviewed more often than older cards, and difficult cards are reviewed more often than easier cards. Wozniak developed this system into a commercial product known as SuperMemo[7] (Wolf, 2008) which has since been deployed in at least one experiment in the field of TEFL (Szofer, 2010) and at least one in the field of cognitive psychology (Metzler-Baddeley & Baddeley, 2009).

In recent years, Pavlik has done research exploring the "microeconomics of learning," focusing on maximizing the mnemonic benefits of rehearsal tasks while minimizing their costs in terms time spent and trials failed (Pavlik, 2005; Pavlik & Anderson, 2005, 2008; Pavlik et al., 2007). The innovation of Pavlik's approach is attending specifically to the optimization and utility of individual trials rather than focusing only on bigger-picture questions such as learning objectives, task modality, etc. The individual trials are optimized via adaptive scheduling algorithms which attempt to achieve a high successful response rate and a low response latency based on a user's prior rehearsal data. By optimizing the effectiveness and efficiency of the individual trials—i.e., optimizing the "microeconomics"—we may hope to achieve a better educational outcome. The present study uses software developed by Pavlik which attends to such microeconomics; see section 2.2.8 below for a more detailed account of this software and its theoretical underpinnings.

[7] http://www.supermemo.com/index.htm

1.3 CLOZE EXERCISES

The present experiment uses CALL software to guide participants in the completion of cloze exercises. A cloze exercise is a type of task in which a language learner must supply a target vocabulary or grammatical item in order to complete a sentence. The act of creating a cloze exercise—viz., of deleting words from a passage—is known as the *cloze procedure*. A brief history of the use of this procedure, as well as justification for its relevance to the present research, is presented below.

Taylor (1953) was the inventor of the cloze procedure and initially used it to measure the readability of written passages. Essentially, the more accurately an adult native speaker could fill in the missing words, the more readable a passage was determined to be. Later, Taylor (1956) focused on the individual differences between his participants and attempted to use cloze exercises as a psychometric test. He reported that it correlated well with other comprehension and intelligence test scores. In neither case did Taylor intend to use the cloze procedure to create educational cloze exercises. However, following the psychometric work of Taylor (1956), later researchers found the concept of the cloze procedure useful in designing assessments of language proficiency.

Educational research picked up in the 1960s for first-language teaching and assessment (Jongsma, 1971; Rye, 1982) and in the 1970s for second-language assessment (Alderson, 1979; Oller & Conrad, 1971). However, the use of cloze exercises as a rehearsal exercise (as opposed to an assessment exercise) in second-language learning remained poorly studied until recent years, perhaps owing in part to the generally poor reputation of that method of studying—Folse (2006) notes that "[m]any educators see fill-in-the-blank exercises as a superficial or passive use of the vocabulary, especially when compared to writing original sentences" (p. 286).

Folse (2006) conducted an experiment to determine the effectiveness of cloze exercises as a type of written L2 vocabulary practice, and his positive results motivate the use of close exercises in the present study. Folse contrasted two types of written exercise: cloze exercises and original sentence writing. He further subdivided the cloze exercise rehearsal into two conditions: one in which each target word was rehearsed by one sentence context, and one in which each target word was rehearsed in three sentence contexts. These written practice conditions are summarized in Table 2 below.

Table 2. Written practice conditions of Folse (2006)

Condition	Type of written practice
1	One cloze exercise per target word
2	Three cloze exercises per target word
3	One original sentence per target word

Each participant in Folse's study completed each condition, with each of the conditions focusing randomly on one of three possible target word lists (A, B, and C) to ensure that participants never studied the same word in two different written practice conditions. Furthermore, the ordering of the conditions was randomized for each participant for the sake of counterbalancing, for a total of 36 possible conditions. Reflecting these 36 conditions, 36 versions of a practice booklet were printed and randomly distributed to 154 participants.

Folse examined two dependent variables: post-test outcome and time on task. Scores on the post-test were significantly higher for words studied in practice condition 2 (three cloze exercises) than for words studied in practice conditions 1 (one cloze exercise) or 3 (original sentence), $p < .0001$. The effect sizes, as calculated using Cohen's *d*, were 1.01 and 0.91, respectively, indicating very large effects. However, students generally spent more time on task in condition 2 than in condition 3. To control for this variable, Folse conducted a post-hoc analysis upon a subset of 31 participants who had equivalent times on task in conditions 2 and 3.

Their post-test scores reflected the trend seen in the general population: condition 2 yielded higher post-test scores than condition 3, $p < .0001$. Folse did not report Cohen's *d* for this comparison, but notes that the mean score on items studied under condition 2 was more than twice the mean score of items studied under condition 3.

Folse's (2006) experiment is relevant to the present study because it demonstrates that cloze exercises—which are a one-step, retrieval-based task easily graded by computer—are efficient in terms of time on task and effective in terms of learning outcome. Folse himself proposes the following:

> Multiple encounters using fill-in-the-blank activities is a task that not only can be done extremely efficiently by the computer but also produces superior retention results. Therefore, it behooves L2 vocabulary software designers to ensure that multiple encounters with the target item is an integral part of their learning software; likewise, educators should look for this feature in software that they might purchase for their learners. (Folse, 2006, p. 289)

The present study thus expands on the work of Folse by implementing the type of cloze-exercise-based computer-assisted rehearsal tool which provides multiple encounters with target forms in different sentence contexts.

1.4 THE ASSISTANCE DILEMMA

In any instructional activity, the question of how much support to provide a student in order to achieve an optimal learning outcome has been a pressing issue for decades. Vygotsky's (1978) theory of the Zone of Proximal Development—defined as "the distance between the actual

developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance, or in collaboration with more capable peers" (p. 86)—yielded the concept of scaffolding (see, e.g., Wood, Bruner, & Ross, 1976), which may be broadly defined as providing a student with some form of support in order to help him or her successfully complete a learning task. Scaffolding is an influential notion in the field of second language teaching and of education in general.

The wrong level of scaffolding for a task, however, may be harmful and inhibit an optimal learning outcome. This problem is known as the assistance dilemma (Koedinger & Aleven, 2007; Koedinger, Pavlik, McLaren, & Aleven, 2008). Koedinger, Pavlik, McLaren, & Aleven (2008) provide the following table to illustrate the assistance dilemma, reproduced below as Table 3:

Table 3. Assisting performance during instruction may aid or harm learning (K. R. Koedinger et al., 2008, p. 2155)

Instructional Support	Poor learning outcome	Better learning outcome
High assistance (less demanding)	crutch	scaffold
Low assistance (more demanding)	undesirable difficulty; extraneous load	desirable difficulty; germane load

As Table 3 above suggests, depending on the demands of the task, the needs of the individual student, and the type of scaffolding provided, it is possible for high levels of scaffolding to have good or bad effects on the learning outcome, just as it is possible for low levels of scaffolding to have good or bad effects on the learning outcome.

For example, imagine two language learners, one of low-intermediate skill and one of advanced skill, watching a film in the target language. Subtitles presented in the L2 would be an appropriate form of support for the low-intermediate student, because they would help the

student to at least understand the gist of the film—a good educational outcome. A lack of subtitles might lead this student to feel completely lost and frustrated, however—a poor educational outcome. For the advanced student, subtitles presented in the L2 may be treated as a crutch, as they enable the student to rely on visual support rather than practicing his or her listening skills, leading to a poor educational outcome. If the subtitles are removed, however, this student will experience a desirable level of difficulty with the listening task, leading to a good educational outcome. The trick, then, is to use the correct level of scaffolding proportional to the demands of the task and to the abilities of the students in order to achieve an optimal learning outcome.

Although the general concept of the assistance dilemma is quite simple and intuitive, the dizzying range of factors present in educational and psychological research complicate efforts to operationalize "assistance" in any sort of universal, context-independent way. Thus, Koedinger & Aleven (2007) note that "although we have argued that good instruction should find an appropriate balance between the giving and withholding of information, there is not as yet a strong theoretical basis for predicting where the balance should lie—currently, finding the right balance is an empirical question" (p. 252).

One of the factors in the present experiment is that of the level of support which participants receive when completing cloze exercises. There are two levels to this factor. In the first condition, a "no-hint" condition, participants are provided only with the cloze exercise. In the second condition, a "hint" condition, participants are provided with the cloze exercise plus a sort definition of the target word (see Appendix E). The rationale behind providing a hint is that it may provide benefits while reducing some of the costs inherent in completing cloze exercises.

Koedinger & Aleven (2007) provide a table conducive to a more detailed discussion of the benefits and costs associated with scaffolding, reproduced below as Table 4.

Table 4. The assistance dilemma: finding the balance between information or assistance giving and withholding is a fundamental challenge in designing effective instruction (K. R. Koedinger & Aleven, 2007, p. 242)

	Benefit	**Cost**
Giving information or assistance	Accuracy Efficiency of communication Thrill of (supported) success	Shallow processing Lack of attention May not engage long-term memory Stealing chance to shine
Withholding information or assistance	Generation effect Forces attention Engages long-term memory Thrill of independent success	Cost of errors Floundering, confusion, wasted time Frustration of failure

In terms of the vocabulary provided by K. R. Koedinger & Aleven (2007) in Table 4 above, it would be most desirable in the present study for scaffolding to enhance accuracy and the thrill of supported success while decreasing wasted time and the frustration of failure.

2.0 METHOD

2.1 PARTICIPANTS

The participants in this study were 46 students enrolled in the level 4 writing course of the University of Pittsburgh's English Language Institute (ELI). Students who enroll in level 4 courses are at an intermediate level of proficiency in English, with TOEFL iBT scores of approximately 45. Eight native languages were represented. These languages and their frequencies of representation are given below in Table 5.

Table 5. Native languages of total pool of participants

Language	N	Percent of total
Arabic	29	63.0%
Chinese	6	13.0%
Japanese	1	2.2%
Korean	6	13.0%
Russian	1	2.2%
Spanish	1	2.2%
Turkish	1	2.2%
Thai	1	2.2%
Total	**46**	**100%**

Although attrition was problematic in this study in general, a certain portion of participants completed so few study trials that the decision was made not to analyze their rehearsal data. These discarded participants (n = 6) each completed fewer than 20 trials. Six of

these participants were Arabic speakers and one was a Turkish speaker. With these participants eliminated, we may revise Table 5 as Table 6 below:

Table 6. Native languages of analyzed participants

Language	N	Percent of total
Arabic	24	60%
Chinese	6	15%
Japanese	1	2.5%
Korean	6	15%
Russian	1	2.5%
Spanish	1	2.5%
Thai	1	2.5%
Total	**40**	**100%**

Participants were not offered compensation for their participation in this study. Rather, the study was implemented as a part of the Writing 4 course curriculum. General class participation and work habits comprise 5% of students' final grades in the Writing 4 class, and participation in this experiment constituted a portion of that 5% course grade. (The participants' teachers individually decided how much to weigh participation in this experiment in the students' class participation grades.) Although all students participated in the instructional intervention, participation in the research component of the intervention was optional. In other words, students had the option of withdrawing their rehearsal and learning data from analysis.

2.2 MATERIALS

2.2.1 Word lists

The present study utilized two word lists: a target word list and a control word list. The target word list was the English Language Institute (ELI)'s Level 4 Core Vocabulary list (Appendix A), while the control word list was the ELI's Level 5 Core Vocabulary list (Appendix C). The Core Vocabulary system was developed in 2007 ("English Language Institute Vocabulary List," 2007) to help provide students with more consistent exposure to academic and professional vocabulary across their ELI classes. The Core Vocabulary system is split into three lists of 50 words each: one list for all Level 3 courses, one list for all Level 4 courses (Appendix A), and one list for all Level 5 courses (Appendix C). These words are derived from the Coxhead Academic Word List (Coxhead, 2000). Both the target words and the control words were tested on the pre-test and post-test. However, only target words were used in the instructional intervention.

The utilization of the Level 5 word list for the control items was a decision of convenience, as the Level 5 words are not taught in the Level 4 courses. Further investigation following the conclusion of the experiment, however, revealed that the words comprising the two lists have substantially different frequency values according to the British National Corpus (Cobb, n d; *The British National Corpus, version 3,* 2007). The frequency values of the target items are presented in Appendix B, while the frequency values of the control items are presented in Appendix D. An abbreviated comparison is presented below as Table 7:

Table 7. Frequency values of target and control vocabulary items

Frequency	Target words	Control words
BNC-1,000	8	0
BNC-2,000	26	9
BNC-3,000	5	8
BNC-4,000	8	17
BNC-5,000	2	12
BNC-6,000	1	2
BNC-7,000	0	0
BNC-8,000	0	2

The table above demonstrates that the target word list contains substantially more frequent words than does the control word list. Thus, the word lists were generally poor matches for one another. Section 3.2 below confirms that participants' knowledge of these two lists of words were not comparable.

2.2.2 Project description and schedule

Participants in this intervention were given a project description and schedule which explained the purpose of the intervention, what its context was within the course, and provided instructions on how participants could "opt out" from the research component. The schedule detailed exactly what actions were expected of participants on each day. See Appendix F for the complete document.

2.2.3 Background survey

At the beginning of the study, a background survey was administered to the participants (see Appendix G).

2.2.4 Pre-test and post-test

At the beginning of the study, a vocabulary pre-test was administered to the participants. At the conclusion of the study, an identical post-test was administered (see Appendix H).

2.2.5 Opinion survey

At the conclusion of the study, an opinion survey was administered to the participants (see Appendix I).

2.2.6 Cloze exercise contexts

A total of 150 cloze exercise contexts were created for this study: three contexts for each of the 50 target words. These cloze exercises are presented below in Appendix E. Most of the contexts were written personally by the author, but a portion of the sentences were written in whole or in part by anonymous paid contributors on Amazon.com's Mechanical Turk[8] (see Appendix F). In his experiment utilizing cloze exercises, Folse (2006) noted how difficult it is to operationalize the appropriateness of a cloze exercise context: "The challenge was to compose sentences that sounded native-like, were at a very low level of ESL proficiency, [and] did not use any new or difficult vocabulary" (p. 282). Folse does not describe whether he attempted to measure or balance these factors. In the present study, the only quantitatively-operationalized restriction on possible cloze exercises was in the length of the sentence: each cloze exercise context was required to be between 8 and 16 words in length, inclusive. A more meticulous study might

[8] https://www.mturk.com/

choose to control for, e.g., the specific syntactic structures and vocabulary frequencies allowed in the sentences, but in the present study, no other systematic restriction beyond the length restriction was enforced. Thus, the appropriateness of a given sentence—in terms of sounding native-like while being written at a level comprehensible to the students—was determined subjectively.

2.2.6.1 Per-participant inventories of cloze exercise contexts

Every participant's inventory of cloze exercises was unique, albeit drawn from the pool of 150 exercises given below in Appendix E. Recall that the list of 50 target vocabulary items is divided into ten weekly lists of five words each in the ELI Level 4 courses. Because three cloze exercise contexts were written for every vocabulary item, this yields 15 possible cloze exercise contexts per weekly list of five words.

Every participant's personal inventory of cloze exercises was arranged automatically by the FaCT System software. The software followed this process:

1. Individually scramble each of the ten clusters of 15 cloze exercises so that the sentences within a cluster are shuffled randomly, but that the clusters remain chronologically ordered relative to one another—with week 2's words always preceding week 3's words, and so on.
2. Discard the last 5 sentences of every 15-sentence chunk for a total inventory of 100 sentences.
3. Out of the inventory of 100 sentences, assign the "even" sentences to the "hint" condition and the "odd" sentences to the "no-hint" condition.

The final result of this manipulation is a list of 100 cloze exercises balanced between the two experimental conditions which are ordered such that participants must demonstrate mastery of earlier material before being exposed to later material.

2.2.7 Target word definitions

Each vocabulary item was paired with one or more target definitions (see Appendix E). Each target definition was utilized in the "hint" condition and also informed the choice of definition to illustrate in the cloze exercise contexts. The source of most of the target definitions was the official Core Vocabulary list of the English Language Institute to ensure a good fit between the material taught in the classroom and the material presented in the tutor software. However, many of the definitions were modified for length or clarity. In addition, definitions which themselves included the target word were rewritten not to include the target word, as shown in Table 8 below.

Table 8. Example of rewritten definition.

Original and rewritten definitions of the word "alternative"	
Original definition	an alternative idea, plan etc. is different from the one you have and can be used instead
Rewritten definition	Something like an idea or plan that is different from the main one you have and can be used instead.

2.2.8 FaCT (Fact and Concept Training) System

The present study uses a modified version of a piece of software called the FaCT (Fact and Concept Training) System (Pavlik et al., 2007). The FaCT System was developed at Carnegie Mellon University by Pavlik and a team of contributors. It is designed to help students practice

single-step problems, such as paired-associate items or multiple-choice questions. Because cloze exercises are single-step problems, they are also appropriate to implement using this software.

The FaCT System implements a spaced repetition task scheduler to ensure that every individual study item is cycled back for review when necessary, yielding comprehensive retrieval-based rehearsal. The practice scheduling module implements a modified version of the ACT-R model of declarative memory (Anderson & Lebiere, 1998; Pavlik, 2005). More specifically, it utilizes the equations from ACT-R which model the strength of a memory chunk as a function of practice (Pavlik & Anderson, 2008). These equations predict both correctness and response latency as a function of prior performance in rehearsal drills, and rehearsal scheduling is structured to optimize both of these variables (i.e., to achieve trials which are both quick and accurate). Thus, one important theoretical difference between the FaCT System and most other computer-aided methods of spaced-repetition rehearsal is that the FaCT System aims to settle on the optimal balance between multiple different measurable aspects of item rehearsal; it does not focus solely on theoretical long-term learning outcome, but rather takes into consideration the efficiency of the act of rehearsal itself. Previous work attempting to implement computationally-driven rehearsal scheduling, such as the three-state Markov model of Atkinson (1972), tended to neglect predictions of latency, but ignoring this variable can lead to uneconomical practice (Pavlik, 2005). In practice, the FaCT System tends to predict that sacrificing long spacings between trials for the benefit of achieving higher accuracy and lower response latency is optimal (see, e.g., Pavlik and Anderson 2008, p. 111).

Pavlik & Anderson (2008) tested their ACT-R-based rehearsal scheduling model by implementing it alongside a modified version of Atkinson's (1972) Markov model as well as a "naïve" flashcard-based intervention in a between-groups design. In the experiment, 60

participants studied a set of 180 Japanese-English word pairs using one of the three rehearsal methods during learning sessions on Monday, Wednesday, and Friday of one week. A final rehearsal session occurred the following Friday. By the fourth rehearsal session, the participants using the ACT-R-based system were achieving significantly higher accuracy at significantly lower latency than either of the two other groups. In addition, the total number of trials completed by each participant was significantly higher in the ACT-R-based rehearsal group than in the other two groups, further demonstrating that this method was considerably more time efficient than the other methods. These results are encouraging in informing the selection of the FaCT System to drive the rehearsal scheduling in the present experiment; however, as Pavlik & Anderson (2008) note, "Although the method clearly has implications for the learning of large sets of paired-associate items by young naïve participants, it is less obvious what this implies for different tasks, different populations of learners, or different materials" (p. 112). Although the sentence completion task of the present experiment is similar to the paired-associate drill trials of Pavlik & Anderson (2008) in the sense that both are one-step problems in which a stimulus prompts an intended one-word response, the sentence completion task does not involve a simple one-to-one mapping between items. Rather, each target word may be elicited by up to three possible cloze exercise contexts, with or without a possible hint. Thus, the present task is more complex in some ways than a paired-associate drill trial.

In the context of the present experiment, the most important features of the FaCT System are the following: it provides a computer-driven interface through which students may complete single-step written exercises; it implements comprehensive retrieval-based rehearsal based on an adaptive expanding schedule; and it generates detailed log files which track the response time

and accuracy on every trial. The specific constructs and variable weightings of the practice scheduler are beyond the scope of the present study.

2.3 PROCEDURE

This study was conducted over the course of 9 weeks in parallel with the Writing 4 course at the University of Pittsburgh English Language Institute. The intervention began in the third full week of the course, by which time the participants had already studied the target words of weeks 2 and 3 and were in the process of studying the target words of week 4 (see Appendix C). Due to scheduling issues, it was not feasible to begin the study earlier than this.

There were four different sections of the Writing 4 course. During the first week of the study, each course section spent one class period in a computer lab for an orientation session regarding the instructional intervention. Because all of the course sections met at the same time of day, it was necessary to distribute the orientation sessions across Monday, Tuesday, Thursday, and Friday.

During this orientation session, participants received the project description and schedule. Then, the participants completed the background survey and the pre-test, followed by a 10-minute session with the FaCT System rehearsal software. During this time, participants were monitored to ensure they understood the software and were using it correctly. Participants were encouraged to ask for help or clarification if needed.

