International Journal of Artificial Intelligence in Education 20 (2010) 73-98
DOI 10.3233/JAI-2010-0003
IOS Press

Personalization of Reading Passages Improves Vocabulary Acquisition

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, Maxine Eskenazi
Language Technologies Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA
mheilman@cs.cmu.edu, kct@cs.cmu.edu, callan@cs.cmu.edu, max@cs.cmu.edu

Alan Juffs, Lois Wilson, *English