During weeks two through eight, on Mondays and Thursdays, participants were to go to a class website to access the rehearsal software, log in, and engage in 10 minutes of cloze procedure drill rehearsal as a form of homework. The teachers of the Writing 4 classes were

instructed to remind participants on Mondays and Thursdays to complete this online vocabulary rehearsal. On the Thursday of week eight, after 16 total sessions using the software, all participants were to access the class website to take the vocabulary post-test and to take the opinion survey regarding the vocabulary rehearsal software.

Due to the low number of post-tests and opinion surveys completed by participants during week eight, the deadline for these materials was extended. Week nine was spent attempting to solicit more post-tests and opinion surveys from participants.

2.3.1 Description of a rehearsal session using the FaCT System software

The twice-weekly rehearsal sessions completed by participants followed this general procedure:

1. The participant would go to the class website, launch the tutor, and log in. (See Figure 5)
2. The participant would then be presented with a cloze exercise, either with or without a hint. (Any given cloze exercise would always be presented in the same condition, so that a cloze exercise presented with a hint would always be presented with a hint, and vice versa.)
3. The participant would type his or her answer in the answer box. (See Figure 6)
4. The participant would submit his or her answer. If he or she answered correctly, then the program would display a check mark and proceed to the next cloze exercise. If he or she answered incorrectly, the program would display a magnifying glass icon and present the participant with the correct answer. (See Figure 7)
5. After 10 minutes of studying, the participant was free to quit the session.

Figure 5, Figure 6, and Figure 7 below illustrate the software and those phases of the cloze exercise rehearsal task given above.

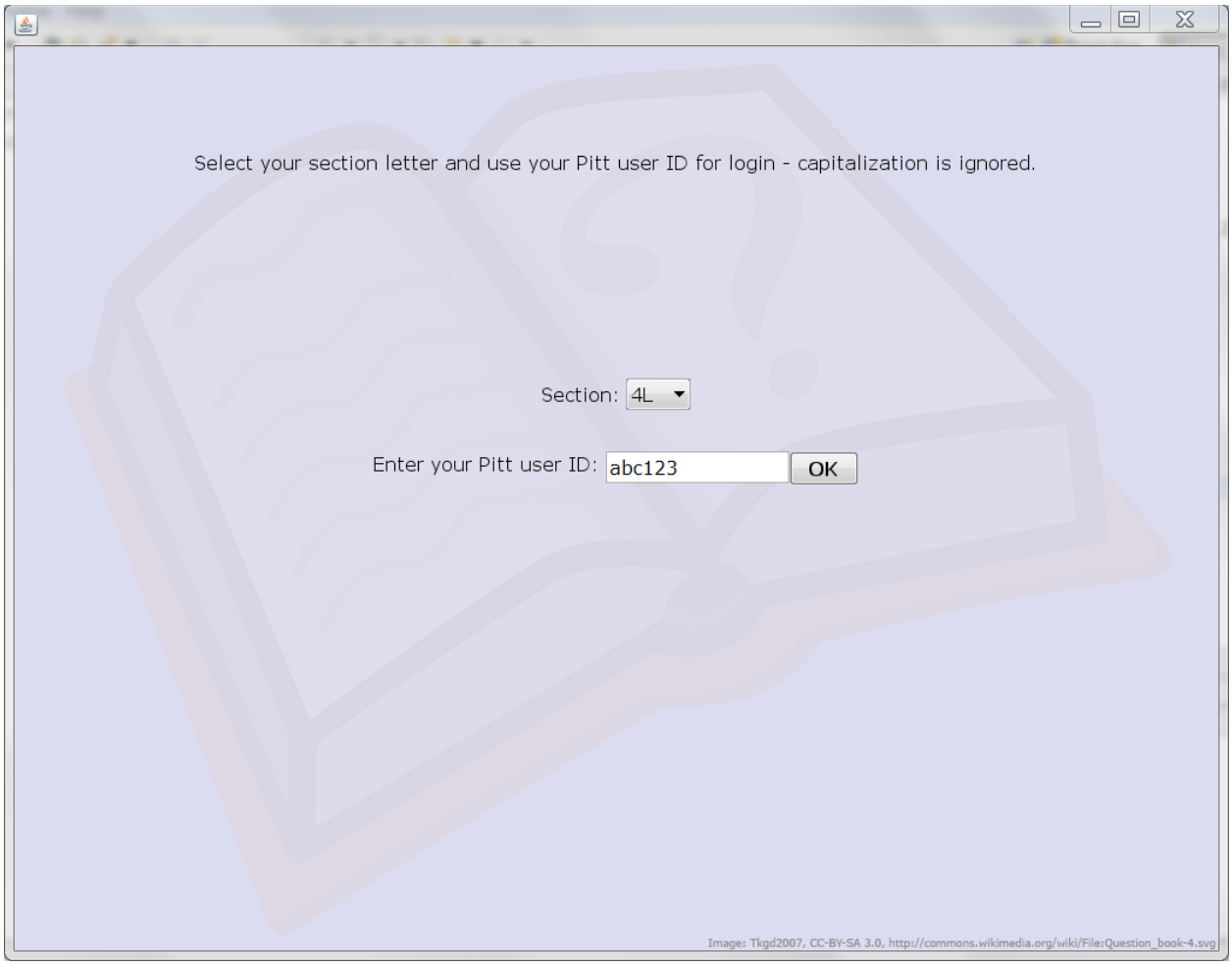


Figure 5. Login screen of the tutor software

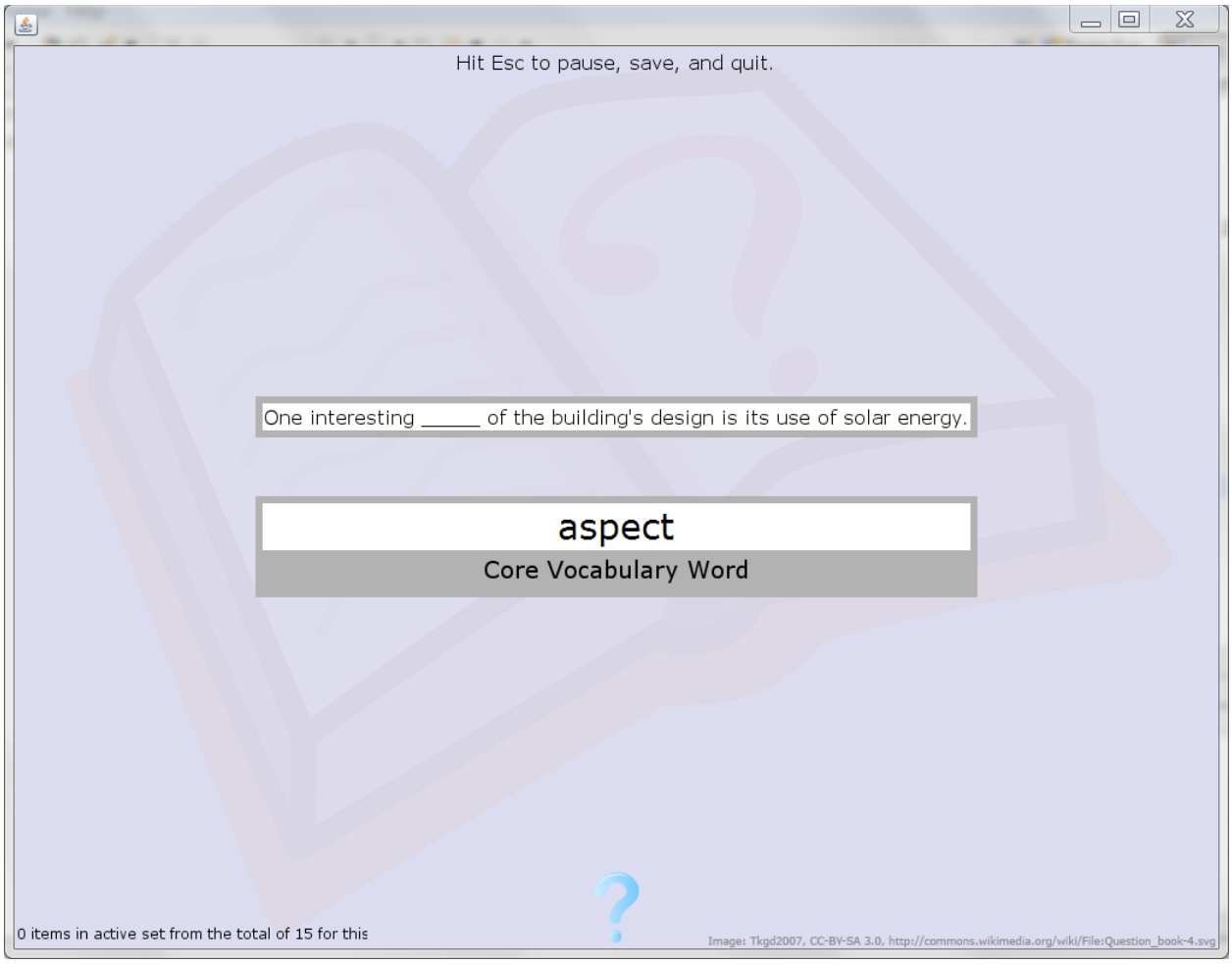


Figure 6. Cloze exercise rehearsal screen, with participant's answer filled in

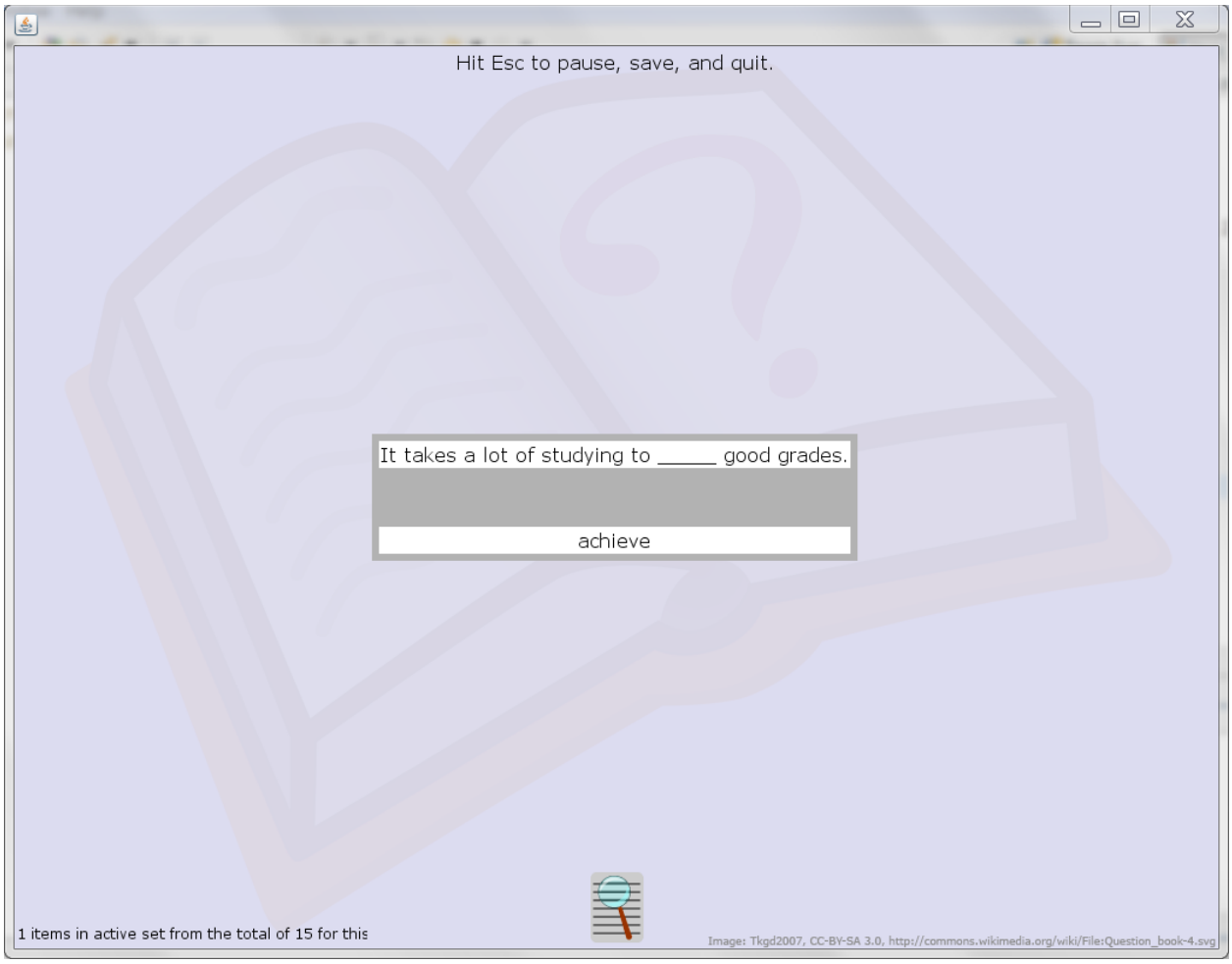


Figure 7. Incorrect trial review screen, with correct target word displayed

3.0 RESULTS

3.1 BACKGROUND SURVEY

The background survey (Appendix G) was completed by 45 out of 46 participants. The survey revealed several important details about the participants in this study. The majority of participants (63%) were L1 speakers of Arabic.[9] The next largest L1 groups were Chinese (13%) and Korean (13%). Five other languages were attested by one speaker each. These results are summarized in Table 7 below.

Table 9. Native languages of total pool of participants

Language	N	Percent of total
Arabic	29	63.0%
Chinese	6	13.0%
Japanese	1	2.2%
Korean	6	13.0%
Russian	1	2.2%
Spanish	1	2.2%
Turkish	1	2.2%
Thai	1	2.2%
Total	**46**	**100%**

The background survey also revealed that 37% of the participants (N = 17) had taken at least one Level 4 course in the ELI in a previous semester. This fact is important because it reveals that a substantial portion of the participant pool had previously been taught the list of

[9] The one participant who failed to complete the background survey was a speaker of Arabic.

target words. Their prior knowledge is borne out in the pre-test results presented in section 3.2 below.

Another important piece of information from the background survey was from participants' responses to the question, "Which do you enjoy more, speaking English or writing in English?" Participants answered using a six-point Likert scale, with '1' corresponding to "Speaking" and '6' corresponding to "Writing". A summary of participants' responses is given below in Table 10.

Table 10. Participants' preferences of speaking or writing in English (lower score indicates affinity for speaking)

Language	**N**	**Mean**	**Standard Deviation**
Arabic	28	2.18	1.090
Chinese	6	3.17	.753
Japanese	1	6.00	
Korean	6	3.00	1.673
Russian	1	2.00	
Spanish	1	1.00	
Turkish	1	3.00	
Thai	1	1.00	
Total	**46**	**2.47**	**1.290**

A box plot of the data in Table 10 are presented below as Figure 8.

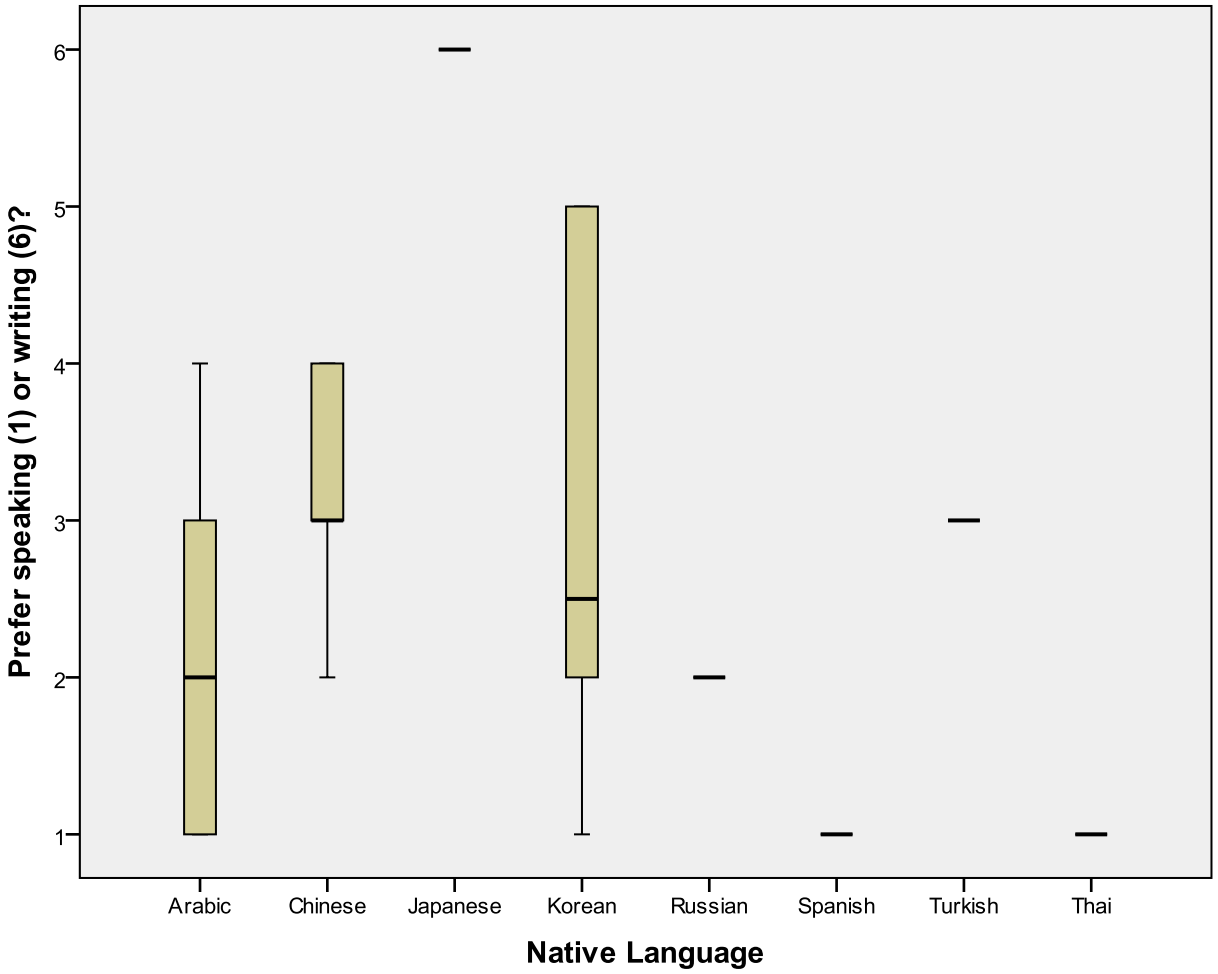


Figure 8. Participants' preferences of speaking or writing in English, by native language

Out of the eight possible L1 backgrounds, five (Japanese, Russian, Spanish, Turkish, and Thai) were attested by only one speaker each. This between-groups fragmentation and glut of small N sizes makes direct statistical analysis of L1 effects impossible. The decision was thus made to group students of similar sociocultural backgrounds together. Arabic speakers formed one cohort; Chinese, Japanese, and Korean speakers formed an East Asian cohort (following Juffs et al., forthcoming, who similarly grouped Chinese and Korean speakers); and the remaining four individuals—one speaker each of Russian, Spanish, Turkish, and Thai—were

excluded from analysis. The descriptive statistics for these two sociocultural groups are presented below in Table 11.

Table 11. Arabic and East Asian participants' preference of speaking (1) or writing (6)

Sociocultural groups	N	Mean	Standard Deviation
East Asian	13	3.31	1.437
Arabic	28	2.18	1.090

A box plot of the data in Table 11 is presented below as Figure 9.

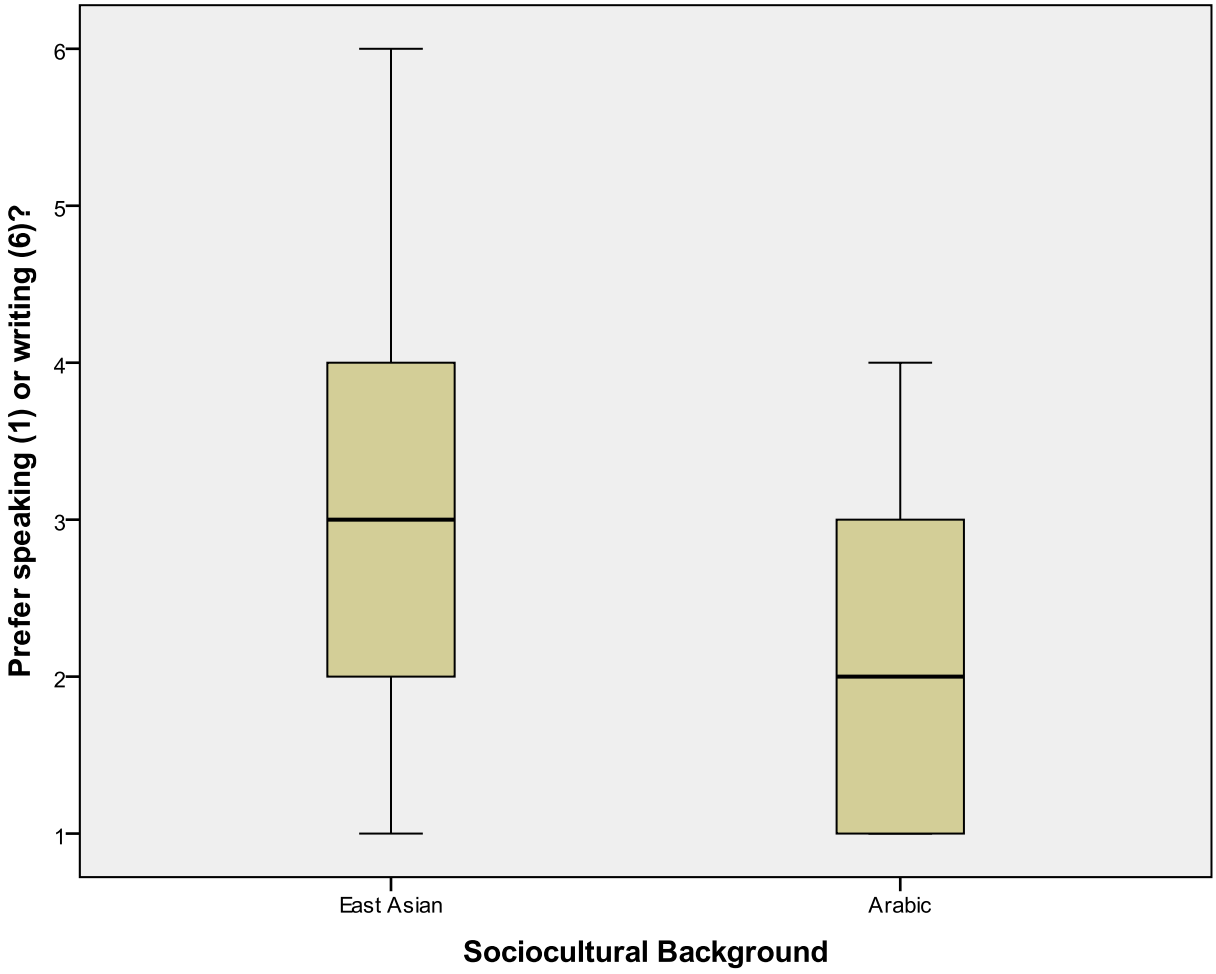


Figure 9. Arabic and East Asian participants' preferences of speaking or writing in English

The box plot in Figure 9 demonstrates that out of the 28 Arabic participants, not a single one had a strong preference for writing in English; rather, the interquartile range spans only from '1' to '3' (out of 6) on the Likert scale, indicating an almost universal preference among these participants for speaking over writing. Their mean score of 2.16 indicates a solid preference for speaking. By comparison, the East Asian group's interquartile range lies closer to the center of the scale, and the group's distribution demonstrates a positive skew extending all the way up to a rating of '6'. Their mean score of 3.31 demonstrates a much more balanced range of preferences between speaking and writing, lying only 0.19 points away from the midpoint of the scale.

An independent-samples t-test was conducted on the data from Table 11 above to see if the Arabic group and the East Asian group differ significantly in their preferences toward speaking or writing in English. For the test, the 95% CI for the difference in means is .309 to 1.949 ($t(39) = 2.786$, $p = .008$, two-tailed, Cohen's $d = 0.89$). The results of this test suggest that Arabic speakers prefer speaking in English (as opposed to writing in English) significantly more than speakers of East Asian languages do, and that this effect is large. This result is important because the rehearsal method employed in this intervention is a writing task, and if Arabic speakers are particularly prone to disliking writing in English in general, we could predict that these individuals would also dislike the studying method. For any instructional intervention deployed over a period of weeks—especially one presented as a recurring homework assignment—a participant's affinity toward the assignment could be a predictor of continued participation. This issue is explored further in section 3.6 below.

Another question on the background survey asked whether participants preferred listening to or reading English. This question is similar to the question regarding speaking vs. writing insofar as it contrasts a skill grounded in orthography with a skill grounded in speech.

However, by focusing on receptive skills, this question helps to avoid potential confounds caused by participants' personalities, such as introversion versus extraversion, as well as potential confounds caused by participants' aversion to difficulties specific to writing or speech production. In other words, this question more neutrally gauges participants' attitudes toward orthography-based or speech-based media independently of other factors. The descriptive statistics for East Asian and Arabic students' responses to this question are presented below as Table 12.

Table 12. Arabic and East Asian participants' preference of listening (1) or reading (6)

Sociocultural groups	N	Mean	Standard Deviation
East Asian	13	3.08	1.256
Arabic	28	3.11	1.315

A box plot of the data in Table 12 is presented below as Figure 10.

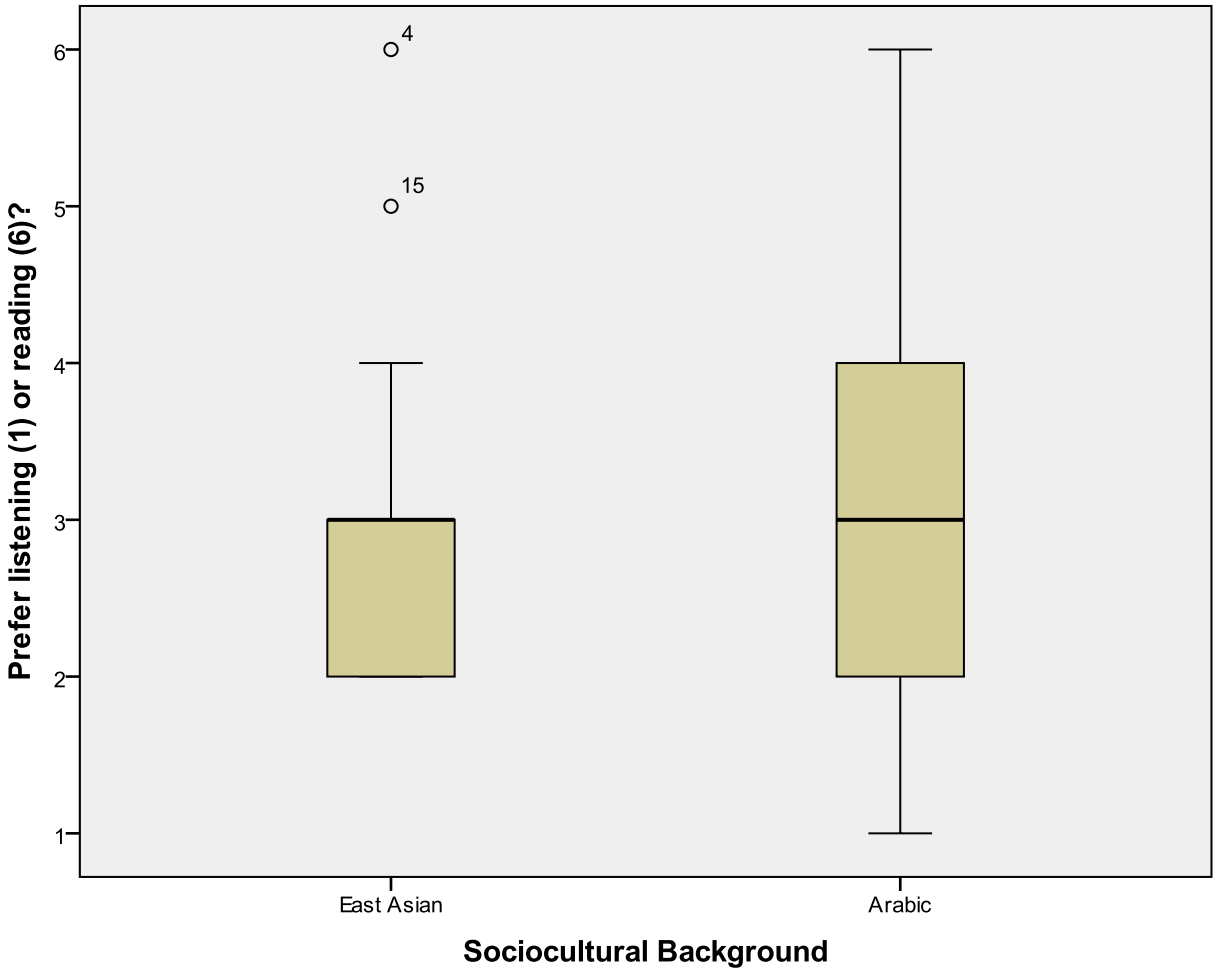


Figure 10. Arabic and East Asian participants' preferences of listening or reading in English

The box plot in Figure 10 demonstrates a substantial overlap between the distributions of the two groups. In addition, the means and standard deviations are virtually identical. An independent-samples t-test was conducted on the data from Table 12 above to see if the Arabic group and the East Asian group differ significantly in their preferences toward speaking or writing in English. For the test, the 95% CI for the difference in means is -.911 to .850 ($t(39) = .069$, $p = .945$, two-tailed). The results of this test demonstrate that there are no significant differences between these two sociocultural cohorts' preferences for listening or reading tasks.

A final comparison was conducted within the Arabic cohort to compare these participants' attitudes toward reading and writing. The descriptive statistics for this comparison are provided below as Table 13.

Table 13. Arabic participants' preferences regarding reading and writing

Sociocultural groups	N	Mean	Standard Deviation
East Asian	13	3.08	1.256
Arabic	28	3.11	1.315

A box plot of the data in Table 13 is presented below as Figure 11.

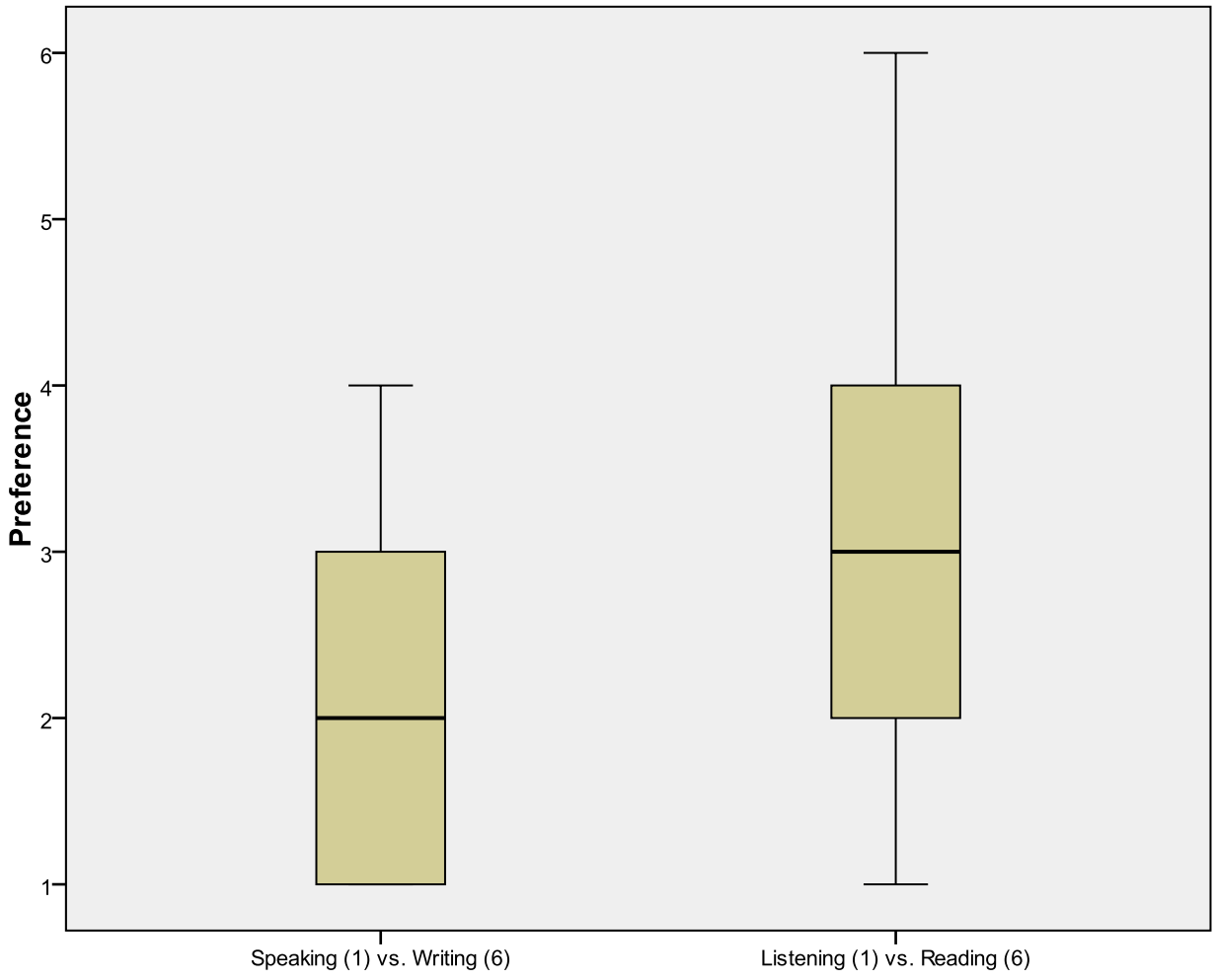


Figure 11. Arabic participants' preferences regarding reading and writing

The box plot in Figure 11 reveals a certain amount of overlap between these variables, but the "listening vs. reading" variable clusters more closely to the center and has a larger positive skew than does the "speaking vs. writing" variable. A paired-samples t-test was conducted on the data from Table 13 above to see if the Arabic participants have significantly different preferences between productive (speaking vs. writing) and receptive (listening vs. reading) skills. For the test, the 95% CI for the difference in means is .287 to 1.570 ($t(27) = 2.971$, $p = .006$, two-tailed, Cohen's $d = .565$). The result confirms a significant difference between Arabic speakers' preferences for orthography-based and speech-based tasks across productive and receptive modalities. Whatever the source of this difference is, its effect size is medium.

3.2 PRE-TEST

The pre-test (Appendix H) was completed by 45 out of 46 participants. The pre-test consisted of 50 target items as well as 50 control items. Descriptive statistics of the results of the pre-test are provided below in Table 14:

Table 14. Descriptive statistics of pre-test

N	Word group	Mean	Std. Dev.
45	Target	35.76	11.109
45	Control	20.44	10.610

The data in Table 14 show a large division between knowledge of target words and knowledge of control words. A paired-samples t-test was conducted on the scores on target

words and control words on the pre-test. For the test, the 95% CI for the difference in means is 13.084 to 17.538 ($t(44) = 13.856$, $p < .001$, two-tailed, Cohen's $d = 1.41$). This result demonstrates that an enormous difference in performance existed between the target word list and the control word list on the pre-test.

As noted in section 3.1 above, 37% of the participants (N = 17) had previously taken at least one Level 4 course in the ELI and had thus been taught the target vocabulary. The descriptive statistics of these participants' scores on the target words versus other participants' scores is given below as Table 15:

Table 15. Descriptive statistics of target words on the pre-test (prior Level 4 experience vs. no experience)

Prior Level 4 experience?	N	Mean	Std. Dev.
Yes	17	41.71	10.415
No	28	32.14	10.051

An independent-samples t-test was conducted on the data in Table 15 above. For the test, the 95% CI for the difference in means is 3.246 to 15.880 ($t(43) = 3.053$, $p = .004$, two-tailed, Cohen's $d = 0.94$). This result indicates that the participants with prior Level 4 experience scored significantly higher than participants new to Level 4, with a very large effect size. This result may appear obvious, but it is important in the context of this study insofar as it demonstrates a limit in the ability of this study to detect positive educational outcome effects. The fact that even the group without any prior Level 4 experience scored an average of 32.14 out of 50 (64%) on the target items demonstrates that participants began this intervention already quite confident with their grasp of the target vocabulary items.

3.3 POST-TEST

The post-test (Appendix H) was completed by only 4 out of 46 participants. This minimal level of participation reflects the generally high rate of attrition witnessed during this study (see section 3.6). These participants' scores on the pre-test and the post-test are presented below as Table 16.

Table 16. Pre-test and post-test results for participants who completed the post-test

Test	Words	N	Mean	Std. Dev.
Pre-test	Target	4	42.75	6.602
	Control	4	24.75	5.500
Post-test	Target	4	47.50	3.109
	Control	4	29.50	8.583

A box plot of these participants' test results are presented below as Figure 12.

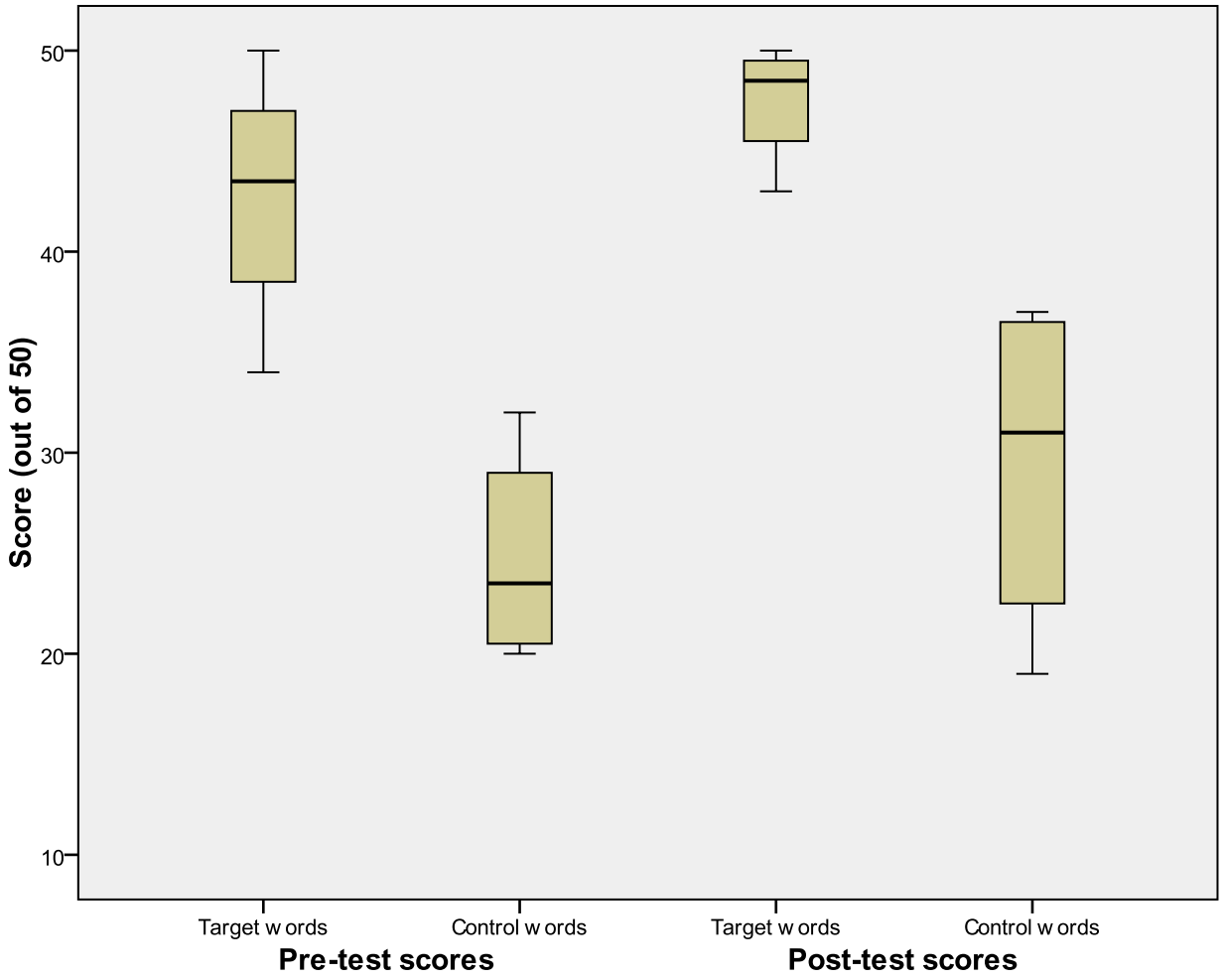


Figure 12. Pre-test and post-test results for participants who completed the post-test

The box plot in Figure 12 reveals that the four participants who completed the post-test were among the highest-scoring students out of the 45 who took the pre-test. Notably, one participant had scored at ceiling level on target words on the pre-test, making it impossible *a priori* to detect a positive learning outcome from that student.

Due to the very small N size, it was not possible to conduct paired-samples t-tests on these data. The Wilcoxon signed ranks test was selected as a suitable nonparametric substitute. A related-samples Wilcoxon signed rank test was conducted on the pre-test and post-test scores on the target word list—i.e., the list of words studied in the instructional intervention. The test

indicated that there was no statistically significant difference between pre-test and post-test scores, $z = -.736$, $p = .461$.

A related-samples Wilcoxon signed rank test was conducted on the pre-test and post-test scores on the control word list. These words were not studied in the instructional intervention. The test indicated that there was no statistically significant difference between pre-test and post-test scores, $z = -1.461$, $p = .144$.

3.4 OPINION SURVEY

The opinion survey (Appendix I) was completed by only 4 out of 46 participants—the same participants who completed the post-test, as these instruments were administered sequentially. Out of seven six-point Likert scale items, one participant gave the highest score, '6', six times, including on statements that appear to contradict each other (for example, strongly agreeing both that the tutor was "too easy" and also "very useful"). Thus, only three of the respondents appear to have completed the survey in good faith.

Of specific interest to the present study was the question of whether the presence of hints helps or hinders participants' rehearsal using the tutor software. On the six-point Likert scale question, "In the vocabulary tutor program, did you prefer to study vocabulary WITH hints or WITHOUT hints?", all three participants gave a score of '3', indicating a slight preference for studying using hints.

Only one of the respondents on the opinion survey was a speaker of Arabic. On the six-point Likert scales, this participant only selected answers of '3' or '4', indicating very mild opinions. However, the participant found the tutor to be a bit too difficult; a little bit useful; and

not very enjoyable. This participant also marginally agreed that the tutor helped him or her to learn a lot of vocabulary and to study it more quickly than he or she used to, but marginally disagreed that the tutor helped him or her to use the new vocabulary words more easily.

3.5 REHEARSAL STATISTICS

The rehearsal statistics logged by the vocabulary tutor (i.e., the FaCT System program) yielded a total of 4,198 core vocabulary trials across 49 participants. These trials were far from evenly distributed across participants, however. The fewest number of trials completed by a participant was 2, while one exceptionally highly motivated participant accounted for fully 851 trials—20.3% of the total number recorded.

Before further analysis was conducted, the decision was made to eliminate some of the less reliable sources of data. Two different sets of trials were excluded: all trials from participants who completed fewer than 20 total trials were excluded from analysis, and all trials which were completed in under 2000 milliseconds were excluded. (In the latter case, it is likely that the participant either hit the "Enter" key by accident.)

After this initial filtering of data, 3,986 trials and 40 participants remained. For each participant, three critical measurements were calculated from their rehearsal data: the average amount of time they spent reading the prompt before pressing a key ("latency 1"); the average amount of time which elapsed between the first keystroke and the end of the trial, assuming that

the student made an attempt to type an answer[10] ("latency 2"); and the average accuracy the participant achieved on the cloze exercises ("accuracy"). These three measurements and their descriptions are provided below in Table 17.

Table 17. Dependent measures from the tutor software

Name	Simple description	Operational description
Latency 1	Average time spent reading and thinking	The average amount of time elapsed between the display of the cloze exercise prompt and the first keystroke made by the participant
Latency 2	Average time spent typing and checking work	The average amount of time elapsed between the first keystroke made by the participant and the end of the trial
Accuracy	Average accuracy	The total number of correct answers divided by the total number of trials completed

3.5.1 Tests for the effect of level of support on participants' performance

The three dependent measures, Latency 1, Latency 2, and Accuracy were further divided into the two levels of the "Support" factor, namely whether the participant was presented with a hint or not in the trial. Descriptive statistics for these measurements are presented below in Table 18.

[10] In other words, trials in which the student never touched the keyboard did not count this phase as being 0ms in length; rather, such trials were excluded from the calculation of this variable.

Table 18. Rehearsal data of analyzed subjects

Condition	Measurement	Mean	Std. Dev.	N
Hint	Latency 1	10179 ms	2331.00	40
	Latency 2	4291 ms	1510.31	40
	Accuracy	68.9%	0.15276	40
No-hint	Latency 1	9684 ms	2235.13	40
	Latency 2	3723 ms	1105.92	40
	Accuracy	63.6%	0.13732	40

Box plots of these participants' rehearsal data are presented below in Figures Figure 13 and Figure 14.

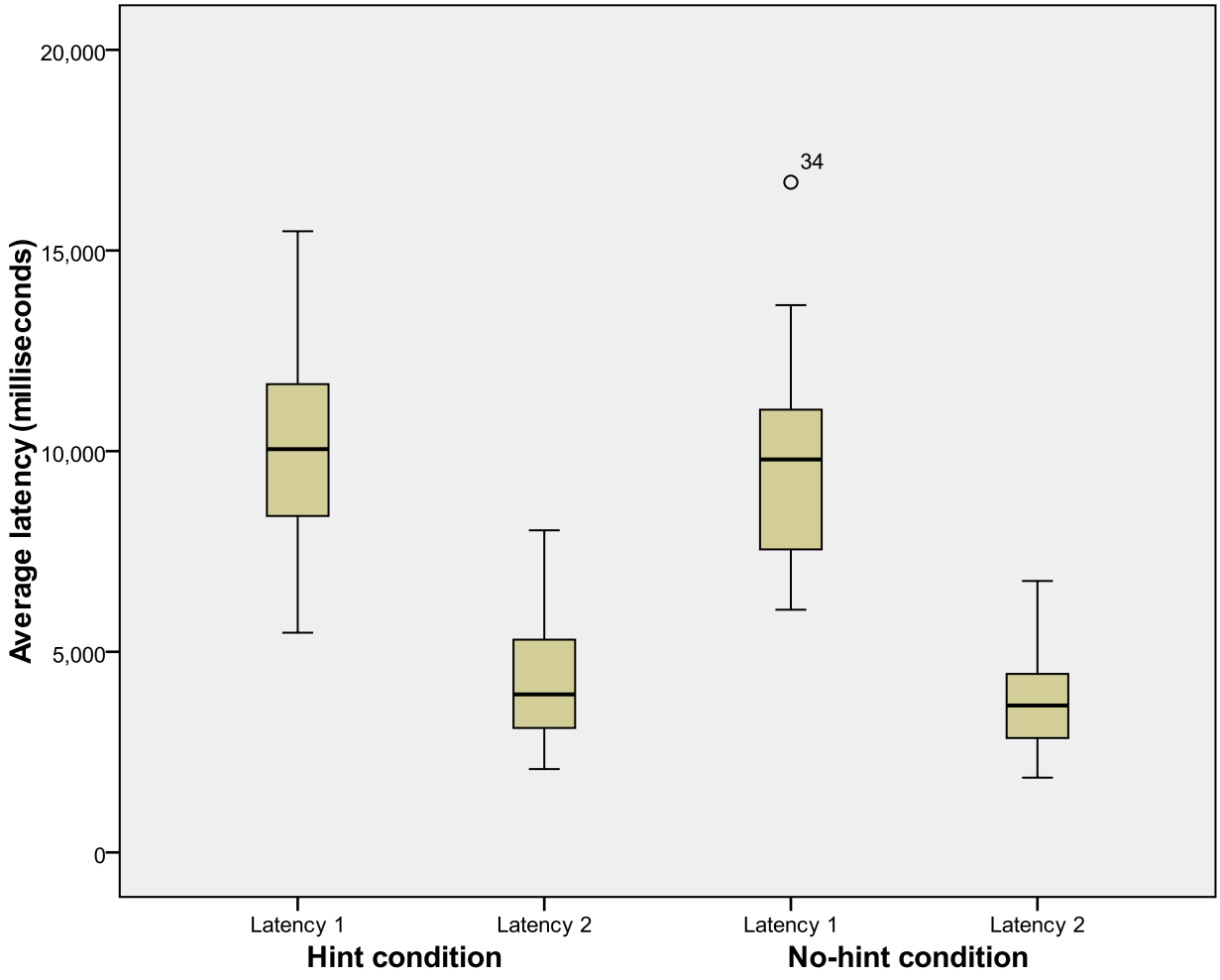


Figure 13. Latency measurements in the hint condition and no-hint condition

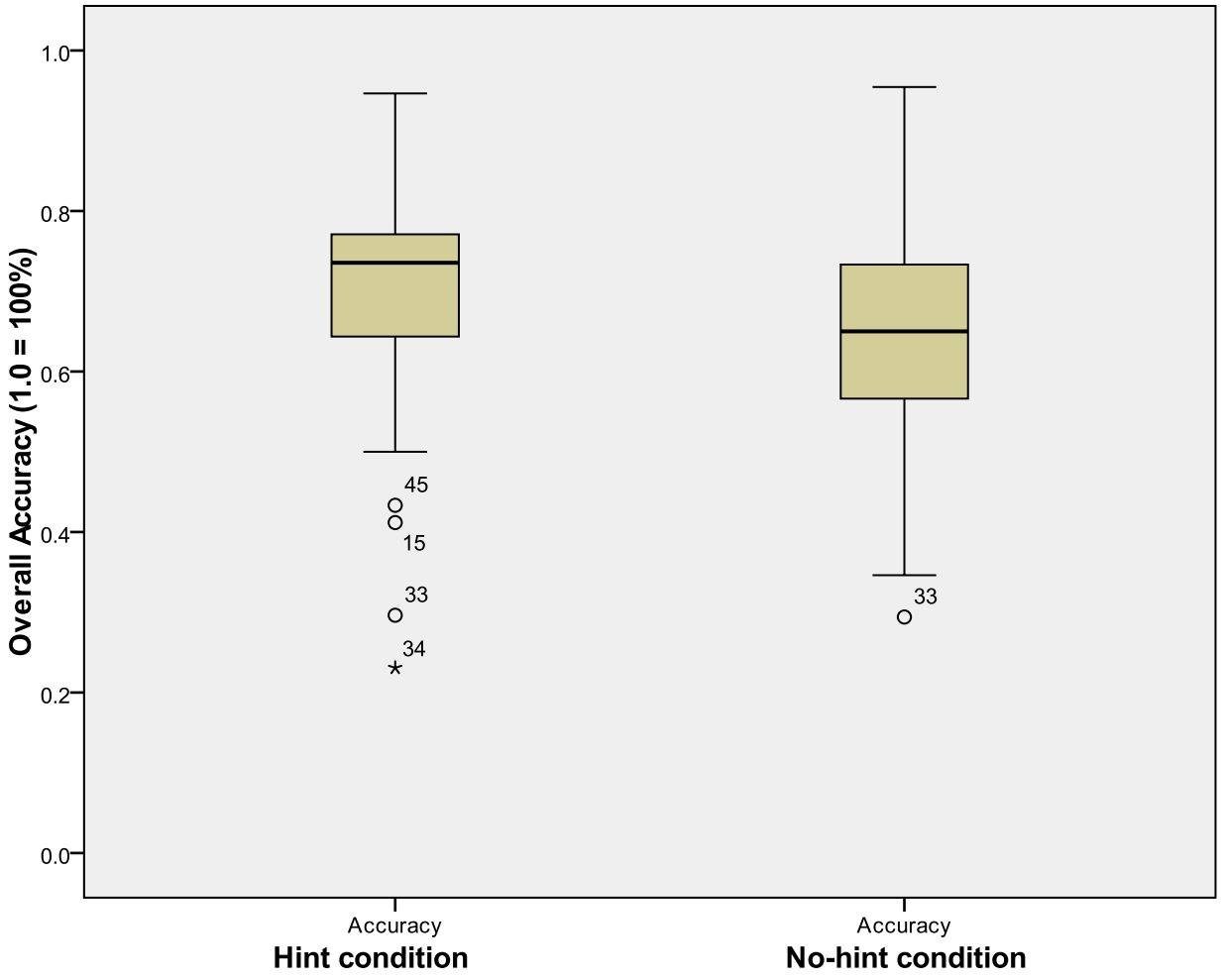


Figure 14. Accuracy measurements in the hint condition and no-hint condition

Three paired-samples t-tests were conducted to determine the effect of level of support on latency 1, latency 2, and accuracy in the hint and no-hint conditions. For the paired samples t-test of the latency 1 measurement, the 95% CI for the difference in means is -99.52, 1090.24 ($t(39) = 1.684$, $p = .100$, two-tailed, Cohen's $d = 0.22$). Although there is a general trend of latency 1 being longer in hint trials than in no-hint trials, the difference in the means fails to reach significance at the level of $\alpha = .05$.

For the paired samples t-test of the latency 2 measurement, the 95% CI for the difference in means is 266.20, 870.38 ($t(39) = 3.805$, $p < .001$, two-tailed, Cohen's $d = 0.43$). The

difference in the means is highly significant with a medium effect size. It is clear that latency 2 is longer in hint trials than in no-hint trials, and that the magnitude of the effect is considerable.

Finally, for the paired samples t-test of the accuracy measurement, the 95% CI for the difference in means is .01893 to .08767 ($t(39) = 3.137$, $p = .003$, two-tailed, Cohen's $d = 0.04$). The difference in the means is highly significant, but with a very small effect size.

3.5.2 Test for the effect of Arabic or Non-Arabic background

Due to Arabic speakers' preference for speaking tasks over writing tasks (see section 3.1 above) and their difficulties in processing English orthography (Martin, 2011; Thompson-Panos & Thomas-Ružić, 1983), the decision was made to compare Arabic speakers' performance in using the tutor against all non-Arabic speakers' performance. The latter group is not expected to comprise a single cohesive population, of course, but if we accept the notion that Arabic speakers experience truly exceptional difficulties with English orthography compared to speakers of other L1s, then this comparison is warranted. Descriptive statistics of this comparison are presented below in Table 19.

Table 19. Rehearsal data of Arabic vs. non-Arabic subjects

Population	Condition	Measurement	Mean	Std. Dev.	N
Arabic	**Hint**	Latency 1	10709 ms	2515.15	23
		Latency 2	4480 ms	1421.23	23
		Accuracy	64.8%	.15565	23
	No-hint	Latency 1	10198 ms	2501.40	23
		Latency 2	3882 ms	969.68	23
		Accuracy	62.9%	.13481	23
Non-Arabic	**Hint**	Latency 1	9346 ms	1897.75	16
		Latency 2	3958 ms	1652.04	16
		Accuracy	74.5%	.13839	16
	No-hint	Latency 1	8937 ms	1676.82	16
		Latency 2	3466 ms	1295.42	16
		Accuracy	63.9%	.14685	16

Descriptively, the most obvious difference between the two groups is that Arabic speakers' accuracy improved by 3% in the hint condition (62.9% to 64.8%), whereas non-Arabic speakers showed an average improvement in accuracy of 16.6% (63.9% to 74.5%). Box plots of these participants' rehearsal statistics are given below as Figure 15, Figure 16, Figure 17, and Figure 18.

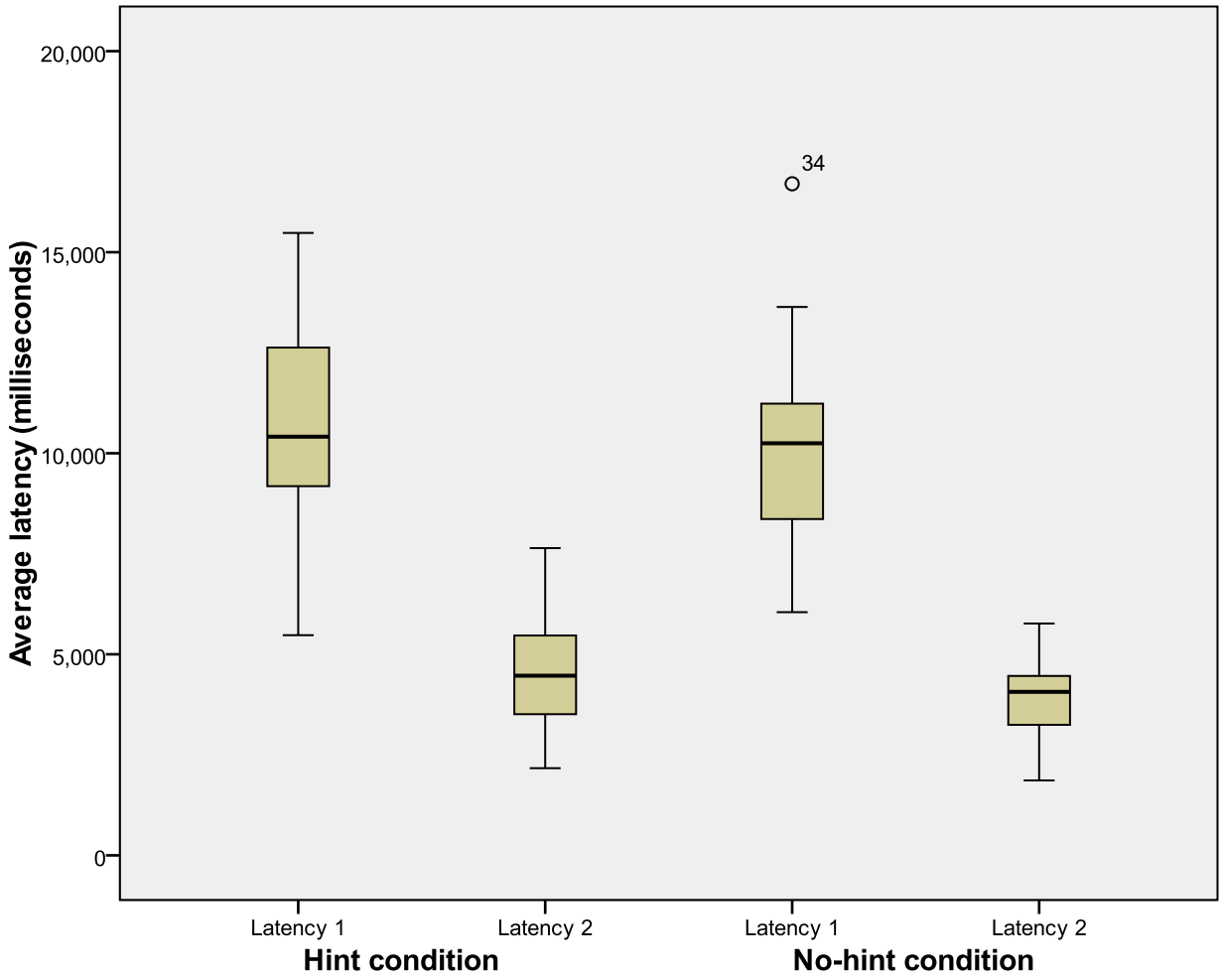


Figure 15. Latency measurements in the hint condition and no-hint condition for Arabic speakers

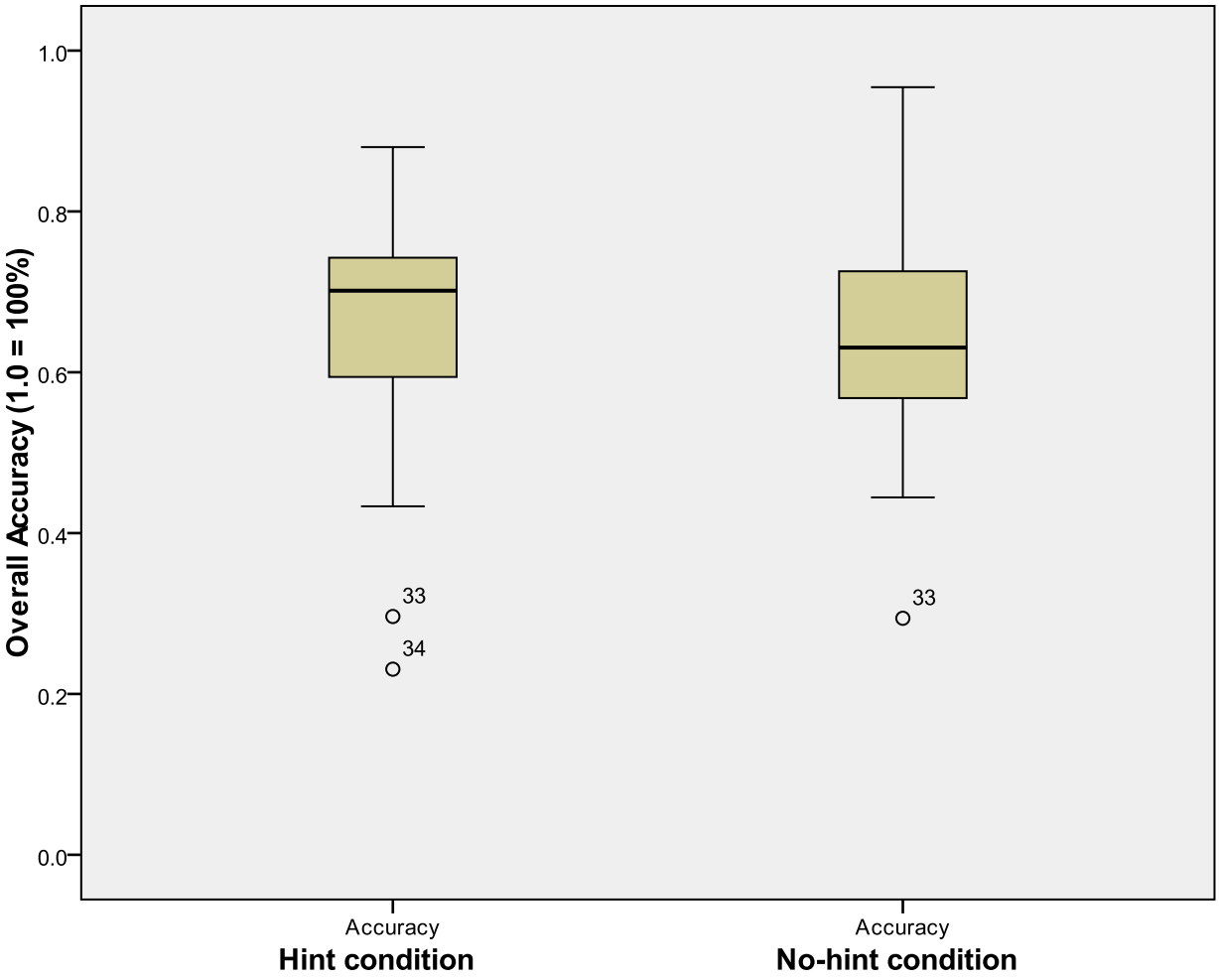


Figure 16. Accuracy measurements in the hint condition and no-hint condition for Arabic speakers

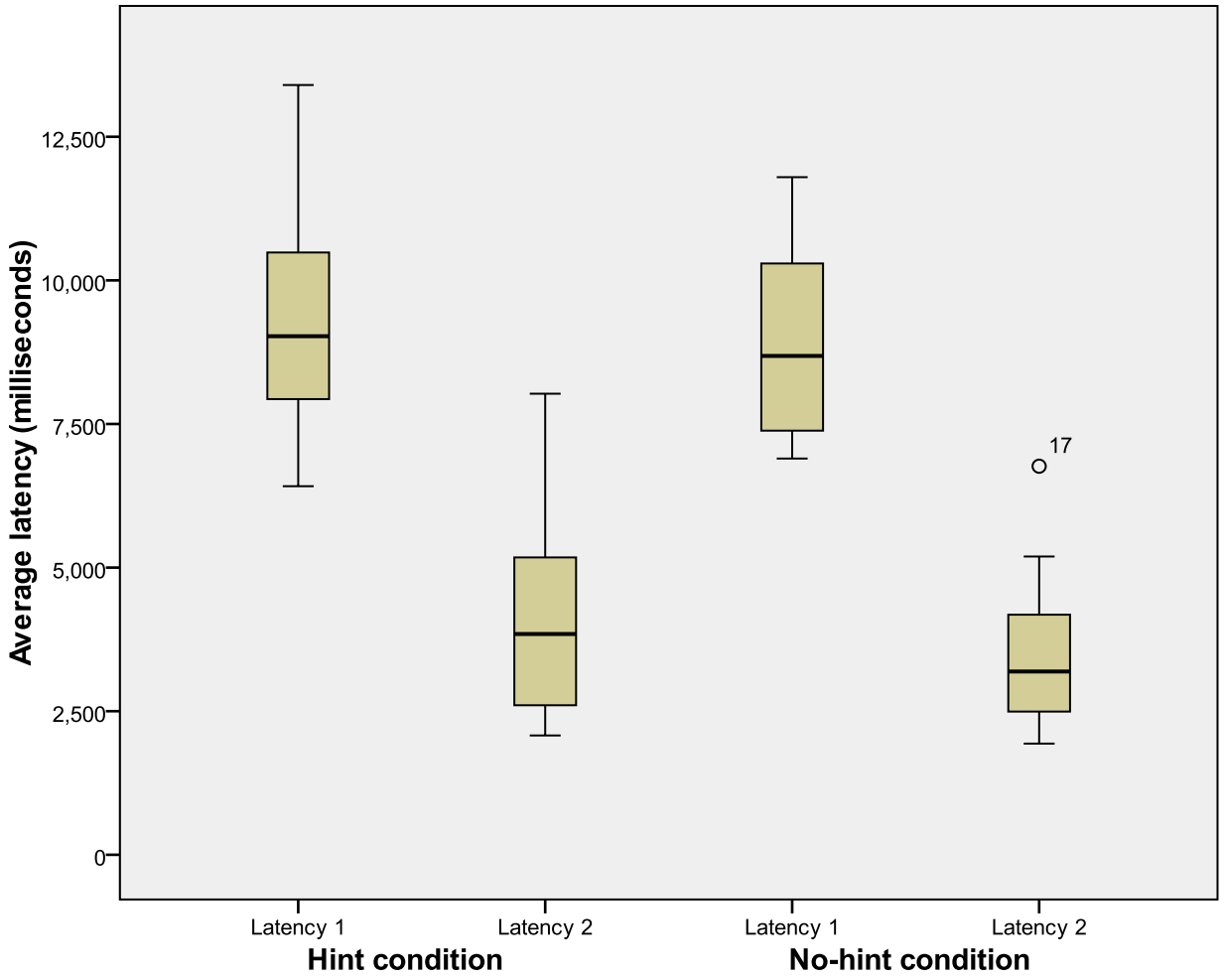


Figure 17. Latency measurements in the hint condition and no-hint condition for non-Arabic speakers

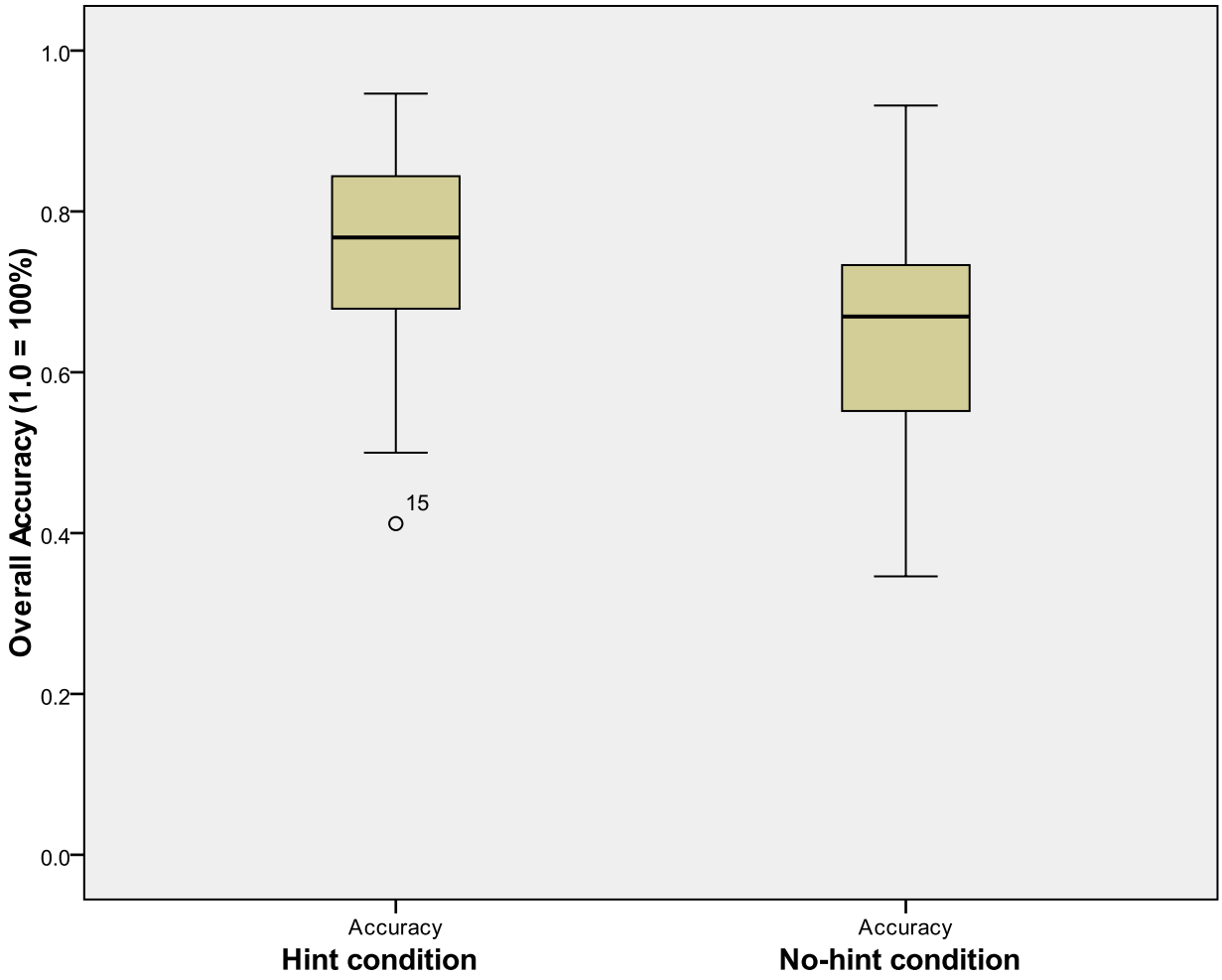


Figure 18. Accuracy measurements in the hint condition and no-hint condition for non-Arabic speakers

A repeated-measures factorial ANOVA was conducted on the data in Table 19 above. Participants' pre-test scores on the target vocabulary items were analyzed as a covariate. A two-way analysis of variance yielded a significant main effect for level of support, $F(3, 34) = 3.016$, $p = .043$, $\eta^2 = .210$. The main effect of language background (i.e., Arabic vs. non-Arabic) was non-significant, $F(3, 34) = 1.451$, $p = .24$. However, the interaction effect was significant, $F(3, 34) = 3.030$, $p = .043$, $\eta^2 = .211$, indicating that the support effect was less for Arabic speakers than for non-Arabic speakers. More specifically, a test of within-subjects contrasts yielded a significant effect of the interaction of level of support and language background upon accuracy,

$F(1, 36) = 8.757$, $p = .005$, $\eta^2 = .196$. This interaction did not yield a significant effect on latency 1, $F(1, 36) = 0.002$, $p = .97$; nor did it yield a significant effect on latency 2, $F(1, 36) = 0.309$, $p = .58$.

Following the discovery that the interaction effect between level of support and language background was significant with regard to accuracy, a pair of paired-samples t-tests were conducted on accuracy in the hint condition versus accuracy in the no-hint condition for Arabic speakers and non-Arabic speakers. The descriptive statistics for these comparisons are presented below as Table 20 and Table 21:

Table 20. Arabic speakers' accuracy in the hint and no-hint conditions

Condition	N	Mean Accuracy	Std. Dev
Hint	24	65.22%	.15339
No-hint	24	63.39%	.13379

Table 21. Non-Arabic speakers' accuracy in the hint and no-hint conditions

Condition	N	Mean Accuracy	Std. Dev
Hint	16	74.45%	.13839
No-hint	16	63.87%	.14685

For the Arabic speakers, the 95% CI for the difference in means is -.02748 to .06414 ($t(23) = .828$, $p = .416$, two-tailed). This result demonstrates that the Arabic speakers in fact show no significant improvement in accuracy at all when provided with a hint. For the non-Arabic speakers, the 95% CI for the difference in means is .06087 to .15063 ($t(15) = 5.022$, $p < .001$, two-tailed, Cohen's $d = 1.261$). This comparison confirms the significant affect of level of support on accuracy found above, but whereas the previously-reported effect size was minuscule (Cohen's $d = 0.04$), the effect size in this comparison is very large (Cohen's $d = 1.261$), as it is no longer diluted by the Arabic speakers' homogeneity between the two support conditions. These

two t-tests provide strong evidence that Arabic speakers responded to hints in a qualitatively different way than did non-Arabic speakers.

3.6 ATTRITION STATISTICS

Only two participants out of 46 remained faithful to the studying schedule (see Appendix F). The remainder (N = 44) stopped using the rehearsal software at some point during the experiment. Descriptive statistics for attrition from the study are presented below as Table 22.

Table 22. Attrition statistics

Dropped out by the end of…	N	Percent of participants
Week 1	**31**	**67.4%**
Week 3	**8**	**17.4%**
Week 6	**5**	**10.9%**
Total	**44**	**95.7%**

The attrition followed a general pattern: a very large number of subjects dropped out almost immediately; a medium number dropped out during the first three weeks; and a small number dropped out in the second three weeks. Only two participants persisted all the way through to week 8. Put differently, only 15 participants (32.6% of the total pool) chose to use the rehearsal software at all past the first week of the experiment. One might suppose that the 31 participants who did not continue to use the software had perhaps misunderstood the directions or were otherwise confused about the project, whereas the 15 participants who did continue understood the project and the expectations of the rehearsal schedule. However, participants' teachers continued to give them biweekly reminders to use the rehearsal software, suggesting that communication was not the problem. In addition, by the time week 7 began, only two

participants (4.3% of the total pool) were still using the rehearsal software. This fact suggests that many participants simply did not wish to use the software, and perhaps having realized that their usage of the software constituted 1% or less of their final course grades, they felt little motivation to participate further in the activity.

An analysis was conducted to determine whether participants' self-rated affinity for L2 writing activities (see section 3.1 above) correlated with that participants' date of final rehearsal. Assuming that higher affinity for L2 writing activities would correlate with higher motivation for completing the present computer-assisted language learning task, and assuming that higher motivation for completing the task would correlate with longer periods of time before attrition, then we might expect a positive relationship between affinity for L2 writing activities and date of last rehearsal session using the software. This relationship is plotted below in Figure 19.

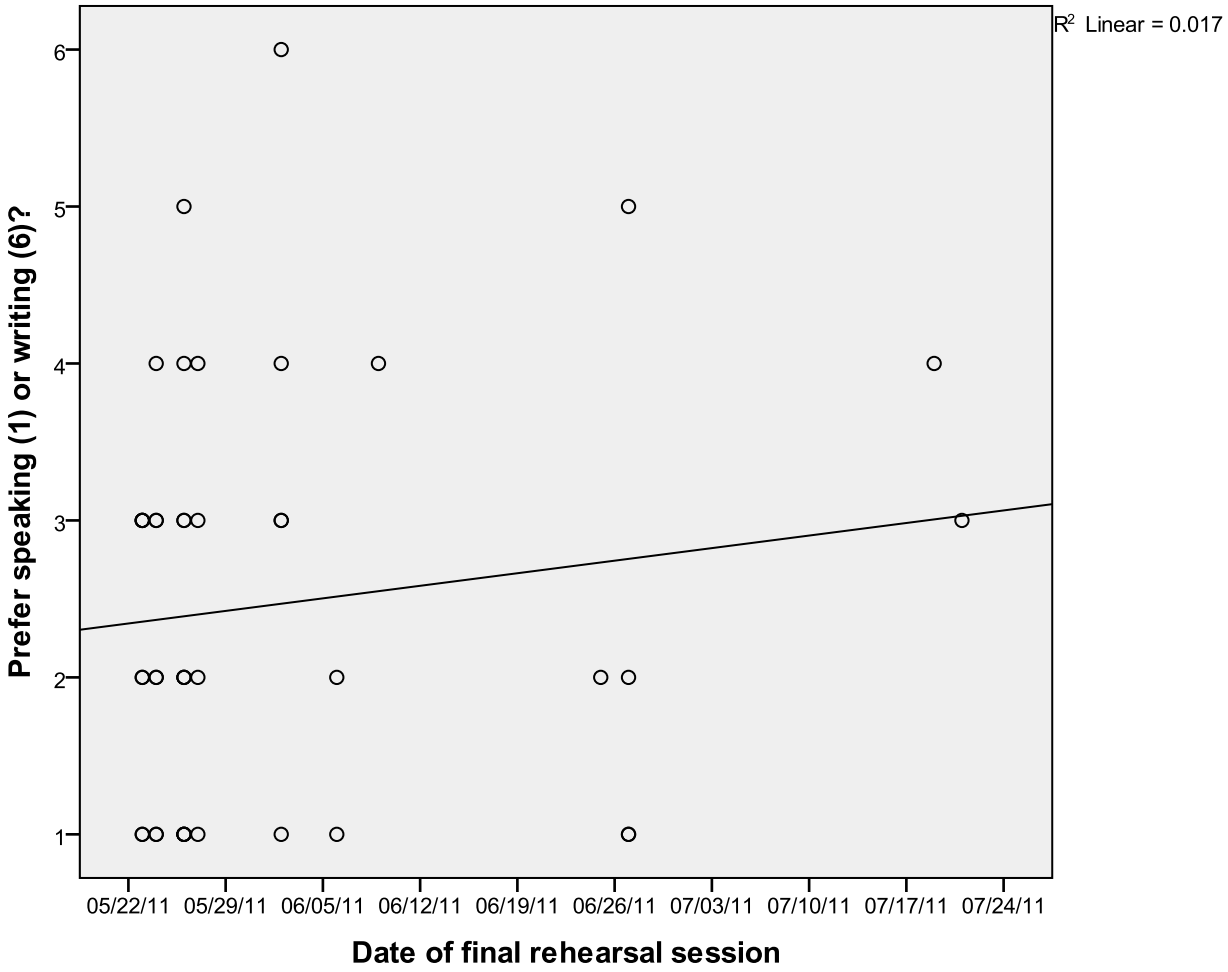


Figure 19. Plot of participants' affinity for writing tasks vs. date of final rehearsal session

Descriptively, Figure 19 above shows a slight positive correlation between a participant's affinity for writing tasks vs. that participant's date of final rehearsal session. A Pearson correlation was carried out on affinity for writing tasks and date of final rehearsal session. The test revealed that there was not a statistically significant correlation between these variables, r = .132, N = 45, p = .387, R^2 = 0.017. Thus, the data from the present experiment contradict the notion that a higher affinity for the modality of the rehearsal task should correlate with participants' continuing motivation to complete that task.

Especially relevant to the assistance dilemma is the question of how accuracy on task affects motivation. A plot of on participants' accuracy as measured at the end of week 1 and the date of each participant's final rehearsal session is presented below as Figure 20.

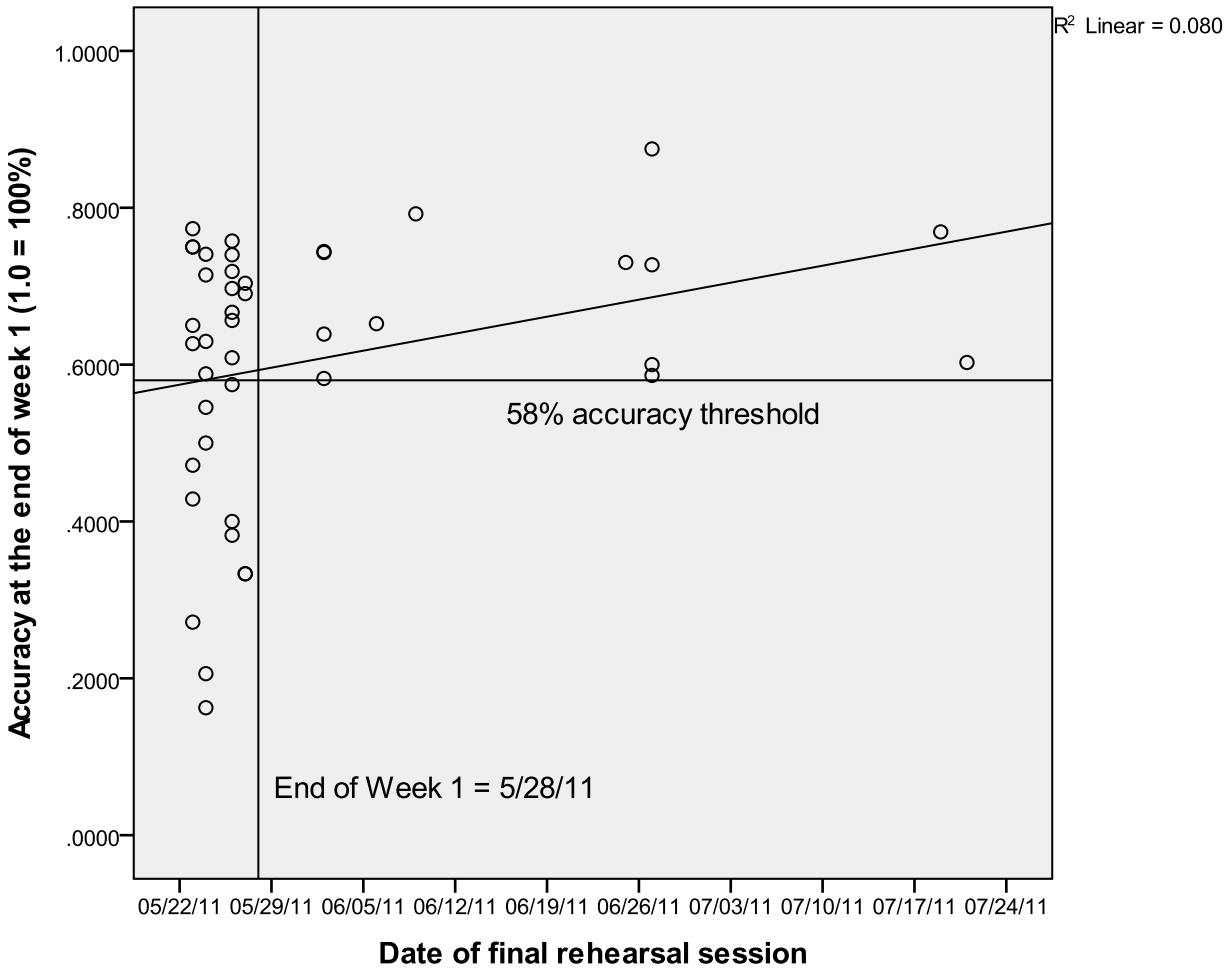


Figure 20. Plot of participants' average accuracy at the end of week 1 vs. date of final rehearsal session

Descriptively, these data form a sort of right angle, with a vertical line of plotted points before the end of week 1 and a horizontal line of plotted points after the end of week 1. This suggests a positive correlation, albeit a somewhat messy one. A Pearson correlation was carried out on participants' accuracy as measured at the end of week 1 and the date of each participant's

final rehearsal session, including those who did not drop out. The test revealed that there was not a statistically significant correlation between these variables, r = .283, N = 43, p = .066, R^2 = 0.08. However, it seems notable that a considerable cohort of low-accuracy participants (ranging from approximately 15% to approximately 57% accuracy) simply vanishes after the first week. After week 1, only participants with 58% or greater accuracy remained in the study, leading to a sharp "right angle" visible in the distribution pivoting on the end of week 1 on the X axis and 58% accuracy on the Y axis. The confounding issue is that a generous number of higher-accuracy participants also dropped out before week 2. It seems likely that accuracy is a good predictor of attrition, but that at least one other motivational factor is not being accounted for in this comparison, leading to a result which falls just short of significance.

Similarly relevant to the assistance dilemma is the question over whether the average amount of time per trial (i.e., latency 1 + latency 2) correlates with attrition. One might imagine a priori that participants who complete tasks more quickly might feel a heightened sense of accomplishment, in which case the delay introduced by the time needed to process additional scaffolding may actually harm motivation. A plot of participants' average latency per trial and the date of each participant's final rehearsal session is presented below as Figure 21.

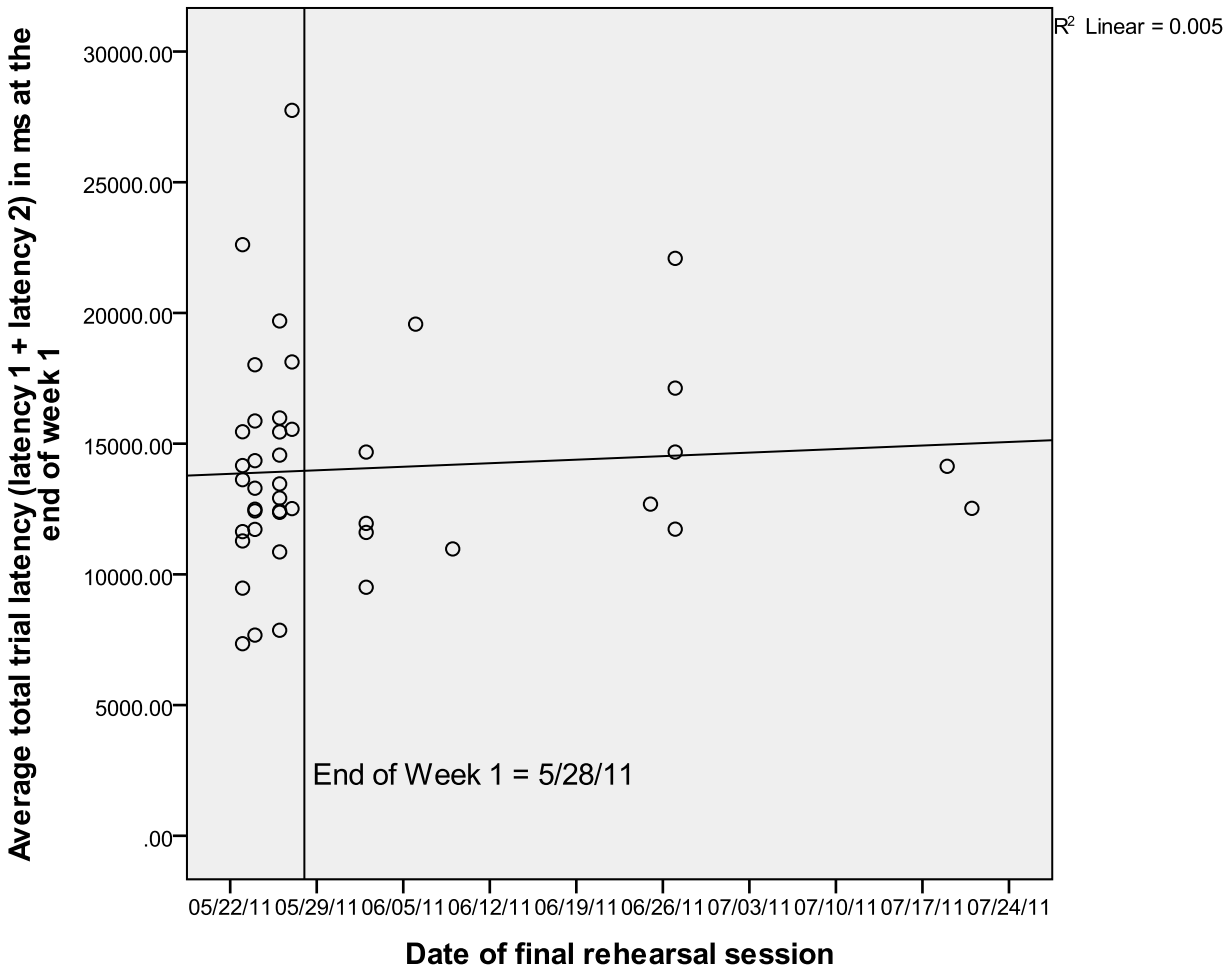


Figure 21. Plot of participants' average time on task per trial vs. date of final rehearsal session

Descriptively, the participants to the left of the "week 1" dividing line appear to cluster similarly to the participants to the right of that line, suggesting no correlation between these measurements. A Pearson correlation was carried out on average latency per trial as measured at the end of week 1 and the amount of time each participant remained in the study. The test revealed that there was not a statistically significant correlation between these variables, $r = .073$, $N = 43$, $p = .641$, $R^2 = 0.005$. However, there may actually be two phenomena at work in these data: outliers who took over 20 seconds per trial tended to drop out before week 2, and outliers who took fewer than 10 seconds per trial also tended to drop out before week 2. Perhaps the

former participants found the task too difficult, while the latter participants did not feel inclined to try very hard on any individual trial, opting instead to intentionally submit a blank answer or a wrong answer so that they could see the target answer—a sort of "gaming the system" which forces the retrieval-based task to act like an encoding-based task. It would not be too surprising for such participants to drop out in either case; however, this is just speculation.

4.0 DISCUSSION

4.1 GENERAL DISCUSSION

The data gathered during the present experiment yielded four primary insights. First, as demonstrated in section 3.1, L1 speakers of Arabic and L1 speakers of East Asian languages differed significantly in their preferences toward speaking and writing in English, with the former group preferring speaking more strongly than the latter group; however, they did not differ significantly in their preferences toward listening or writing in English. Second, as demonstrated in section 3.5.1, the presence of a hint tended to increase the amount of time participants spent reading a prompt (latency 1), but not significantly; it significantly increases the amount of time participants spent typing and checking their work (latency 2); and it significantly increased participants' accuracy. Third, as demonstrated in section 3.5.2, L1 speakers of Arabic differed significantly from all other participants in the study, regardless of L1 background, in that the Arabic speakers' accuracy on the cloze exercise task saw no significant improvement in the hint condition, whereas non-Arabic speakers' accuracy was substantially improved with a very large affect size (Cohen's d = 1.261). Fourth, as demonstrated in section 3.6, the length of time participants continued to use the rehearsal software did not significantly correlate with participants' affinity for writing tasks; nor did it correlate with their average accuracy or average

latency per trial, as measured at the end of Week 1. However, in the case of average accuracy, the correlation closely approached significance, $p = .066$.

Regarding Arabic and East Asian speakers' preferences regarding speaking or writing in English, the results of the background survey corroborate recent findings of Juffs, Friedline, Wilson, Eskenazi, & Heilman (forthcoming). Juffs et al. determined that that Arabic-speaking students were less text-oriented than Korean- and Chinese-speaking students. In their study, Arabic speakers tended to favor oral interaction in vocabulary learning, whereas they strongly disliked reading as a method of acquiring new vocabulary. By contrast, speakers of Chinese and Korean singled out reading and writing as being particularly good ways of learning and practicing vocabulary. However, the picture is somewhat muddled by the finding in this study that Arabic speakers are largely ambivalent regarding the choice between listening and reading tasks; for these receptive tasks, their preferences mirror those of East Asian students. This finding does not corroborate the assertions of Juffs et al. Perhaps Arabic speakers' specific aversion to writing tasks makes sense in light of the fact that Arabic speakers demonstrate exceptional difficulty processing English orthography and thus find English spelling unusually challenging (Martin, 2011). The present study did not accommodate Arabic speakers' attitudes or special needs toward text-based tasks, but this is an avenue likely to be fruitful in future research (see section 5.1.2 below).

Regarding the effect of hints upon participants' rehearsal statistics, the presence of hints was a net positive for participants. Time on task did significantly increase on the "typing and checking work" phase of cloze exercise completion (latency 2), but the difference in absolute terms was small: 3.7 seconds without a hint and 4.3 seconds with a hint. The more interesting effect occurred on accuracy. Without hints, participants achieved 63.6% accuracy on average.

With hints, participants achieved 68.9% accuracy on average, which was still well below ceiling level. This suggests that the hint assisted participants in successfully completing the cloze exercise without making the task trivially easy or shallow. Further evidence for this analysis comes from a sampling of the most prolific studiers' data. Participants who completed more than 500 total trials (n = 2)—indicating extensive review and practice of previously-completed cloze exercises—demonstrated an average accuracy of 85.3% on no-hint trials and an average accuracy of 85.7% on hint trials. This narrowing of the gap between no-hint and hint condition accuracy suggests that the hints are helpful for getting students started on the task, but that their power of scaffolding is not so great as to diminish the difficulty or value of the exercise over the long run.

Regarding the limited benefit hints had for Arabic speakers compared to all other L1 backgrounds, a few possible hypotheses could explain this gap. One possible explanation is that Arabic speakers did not attend to the hints as much as non-Arabic speakers did. This explanation is not supported by the data, however. If one group were to systematically ignore the hints, we would expect that group's time on task not to increase in the hint condition. However, both the Arabic speakers and the non-Arabic speakers' time on task did increase, suggesting that both groups were attending to the hints. Another possible explanation is that Arabic speakers were less likely to use the hints in a beneficial way. For example, they might have had difficulty coping with the increased cognitive load which the hints represented, leading to difficulty comprehending the hints or mapping them onto the vocabulary they were studying. The present study does not have any data which could support or disconfirm this hypothesis. A final explanation would be that the hints did not help to alleviate a major source of Arabic speakers' errors. If, for example, Arabic speakers struggled more with correct spelling than with lexical

choice in the cloze exercise task, then the hint would not be helpful to them. This explanation is plausible, but more specific analysis of Arabic speakers' and non-Arabic speakers' errors would be required to confirm or reject it.

Regarding the lack of a correlation between the amount of time participants stayed in the study and their affinity for L2 writing tasks, there are a few possible explanations. First, the sentence completion task employed in the present study may not have been a prototypical writing task from the perspective of the participants. Perhaps some participants enjoy the creative aspects of writing, whereas the present intervention had no creative aspect to it. Perhaps other participants find writing tasks more enjoyable than speaking tasks because there is usually less time pressure, whereas the present intervention had explicit time pressure—participants had only 30 seconds to complete each trial. In short, the present intervention may have lacked certain affective or processing advantages which participants usually associate with writing tasks. Another possible explanation for the lack of correlation between participants' affinity for L2 writing tasks and the length of time they continued to use the rehearsal software is that the relationship was confounded by the effect of other affective or motivational variables. Regardless of how enjoyable it is, if an activity seems too easy, too hard, or just irrelevant to a participant's goals, a participant may see no reason to continue doing it. Even a text-oriented student who loves reading and writing tasks, for example, may stop participating in an activity if that student does not feel he or she is capable of doing the task or benefitting from it.

Some measure of support for that notion comes from the correlation results between the amount of time participants stayed in the study and their average accuracy as measured at the end of week 1. Although this correlation failed to reach significance (p = .066), it is striking that every participant whose average accuracy was below 58% dropped out before week 2 began.

There is a clear tendency for accuracy and attrition to correlate, and it seems likely that the failure for the correlation to reach significance is due to confounding factors which led even some of the higher-accuracy participants to drop out early. Thus, in terms of the assistance dilemma, we would predict lower (or at least, more gradual) attrition if the level of scaffolding could be increased for the most inaccurate participants. In their cases, the benefit of achieving the thrill of supported success and decreasing the frustration of failure would likely outweigh any costs pertaining to shallower processing or lack of attention (cf. Table 4).

4.2 THEORETICAL IMPLICATIONS

The present study combined multiple strands of research from the fields of TESOL, SLA, and cognitive science into a single program. One of its theoretically-oriented goals was to measure variables relevant to the assistance dilemma, and the results of the present study indicate that the level of support provided on the cloze exercise task—namely, a short written definition of the target word—was appropriate in increasing participants' accuracy without making the task trivially easy. However, this scaffolding was of no apparent utility to Arabic speakers compared to speakers of all other native languages, demonstrating that the selection of scaffolding is not only task-dependent and skill-level-dependent, but also dependent on unique factors pertaining to sociocultural background and/or cognitive processing differences between groups. Thus, it is an oversimplification to view scaffolding from only a quantitative perspective, i.e. to judge "how little" or "how much" scaffolding there is. If different students respond in qualitatively different ways to a single type of scaffolding rather than falling on a continuum, then this issue must be taken into account in providing both the proper *type* of scaffolding for the student while also

correctly manipulating the level of support. In addition, a correlation conducted on the length of time participants remained in the study and their average accuracy as measured at the end of week 1 is strongly suggestive of a connection between these variables, but failed to reach significance ($p = .066$). If we tentatively entertain this connection, there is a case to be made that assistance is not only beneficial to overall learning outcome, but can also be a critical factor in whether a student remains motivated to continue a learning activity or not—a concept relevant to the assistance dilemma.

4.3 IMPLICATIONS FOR PEDAGOGY

The present study largely failed in its capacity as a learning intervention. Most participants chose to drop out before the conclusion of the study, and thus, the intervention did not yield reliable data on educational outcomes due to the paucity of post-test data. In addition, the results of this intervention cannot be applied to a general evaluation of cloze exercises as a method of written vocabulary practice because the computer-based nature of the task and the strict grading employed are not immediately comparable to the format used by, e.g., Folse (2006). This particular intervention did not appear to motivate or engage its participants, leading to high attrition, and most of the problems and weaknesses which arose in the present study stem from that central fact. Therefore, future research would do well to better isolate the motivational and affective factors relevant to this and other forms of computer aided language learning (see section 5.1.1 below) and to establish best practices for the types of features and feedback which ought to be implemented in CALL tools in order to best motivate students. As mentioned above in section 4.1, data gathered in the course of this study suggest that accuracy on task could be a

critical factor in driving motivation to continue a learning activity, though the result failed to reach significance (p = .066).

5.0 CONCLUSION

The present study demonstrates that written definitions are an appropriate form of assistance for cloze exercises, yielding an acceptable tradeoff between increased accuracy and increased time on task. It also establishes that Arabic speakers differ significantly from speakers of other L1 backgrounds, both in terms of their negative attitudes toward writing in English and in terms of their lack of success making use of hints on digital cloze exercises. However, this study is just the first step in a larger research program, and it raises more questions than it answers. Substantial research must be done in the future to address problematic areas of the present study and to expand the scope of the investigation to other areas.

5.1 LIMITATIONS OF THE CURRENT STUDY AND FUTURE DIRECTIONS

One limitation of the present study is that it was not possible to measure accurate educational outcomes due to a number of factors, including a mismatch in frequencies between the target and control word lists, the large proportion of participants who had previously studied the target word list in earlier classes, the high attrition during the weeks of assigned tutor usage, and the anemic level of response to the post-test. Thus, although rehearsal data was gathered and analyzed for 40 participants, it is impossible to view most of that rehearsal data within the greater context of a pre-test/post-test design, because only four participants opted to complete the

post-test. It is clear that assigning the post-test as a homework assignment was a mistake; in the future, any measurement of learning outcome ought to be done as an in-class activity.

It is also difficult to make generalizations about participants' attitudes toward the tutor and what factors led participants to either continue to use the tutor or drop out of the study. Only three participants responded in good faith to the post-intervention opinion survey, and because these participants tended to be highly-motivated students who either did not drop out of the study or else dropped out unusually late, their opinions were correspondingly quite mild. Thus, at present, it is difficult to target any specific areas for improvement.

There are several modifications and directions for future study proposed in the sections below.

5.1.1 Stronger focus on student motivation and engagement

An immediate future direction for research is to determine what about the present tutor is repellant to students and what modifications to the rehearsal interface and methodology can be made to increase participants' engagement with the learning tool. Juffs et al. (forthcoming) assert that if students believe that a language-learning task does not mesh with their own goals and desires for learning English, either in terms of its face validity ("Will this help me to learn what I want to learn?") or its modality ("Do I enjoy learning this way?"), then the effect on motivation and engagement with the learning tool can be catastrophic. Although none of the participants who gave their opinion of the vocabulary tutor on the post-intervention survey were harsh in their appraisals, the fact that 42 out of 46 participants failed to complete the online survey after several reminders suggests that those participants who did drop out felt little to no personal investment in their vocabulary tutor rehearsal sessions, and thus did not wish to be bothered with

a survey about a tool they viewed as irrelevant or boring. This lack of engagement likely underlies the attrition statistics witnessed in section 3.6: most participants likely had a negative first or second impression with the software and simply decided not to use it, regardless of the fact that their participation was counted for course credit. To that end, future efforts will focus on piloting interventions with smaller groups and with a stronger emphasis on affective and motivational variables.

One specific issue with the present tutor is that it may have been poorly calibrated relative to the difficulty of the task and may have been too eager to present new material before participants had sufficiently mastered older material. This tendency would lead to poorer accuracy on task. Results presented in section 3.6 demonstrated a relationship between participants' accuracy and their tendency to drop out of the study, with more inaccurate students dropping out more quickly than more accurate students, but this correlation fell just short of significance, $p = .066$. On average, participants who completed at least 20 total trials demonstrated accuracy levels of 68.9% in the hint condition and 63.6% in the no-hint condition. Success rates this low were likely demoralizing to participants. In the future, such a tutor ought to be calibrated better to ensure that all participants achieve success rates closer to 85-95%. In addition, although it is promising that the hints helped participants to achieve greater success in the task, it is likely that the small magnitude of the boost was not intuitively felt by participants. Thus, they may have received the practical benefits of scaffolding (higher accuracy) without experiencing the corresponding affective boost (enjoying a greater rate of success).

5.1.2 More specific focus on the needs of Arabic speakers

It has been demonstrated that Arabic speakers have exceptional difficulty processing English orthography due to some form of L1 interference (Martin, 2011) and that Arabic speakers are pointedly not text-oriented when it comes to learning and studying English vocabulary (Juffs et al., forthcoming). The present study addressed neither of these issues. One specific issue which Arabic speakers likely struggled with is that of correct spelling. Due perhaps to their difficulties processing written vowels of English vowels, Arabic speakers are notoriously poor spellers. The vocabulary tutor, however, did not accept misspelled input as valid—misspelled words were rejected the same way that incorrect words were. Although the 1,038 failed cloze exercise trials were not coded for type of error (e.g., misspelling, typo, incorrect word, no answer, etc.), misspellings certainly exist among the incorrect responses. If Arabic speakers had encountered fewer cases in which their correct—but incorrectly spelled—answer was counted wrong, perhaps they would have felt more motivation to continue using the tutor.

5.1.3 Incorporation of other measures of participants' actions, such as eyetracking

One of the puzzles of the present study is why the addition of a hint did not significantly increase the amount time participants spent reading each prompt before pressing their first keystroke (latency 1), but did significantly increase the amount of time they spent typing and checking their work (latency 2). This seems to contradict common sense: a significantly larger prompt (in the form of a cloze exercise plus a hint) ought to take longer to read than a shorter prompt, and it ought not have any effect on the amount of time a participant spends typing a response. An

eyetracking study of a small group of participants would help to elucidate exactly where the participants are directing their attention during each phase of the task.

5.1.4 Focus on other aspects of vocabulary learning, such as the development of polysemy and derivational word families

In the present study, any given target vocabulary item was only used in one part of speech and in only one or two highly-frequent definitions. In other words, this intervention focused on breadth of practice across many target items, not depth of practice within a single word or family of words. Modifications of this study could instead focus on providing deep practice of just a few target words or target word families, helping students to explore the different derived forms or different definitions of a small word list.

APPENDIX A

TARGET ITEMS: CORE VOCABULARY LEVEL 4

Table 23. Target items: Core Vocabulary Level 4

Week	Words
2	achieve aspect assess concept cooperate
3	acquire initial obvious relevant target
4	approximate demonstrate evaluate modify previous
5	factor imply method resource valid
6	affect category distinct evident perspective

7	establish feature potential range secure
8	complex constant distribute equate indicate
9	alternative correspond exclude incorporate rely
10	compatible consequence principal region restrict
11	considerable illustrate impact perceive regulate

APPENDIX B

FREQUENCY VALUES OF CORE VOCABULARY LEVEL 4

Table 24. Frequency values of Core Vocabulary level 4

Frequency Value	Word count	Words
BNC-1,000	**8**	achieve affect obvious previous range region resource secure
BNC-2,000	**26**	alternative aspect assess category complex concept considerable constant demonstrate distinct distribute establish exclude factor feature impact indicate initial

		method potential regulate relevant rely restrict target valid
BNC-3,000	**5**	approximate correspond illustrate incorporate principal
BNC-4,000	**8**	acquire consequence cooperate evident imply modify perceive perspective
BNC-5,000	**2**	compatible evaluate
BNC-6,000	**1**	equate

APPENDIX C

CONTROL ITEMS: CORE VOCABULARY LEVEL 5

Table 25. Control items: Core Vocabulary Level 5

Week	Words
2	acknowledge clarify discriminate eliminate reluctant
3	abstract crucial inherent rational stable
4	ambiguous coordinate generate parallel trigger
5	access anticipate enhance reveal underlie
6	attribute bias exhibit philosophy violate

7	adapt emerge framework justify trend
8	compensate diverse impose sustain unify
9	arbitrary fluctuate inevitable manipulate transform
10	implement integrate maximize pursue reinforce
11	implicit negate persist proportion retain

APPENDIX D

FREQUENCY VALUES OF CORE VOCABULARY LEVEL 5

Table 26. Frequency values of Core Vocabulary Level 5

Frequency Value	Word count	Words
BNC-1,000	**0**	
BNC-2,000	**9**	access clarify compensate exhibit generate impose justify proportion stable
BNC-3,000	**8**	adapt coordinate crucial discriminate framework reveal transform trigger
BNC-4,000	**17**	acknowledge anticipate attribute bias enhance fluctuate implement inevitable

		integrate manipulate persist philosophy pursue retain sustain trend underlie
BNC-5,000	**12**	abstract ambiguous arbitrary diverse eliminate emerge inherent maximize parallel reinforce reluctant
BNC-6,000	**2**	implicit rational
BNC-7,000	**0**	
BNC-8,000	**2**	negate violate

APPENDIX E

INVENTORY OF CLOZE EXERCISES

Table 27. Inventory of cloze exercises

Item	Week	Word	Cloze Exercise	Hint
1	2	achieve	It takes a lot of studying to _____ good grades.	To succeed in doing something important by your own effort.
2	2	achieve	You can _____ any goal if you work hard and believe in yourself.	To succeed in doing something important by your own effort.
3	2	achieve	It is important to set realistic goals that you can _____.	To succeed in doing something important by your own effort.
4	2	aspect	The cost of tuition is an important _____ to consider when researching colleges.	One part of a situation, idea, plan, etc that has many parts.
5	2	aspect	One interesting _____ of the building's design is its use of solar energy.	One part of a situation, idea, plan, etc that has many parts.
6	2	aspect	One important _____ of owning a dog is training it properly.	One part of a situation, idea, plan, etc that has many parts.
7	2	assess	In an emergency it is important to _____ the situation quickly.	To make a judgment about a person or situation after thinking carefully about it.
8	2	assess	A manager should always _____ the skills of his or her employees.	To make a judgment about a person or situation after thinking carefully about it.
9	2	assess	A judge must _____ whether those laws are relevant to the case.	To make a judgment about a person or situation after thinking carefully about it.
10	2	concept	The _____ of a paperless society will help preserve forests.	An idea of how something is, or how something should be done.
11	2	concept	The _____ of human rights is important in most nations.	An idea of how something is, or how something should be done.
12	2	concept	Young children do not understand the _____ of death.	An idea of how something is, or how something should be done.

13	2	cooperate	We all had to _____ to get the project done on schedule.	To work with someone else to achieve something that you both want.
14	2	cooperate	The political parties should _____ for the benefit of the people.	To work with someone else to achieve something that you both want.
15	2	cooperate	If only he would _____, we could finish the project in no time!	To work with someone else to achieve something that you both want.
16	3	acquire	Some people _____ wealth through luck instead of hard work.	To get or gain something.
17	3	acquire	A good way to _____ knowledge is to read books.	To get or gain something.
18	3	acquire	It is not difficult to _____ new friends if you are sociable and helpful.	To get or gain something.
19	3	initial	My _____ reaction was fear, but I quickly calmed down.	Happening at the beginning.
20	3	initial	Our _____ goal is to sell 2,000 books, but I think we can sell 5,000.	Happening at the beginning.
21	3	initial	My _____ idea was to make a sandwich, but we have no bread.	Happening at the beginning.
22	3	obvious	It is _____ that you should dress warmly before playing in the snow.	Easy to notice or understand.
23	3	obvious	Cynthia's blush made her embarrassment _____ to everyone.	Easy to notice or understand.
24	3	obvious	It is _____ that we need to work together to solve this problem.	Easy to notice or understand.
25	3	relevant	Understanding the dangers of texting while driving is especially _____ for teen drivers.	Directly relating to the subject or problem being discussed or considered.
26	3	relevant	The school allowed teachers to show _____ films to their students.	Directly relating to the subject or problem being discussed or considered.
27	3	relevant	To make a good decision, you must know which information is _____ and which is not.	Directly relating to the subject or problem being discussed or considered.
28	3	target	Knowing when she wanted to retire allowed her to set a(n) _____ amount to save.	Something that you are trying to achieve, such as a total, an amount, or a time.
29	3	target	In exercising, my _____ is to take at least 10,000 steps per day.	Something that you are trying to achieve, such as a total, an amount, or a time.
30	3	target	Keep your _____ in mind to help yourself stay motivated.	Something that you are trying to achieve, such as a total, an amount, or a time.
31	4	approximate	The _____ answer is not good enough, because it has to be exact.	Close to the exact number or amount, but could be a little bit more or less than it.

32	4	approximate	The _____ distance to the mall is about five kilometers.	Close to the exact number or amount, but could be a little bit more or less than it.
33	4	approximate	I think this job will take about three hours to do, but that is only _____.	Close to the exact number or amount, but could be a little bit more or less than it.
34	4	demonstrate	The student's high test scores _____ her abilities.	To show or prove something clearly.
35	4	demonstrate	The salesman can _____ how the product works in under five minutes.	To show or prove something clearly.
36	4	demonstrate	I will now _____ how to solve the math problem.	To show or prove something clearly.
37	4	evaluate	The taste testers will _____ the flavor and texture of the new strawberry ice cream.	To judge how good, useful, or successful something is.
38	4	evaluate	The inspector had to _____ the safety of the elevator.	To judge how good, useful, or successful something is.
39	4	evaluate	It is time to _____ the employees' abilities and strengths.	To judge how good, useful, or successful something is.
40	4	modify	We should _____ the recipe to reduce the amount of sugar.	To make small changes to something in order to improve it and make it more suitable or effective.
41	4	modify	It can be very difficult to _____ our habits once we are adults.	To make small changes to something in order to improve it and make it more suitable or effective.
42	4	modify	The teacher can _____ the assignment to suit the disabled student.	To make small changes to something in order to improve it and make it more suitable or effective.
43	4	previous	You can use a(n) _____ employer as a reference when filling out a job application.	Having happened or existed before the event, time, or thing that you are talking about now.
44	4	previous	You should not read the new book until you have read the _____ one.	Having happened or existed before the event, time, or thing that you are talking about now.
45	4	previous	The tall building broke the _____ height record by ten meters.	Having happened or existed before the event, time, or thing that you are talking about now.
46	5	factor	The temperature was only one _____ affecting the result of the experiment.	One of several things that influence or cause a situation.
47	5	factor	One _____ that can harm an economy is unemployment.	One of several things that influence or cause a situation.
48	5	factor	The most important _____ in my decision was the cost.	One of several things that influence or cause a situation.
49	5	imply	If we do not protest the plan, it will _____ that we agree with it.	To suggest that something is true, without saying this directly.
50	5	imply	Although she was not upset, her frown seemed to _____ otherwise.	To suggest that something is true, without saying this directly.

51	5	imply	Are you trying to _____ I don't know what I'm talking about?	To suggest that something is true, without saying this directly.
52	5	method	Using flashcards is a good _____ of preparing for exams.	A planned way of doing something, especially one that a lot of people know about and use.
53	5	method	The new construction _____ will help us to build more houses using fewer resources.	A planned way of doing something, especially one that a lot of people know about and use.
54	5	method	It is easy to solve math problems once you understand the correct _____.	A planned way of doing something, especially one that a lot of people know about and use.
55	5	resource	Fresh water is an important natural _____.	Something such as useful land, or minerals such as oil or coal, that exists in a country and can be used to increase its wealth; or all the money, property, skills, or personal qualities that you have available when you need them.
56	5	resource	A dictionary is a useful _____ when you are learning a new language.	Something such as useful land, or minerals such as oil or coal, that exists in a country and can be used to increase its wealth; or all the money, property, skills, or personal qualities that you have available when you need them.
57	5	resource	If a country lacks a(n) _____ like oil or aluminum, it must trade with its neighbors.	Something such as useful land, or minerals such as oil or coal, that exists in a country and can be used to increase its wealth; or all the money, property, skills, or personal qualities that you have available when you need them.
58	5	valid	You have a(n) _____ passport, so we will let you cross the border.	Legally or officially acceptable; or, a reason or argument that is based on what is reasonable or sensible.
59	5	valid	You should offer a(n) _____ reason for being late for work.	Legally or officially acceptable; or, a reason or argument that is based on what is reasonable or sensible.
60	5	valid	I was not let into the concert because the ticket I purchased was not _____.	Legally or officially acceptable; or, a reason or argument that is based on what is reasonable or sensible.
61	6	affect	If a diet does not _____ someone's weight, they should try exercise.	To cause a change in something or in someone's situation.

62	6	affect	I didn't realize that my actions would _____ him so much.	To cause a change in something or in someone's situation.
63	6	affect	One partner's snoring can _____ the other partner's quality of sleep.	To cause a change in something or in someone's situation.
64	6	category	The women selected silk from the _____ of fabric options.	A group of people or things that are all of the same type.
65	6	category	Do penguins really belong to the same _____ as other birds?	A group of people or things that are all of the same type.
66	6	category	This music is so strange that it does not seem to fit into any _____.	A group of people or things that are all of the same type.
67	6	distinct	Rice milk is quite _____ from cow milk and goat milk.	Clearly different or belonging to a different type.
68	6	distinct	She sorted the mail into three _____ categories.	Clearly different or belonging to a different type.
69	6	distinct	Life in the mountains is very _____ from life in the city.	Clearly different or belonging to a different type.
70	6	evident	Once we saw all of the police cars, it was _____ that something bad had happened.	Easy to see, notice, or understand.
71	6	evident	It was _____ that the project was a total failure.	Easy to see, notice, or understand.
72	6	evident	From the available data, it is _____ that vitamin C does not cure colds.	Easy to see, notice, or understand.
73	6	perspective	Her _____ on the situation was completely different from mine.	A way of thinking about something, especially one which is influenced by the type of person you are or by your experiences.
74	6	perspective	Her work as an elementary school teacher gives her a different _____ on children.	A way of thinking about something, especially one which is influenced by the type of person you are or by your experiences.
75	6	perspective	From my _____, it is important not to waste food.	A way of thinking about something, especially one which is influenced by the type of person you are or by your experiences.
76	7	establish	The group's task was to _____ a new art museum for the town.	To start or to found a company, organization, system, etc. that is intended to exist or continue for a long time.
77	7	establish	You can _____ a business if you have some good ideas and work hard.	To start or to found a company, organization, system, etc. that is intended to exist or continue for a long time.
78	7	establish	I will _____ a company that sells folk music from around the world.	To start or to found a company, organization, system, etc. that is intended to exist or continue for a long time.

79	7	feature	The GPS system was my favorite _____ in the new car.	A part of something that you notice because it seems important, interesting, or typical.
80	7	feature	Her eyes were the _____ that I noticed the first time I met her.	A part of something that you notice because it seems important, interesting, or typical.
81	7	feature	An obvious _____ of the new television was its size.	A part of something that you notice because it seems important, interesting, or typical.
82	7	potential	Global warming is a(n) _____ threat to all living organisms on earth.	Likely to develop into a particular type of person or thing in the future.
83	7	potential	I always felt that she had the _____ to be a great actress.	Likely to develop into a particular type of person or thing in the future.
84	7	potential	He is such a good speaker that I think he has the _____ to become the next president.	Likely to develop into a particular type of person or thing in the future.
85	7	range	The _____ of ice cream flavors available these days is astonishing.	A number of people or things that are all different, but are all of the same general category.
86	7	range	The punishments _____ from hours doing community service to years in prison.	A number of people or things that are all different, but are all of the same general category.
87	7	range	There is a wide _____ of variation in the cost of food around the world.	A number of people or things that are all different, but are all of the same general category.
88	7	secure	My baby feels _____ because he knows I will pick him up if he cries.	Something safe; or something that you can depend on because it is not likely to change.
89	7	secure	My apartment building is _____ because you need a special key to get inside.	Something safe; or something that you can depend on because it is not likely to change.
90	7	secure	You should put your money in the bank to make sure it is _____.	Something safe; or something that you can depend on because it is not likely to change.
91	8	complex	The economy is so _____ that no one completely understands it.	Consisting of many different parts and often difficult to understand.
92	8	complex	The _____ math problem took 45 minutes to solve.	Consisting of many different parts and often difficult to understand.
93	8	complex	English spelling is so _____ that many native speakers are poor spellers.	Consisting of many different parts and often difficult to understand.
94	8	constant	Only _____ efforts and hard work can take you towards success.	Happening regularly or all the time.
95	8	constant	The steady flow of water in the river is _____.	Happening regularly or all the time.
96	8	constant	The one _____ in the world is that nothing stays the same.	Happening regularly or all the time.

97	8	distribute	The company will _____ bonuses to their employees at the end of the year.	To share things among a group of people, especially in a planned way; to give out.
98	8	distribute	We are collecting old blankets to _____ to the homeless.	To share things among a group of people, especially in a planned way; to give out.
99	8	distribute	Some universities _____ new computers to their students.	To share things among a group of people, especially in a planned way; to give out.
100	8	equate	Some people _____ success with money instead of happiness.	To consider two things to be similar or connected.
101	8	equate	Be careful not to _____ knowledge with wisdom.	To consider two things to be similar or connected.
102	8	equate	Shoppers began to _____ the expensive cereal with the cheap one because they tasted the same.	To consider two things to be similar or connected.
103	8	indicate	We're hoping the test results will _____ the patient is healthy.	To show that a particular situation exists, or that something is likely to be true.
104	8	indicate	The warm days _____ that winter is ending.	To show that a particular situation exists, or that something is likely to be true.
105	8	indicate	Tiredness and difficulty concentrating _____ that you need more sleep.	To show that a particular situation exists, or that something is likely to be true.
106	9	alternative	It is good to have a(n) _____ plan in case your main plan does not work.	Something like an idea or plan that is different from the main one you have and can be used instead.
107	9	alternative	Drinking water is a healthy _____ to drinking soda.	Something like an idea or plan that is different from the main one you have and can be used instead.
108	9	alternative	That idea will never work. Here is a(n) _____ solution.	Something like an idea or plan that is different from the main one you have and can be used instead.
109	9	correspond	The information you provided does not _____ with our records.	For one idea or fact to have a relationship or connection with another one.
110	9	correspond	An increase in body weight seems to _____ with diabetes and heart disease.	For one idea or fact to have a relationship or connection with another one.
111	9	correspond	Higher levels of education seem to _____ with higher salaries.	For one idea or fact to have a relationship or connection with another one.
112	9	exclude	If you _____ Harold from the group, he will feel very bad.	To deliberately not include something.
113	9	exclude	Price tags in stores usually _____ the sales tax to make items seem cheaper.	To deliberately not include something.

114	9	exclude	Be careful not to _____ anything important when packing for your vacation.	To deliberately not include something.
115	9	incorporate	We will try to _____ your suggestions into the project.	To include something as part of a group, system, plan, etc.
116	9	incorporate	The new building will _____ "green" technologies like solar panels.	To include something as part of a group, system, plan, etc.
117	9	incorporate	The restaurant's meals _____ ingredients from Indian and Chinese cuisine.	To include something as part of a group, system, plan, etc.
118	9	rely on	It is nice to have a friend you can _____, especially in times of distress.	To trust or depend on someone or something to do what you need or expect them to do.
119	9	rely on	You can _____ her work because she will finish it perfectly.	To trust or depend on someone or something to do what you need or expect them to do.
120	9	rely on	I _____ the bus to get me back and forth to work every day.	To trust or depend on someone or something to do what you need or expect them to do.
121	10	compatible	Is that program _____ with all types of smart phones?	Able to exist or be used together without causing problems.
122	10	compatible	Make sure that your career is _____ with your interests.	Able to exist or be used together without causing problems.
123	10	compatible	If you start a business with someone, make sure your goals are _____.	Able to exist or be used together without causing problems.
124	10	consequence	One _____ of natural disasters is poverty.	Something that happens as a result of a particular action or set of conditions.
125	10	consequence	The _____ of staying out late is being tired in the morning.	Something that happens as a result of a particular action or set of conditions.
126	10	consequence	Heart problems are a(n) _____ of long term drug use.	Something that happens as a result of a particular action or set of conditions.
127	10	principal	My _____ reason for working at that particular store was the employee discount.	Most important; main.
128	10	principal	The _____ reason she went to college was to get a good job.	Most important; main.
129	10	principal	Any government's _____ aim should be the safety and wellbeing of its people.	Most important; main.
130	10	region	This type of coffee bean grows only in one specific _____ of Peru.	A large area of a country or of the world, usually without exact limits; an area.
131	10	region	The Gulf _____ of the United States often experiences hurricanes.	A large area of a country or of the world, usually without exact limits; an area.
132	10	region	The southern _____ of the desert was too dangerous to travel through.	A large area of a country or of the world, usually without exact limits; an area.

133	10	restrict	If my son does not behave, I will _____ his access to the Internet.	To limit or control the size, amount, or range of something.
134	10	restrict	Universities often _____ smoking on campus to encourage students not to smoke.	To limit or control the size, amount, or range of something.
135	10	restrict	The automobile safety law will _____ drivers from talking on their phones.	To limit or control the size, amount, or range of something.
136	11	considerable	Shutting down computers at night will result in _____ savings for the company.	Fairly large or influential, especially enough to have an effect or be important.
137	11	considerable	A(n) _____ number of voters are opposed to that idea.	Fairly large or influential, especially enough to have an effect or be important.
138	11	considerable	I like your plan, but it does have one _____ problem.	Fairly large or influential, especially enough to have an effect or be important.
139	11	illustrate	In order to _____ his points in the meeting, he brought charts and graphs.	To make the meaning of something clearer by giving examples.
140	11	illustrate	The revolution in Egypt helped to _____ the power of the Internet.	To make the meaning of something clearer by giving examples.
141	11	illustrate	Let me _____ the theory by telling you about a real-world example.	To make the meaning of something clearer by giving examples.
142	11	impact	The _____ of the decision was felt by all the employees in the company.	The effect or influence that an event or situation has on someone or something.
143	11	impact	The _____ of the flooding in Pakistan included destroyed homes and farms.	The effect or influence that an event or situation has on someone or something.
144	11	impact	The trauma had such a large _____ on the girl that she did not speak for years.	The effect or influence that an event or situation has on someone or something.
145	11	perceive	Using a different word can change how the listener will _____ what you're saying.	To understand or think of something or someone in a particular way.
146	11	perceive	Although he laughed at the joke, I could _____ that he was offended.	To understand or think of something or someone in a particular way.
147	11	perceive	Some people _____ global warming as a large danger, while others do not worry.	To understand or think of something or someone in a particular way.
148	11	regulate	During wars, governments sometimes _____ the amount of food people can buy.	To control an activity or process, especially by rules.
149	11	regulate	State laws _____ the sale of guns.	To control an activity or process, especially by rules.
150	11	regulate	Factories must be careful not to violate the laws which _____ air and water quality.	To control an activity or process, especially by rules.

APPENDIX F

PROJECT DESCRIPTION AND SCHEDULE

The following two pages show the project description and schedule provided to all participants.

Writing 4 Core Vocabulary Tutor Project Description

The purpose of this research study is to determine the effectiveness of a certain type of computer-assisted vocabulary practice. In other words, we want to discover how helpful a computer program is in helping you to study English vocabulary. We will ask all of you to use the vocabulary practice software two times per week for ten minutes as homework while you are a student in Writing 4. You may use any computer, at home or here on campus, to use this software, as long as you have an Internet connection. You will use the studying software as homework for eight weeks, starting next Monday, May 30th, 2011.

Today, we will ask you to share some basic information about yourself and to take a vocabulary pre-test. At the end of the semester, we will ask you for your thoughts and opinions about the vocabulary software, and we will ask you to complete a vocabulary post-test. This will all be done on computer.

As you use the software, the program will keep track of how quickly you answer questions and whether you answer them right or wrong; but because this is just for practice, you are not being graded for right or wrong answers. The only thing you are being graded on is whether you use the software to study.

There are no anticipated negative effects of participating in this study, and although we expect that the vocabulary practice software will be helpful and useful, we cannot guarantee that it will be effective. There is no payment for participating in this study.

For grading purposes, we will tell your teachers who has been using the software and who has not been using the software. However, your teachers will not see the answers you give to any questions we ask you. All information we ask from you will be kept private and confidential. The only person who will be allowed to see identifiable information from you is Bill Price. If we report any scientific results from this study, the results will be kept anonymous.

Although you must use the software as homework in the Writing 4 course, you can choose whether or not we are allowed to look at your information for research purposes. In other words, participation in the research is voluntary, and you may withdraw at any time by contacting Bill Price. If you choose to withdraw, you will still need to do the work as a student in the ELI, but we will not use your results for scientific research. Withdrawing from research will not affect your grade or your standing with the University.

This study is being conducted by Bill Price, who can be reached at wcp5@pitt.edu if you have any questions or wish to withdraw from the research.

https://sites.google.com/site/corevocabularytutor/

Schedule

Date	Description	Done?
Today	Questionnaire; Pre-test; Tutor session 1 (in class)	✓
Monday 5/30	Tutor session 2 (10 minutes) (on website)	
Thursday 6/2	Tutor session 3 (10 minutes) (on website)	
Monday 6/6	Tutor session 4 (10 minutes) (on website)	
Thursday 6/9	Tutor session 5 (10 minutes) (on website)	
Monday 6/13	Tutor session 6 (10 minutes) (on website)	
Thursday 6/16	Tutor session 7 (10 minutes) (on website)	
Monday 6/20	Tutor session 8 (10 minutes) (on website)	
Thursday 6/23	Tutor session 9 (10 minutes) (on website)	
Monday 6/27	Tutor session 10 (10 minutes) (on website)	
Thursday 6/30	Tutor session 11 (10 minutes) (on website)	
Monday 7/4	Tutor session 12 (10 minutes) (on website)	
Thursday 7/7	Tutor session 13 (10 minutes) (on website)	
Monday 7/11	Tutor session 14 (10 minutes) (on website)	
Thursday 7/14	Tutor session 15 (10 minutes) (on website)	
Monday 7/18	Tutor session 16 (10 minutes) (on website)	
Thursday 7/21	Questionnaire; Post-test (on website)	

https://sites.google.com/site/corevocabularytutor/

APPENDIX G

BACKGROUND SURVEY

The following pages reproduce the background survey.

Pre Vocabulary Self-Evaluation Questionnaire (Summer 2011)

Your responses will be used only for research purposes. They will not be shared with others.

Please enter your Pitt ID *
If your email address is abc123@pitt.edu, enter abc123

Please select your course section. *
Writing 4M

Page 2 After page 1 Continue to next page

Personal Background

What is your native language? *
- Arabic
- Chinese
- Japanese
- Korean
- Other:

How many years have you studied English? *
1

How long have you lived in an English-speaking country? *
e.g. USA, United Kingdom, Canada, Australia,...
Fewer than 3 months

How old are you? *
For example: 25

Page 3 After page 2 Continue to next page

History in the ELI

What ELI courses are you taking this semester? *
Choose all that apply.

- [ ] Speaking 3
- [ ] Listening 3
- [ ] Grammar 3
- [ ] Reading 3
- [ ] Writing 3
- [ ] Speaking 4
- [ ] Listening 4
- [ ] Grammar 4
- [ ] Reading 4
- [ ] Writing 4
- [ ] Speaking 5
- [ ] Listening 5
- [ ] Grammar 5
- [ ] Reading 5
- [ ] Writing 5

IN THE PAST, have you taken any Level 4 or Level 5 classes at the ELI? *
For example, if you took Speaking 4 in Spring 2011, or Listening 5 in Fall 2010, choose "Yes"

- ( ) Yes
- ( ) No

Page 4 After page 3 Continue to next page

Note: "Go to page" selections will override this navigation. Learn more

Past ELI courses

Please type all Level 4 or Level 5 courses you took in the past in the ELI. *
Example: Fall 2010 - Listening 4, Spring 2011 - Listening 5 and Speaking 4

Page 5 After page 4 Continue to next page

Thoughts about Vocabulary

Approximately how much English vocabulary do you believe you know? *

	1	2	3	4	5	6	
I know very little English vocabulary	( )	( )	( )	( )	( )	( )	I know a lot of English vocabulary

Which do you enjoy more, speaking English or writing in English? *

	1	2	3	4	5	6	
I enjoy speaking English more	○	○	○	○	○	○	I enjoy writing English more

Which do you enjoy more, listening to English or reading English? *

	1	2	3	4	5	6	
I enjoy listening to English more	○	○	○	○	○	○	I enjoy reading English more

What strategies do you like to use to learn new vocabulary words?

Choose as many options as you want.

- ☐ I like to listen to how other people use the new word.
- ☐ I like to read the new word in a book or newspaper.
- ☐ I like to look up the new word in the dictionary.
- ☐ I like to try to use the new word in my own speaking.
- ☐ I like to try to use the new word in my own writing.
- ☐ I like to write down new words so that I can study them later.
- ☐ I like to use flash cards (or other studying methods) to study the new word.
- ☐ Other:

APPENDIX H

PRE-TEST AND POST-TEST

The following pages reproduce the pre-test and post-test.

Vocabulary Knowledge Assessment (Part 1 of 10)

Please check off the words that you already know and could use in your own writing.

- ☐ achieve
- ☐ acknowledge
- ☐ aspect
- ☐ clarify
- ☐ assess
- ☐ discriminate
- ☐ concept
- ☐ eliminate
- ☐ cooperate
- ☐ reluctant

Vocabulary Knowledge Assessment (Part 2 of 10)

Please check off the words that you already know and could use in your own writing.

- ☐ acquire
- ☐ abstract
- ☐ initial
- ☐ crucial
- ☐ obvious
- ☐ inherent
- ☐ relevant
- ☐ rational
- ☐ target
- ☐ stable

Vocabulary Knowledge Assessment (Part 3 of 10)

Please check off the words that you already know and could use in your own writing.

- ☐ approximate
- ☐ ambiguous
- ☐ demonstrate
- ☐ coordinate
- ☐ evaluate
- ☐ generate
- ☐ modify
- ☐ parallel
- ☐ previous
- ☐ trigger

Vocabulary Knowledge Assessment (Part 4 of 10)

Please check off the words that you already know and could use in your own writing.

- ☐ factor
- ☐ access
- ☐ imply
- ☐ anticipate
- ☐ method
- ☐ enhance
- ☐ resource
- ☐ reveal
- ☐ valid
- ☐ underlie

Vocabulary Knowledge Assessment (Part 5 of 10)

Please check off the words that you already know and could use in your own writing.

- ☐ affect
- ☐ attribute
- ☐ category
- ☐ bias
- ☐ distinct
- ☐ exhibit
- ☐ evident
- ☐ philosophy
- ☐ perspective
- ☐ violate

Vocabulary Knowledge Assessment (Part 6 of 10)

Please check off the words that you already know and could use in your own writing.

- ☐ establish
- ☐ adapt
- ☐ feature
- ☐ emerge
- ☐ potential
- ☐ framework
- ☐ range
- ☐ justify
- ☐ secure
- ☐ trend

Vocabulary Knowledge Assessment (Part 7 of 10)

Please check off the words that you already know and could use in your own writing.

- ☐ complex
- ☐ compensate
- ☐ constant
- ☐ diverse
- ☐ distribute
- ☐ impose
- ☐ equate
- ☐ sustain
- ☐ indicate
- ☐ unify

Vocabulary Knowledge Assessment (Part 8 of 10)

Please check off the words that you already know and could use in your own writing.

- [ ] alternative
- [ ] arbitrary
- [ ] correspond
- [ ] fluctuate
- [ ] exclude
- [ ] inevitable
- [ ] incorporate
- [ ] manipulate
- [ ] rely on
- [ ] transform

Vocabulary Knowledge Assessment (Part 9 of 10)

Please check off the words that you already know and could use in your own writing.

- [ ] compatible
- [ ] implement
- [ ] consequence
- [ ] integrate
- [ ] principal
- [ ] maximize
- [ ] region
- [ ] pursue
- [ ] restrict
- [ ] reinforce

Vocabulary Knowledge Assessment (Part 10 of 10)

Please check off the words that you already know and could use in your own writing.

- [ ] considerable
- [ ] implicit
- [ ] illustrate
- [ ] negate
- [ ] impact
- [ ] persist
- [ ] perceive
- [ ] proportion
- [ ] regulate
- [ ] retain

APPENDIX I

OPINION SURVEY

The following pages reproduce the opinion survey.

Post Vocabulary Self-Evaluation Questionnaire (Summer 2011)

Your responses will be used only for research purposes. They will not be shared with others.

Please enter your Pitt ID *
If your email address is abc123@pitt.edu, enter abc123

Please select your course section. *
Writing 4M

Page 2 After page 1 Continue to next page

Thoughts about the Vocabulary Tutor (Part 1 of 4)

I thought that the vocabulary tutor program was... *

	1	2	3	4	5	6	
Too hard	○	○	○	○	○	○	Too easy

I thought that the vocabulary tutor program was... *

	1	2	3	4	5	6	
Not useful	○	○	○	○	○	○	Very useful

I thought that the vocabulary tutor program was... *

	1	2	3	4	5	6	
Not enjoyable to use	○	○	○	○	○	○	Very enjoyable to use

Page 3 After page 2 Continue to next page

Thoughts about the Vocabulary Tutor (Part 2 of 4)

I believe that using the vocabulary tutor helped me to learn a lot of vocabulary. *

	1	2	3	4	5	6	
Disagree	○	○	○	○	○	○	Agree

I believe that using the vocabulary tutor helped me to use new vocabulary more easily. *

1 2 3 4 5 6

Disagree ○ ○ ○ ○ ○ ○ Agree

I believe that using the vocabulary tutor helped me to study more quickly than I used to. *

1 2 3 4 5 6

Disagree ○ ○ ○ ○ ○ ○ Agree

Page 4 After page 3 Continue to next page

Thoughts about the Vocabulary Tutor (Part 3 of 4)

In the vocabulary tutor program, did you prefer to study vocabulary WITH hints or WITHOUT hints? *

1 2 3 4 5 6

With hints ○ ○ ○ ○ ○ ○ Without hints

Page 5 After page 4 Continue to next page

Thoughts about the Vocabulary Tutor (Part 4 of 4)

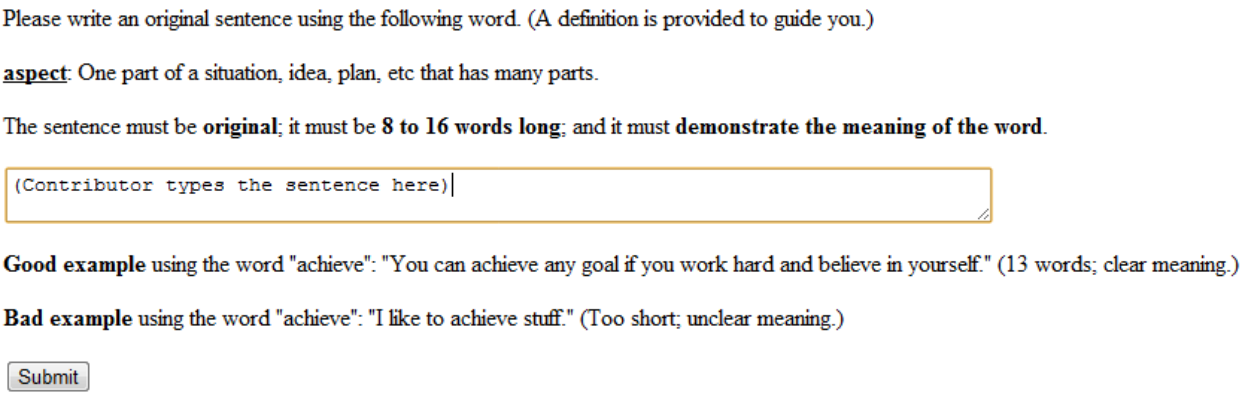
Do you have any comments or suggestions regarding the vocabulary tutor program? For example, what was the best part of it? What was the worst part of it? What would you change?

Your opinion will help us to make the program better for students in the future.

APPENDIX J

SENTENCE WRITING TASK PUBLISHED TO MECHANICAL TURK

The following image shows an exemplar of the sentence writing task published to Amazon.com's Mechanical Turk website to help generate cloze exercise sentences.



BIBLIOGRAPHY

Alderson, J. C. (1979). The Cloze Procedure and Proficiency in English as a Foreign Language. *TESOL Quarterly*, *13*(2), 219-227.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Lawrence Erlbaum Associates. Retrieved from http://books.google.com/books?hl=en&lr=&id=rdxRk3J8z-gC&oi=fnd&pg=PR7&dq=atomic+components+of+thought&ots=CxUmsLQjH2&sig=2nEDU30Z4JvwLW4-UdPB110qZyc

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

Bahrick, H. P., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language*, *52*(4), 566-577. doi:10.1016/j.jml.2005.01.012

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

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological bulletin*, *132*(3), 354-80. doi:10.1037/0033-2909.132.3.354

Cobb, T. (n.d.). Vocabulary Profiler (BNC). Retrieved October 23, 2011, from http://www.lextutor.ca/vp/bnc/

Coxhead, A. (2000). A new academic word list. *TESOL quarterly*, *34*(2), 213-238. Retrieved from http://www.ingentaconnect.com/content/tesol/tq/2000/00000034/00000002/art00002

Cull, W. L., Shaughnessy, J. J., & Zechmeister, E. B. (1996). Expanding understanding of the expanding-pattern-of-retrieval mnemonic: Toward confidence in applicability. *Journal of Experimental Psychology: Applied*, *2*(4), 365-378. doi:10.1037/1076-898X.2.4.365

Dempster, F. N. (1988). The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist*, *43*(8), 627-634. Retrieved from http://psycnet.apa.org/journals/amp/43/8/627/

Ebbinghaus, H. (1885). Uber das Gedächtnis. Untersuchungen zur experimentellen Psychologie. Leipzig: Duncker & Humblot.

Ebbinghaus, H. (1964). Memory: a contribution to experimental psychology. New York: Dover Publications. Retrieved from http://books.google.com/books?id=xT0mAQAAIAAJ

English Language Institute Vocabulary List. (2007). *Language*. English Language Institute, University of Pittsburgh.

Folse, K. S. (2006). The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly*, *40*(2), 273-293. Retrieved from http://www.jstor.org/stable/40264523

Jongsma, E. (1971). *The Cloze Procedure as a Teaching Technique*. Newark, Delaware: The International Reading Association. Retrieved from http://eric.ed.gov/ERICWebPortal/recordDetail?accno=ED055253

Juffs, A., Friedline, B., Wilson, L., Eskenazi, M., & Heilman, M. (n.d.). Activity theory and computer-assisted learning of English vocabulary.

Karpicke, J. D., & Roediger, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(4), 704-719. Retrieved from http://psycnet.apa.org/journals/xlm/33/4/704/

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966-8. doi:10.1126/science.1152408

Koedinger, K. R., & Aleven, V. (2007). Exploring the Assistance Dilemma in Experiments with Cognitive Tutors. *Educational Psychology Review*, *19*(3), 239-264. doi:10.1007/s10648-007-9049-0

Koedinger, K. R., Pavlik, P. I., McLaren, B. M., & Aleven, V. (2008). Is it Better to Give than to Receive? The Assistance Dilemma as a Fundamental Unsolved Problem in the Cognitive Science of Learning and Instruction. *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 2155-2160).

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

Martin, K. I. (2011). *Reading in English: a comparison of native Arabic and native English speakers*. University of Pittsburgh.

Metzler-Baddeley, C., & Baddeley, R. J. (2009). Does adaptive training work? *Applied Cognitive Psychology*, *23*(2), 254-266. doi:10.1002/acp.1454

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

Oller, J. W., & Conrad, C. A. (1971). The cloze technique and ESL proficiency. *Language Learning, 21*(2), 183-195.

Pavlik, P. I. (2005). *The microeconomics of learning: Optimizing paired-associate memory*. Carnegie Mellon University. Retrieved from http://gradworks.umi.com/31/91/3191608.html

Pavlik, P. I., & Anderson, J. R. (2005). Practice and Forgetting Effects on Vocabulary Memory: An Activation-Based Model of the Spacing Effect. *Cognitive Science, 29*(4), 559-586. doi:10.1207/s15516709cog0000_14

Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of experimental psychology. Applied, 14*(2), 101-17. doi:10.1037/1076-898X.14.2.101

Pavlik, P. I., Presson, N., Dozzi, G., Wu, S.-mei, MacWhinney, B., & Koedinger, K. R. (2007). The FaCT (Fact and Concept Training) System: A new tool linking cognitive science with educators. *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society* (pp. 397–402). Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.3595&rep=rep1&type=pdf

Pimsleur, P. (1967). A memory schedule. *The Modern Language Journal*, *51*(2), 73–75. JSTOR. Retrieved from http://www.jstor.org/stable/321812

Rea, C. P., & Modigliani, V. (1985). The effect of expanded versus massed practice on the retention of multiplication facts and spelling lists. *Human Learning, 4*(1), 11-18. Retrieved from http://psycnet.apa.org/psycinfo/1986-07610-001

Rye, J. (1982). *Close Procedure in the Teaching of Reading*. Portsmouth, New Hampshire: Heinemann Educational Books Inc. Retrieved from http://journals.sfu.ca/tesl/index.php/tesl/article/view/544/375

Skinner, B. F. (1968). *The technology of teaching*. East Norwalk, CT: Appleton-Century-Crofts.

Szofer, A. (2010). *Sex and age effects on computer assisted vocabulary and grammar acquisition in adolescence*. Adam Mickiewicz University, Poznan, Poland.

Taylor, W. W. (1953). “Cloze procedure”: a new tool for measuring readability. *Journalism Quarterly, 30*, 415-433.

Taylor, W. W. (1956). Recent developments in the use of “cloze procedure.” *Journalism Quarterly, 33*, 42-48, 99.

The British National Corpus, version 3. (2007). . Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Retrieved from http://www.natcorp.ox.ac.uk/

Thompson-Panos, K., & Thomas-Ružić, M. (1983). The least you should know about Arabic: Implications for the ESL writing instructor. *TESOL Quarterly*, *17*, 609-623.

Tulving, E. (1991). Concepts of human memory. In L. Squire, G. Lynch, N. M. Weinberger, & J. L. McGaugh (Eds.), *Memory: Organization and Locus of Change* (pp. 3-32). Oxford University Press.

Vygotsky, L. S. (1978). *Mind in society: the development of higher psychological processes*. (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.). Cambridge, MA: Harvard University Press. Retrieved from http://books.google.com/books?id=RxjjUefze_oC&pgis=1

Wolf, G. (2008, May). Want to Remember Everything You'll Ever Learn? Surrender to This Algorithm. *Wired Magazine*. Retrieved from http://www.wired.com/medtech/health/magazine/16-05/ff_wozniak?currentPage=all

Wood, D. J., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychiatry and Psychology*, *17*(2), 89-100.

Wozniak, P. A. (1990). *Optimization of Learning*. University of Technology in Poznan.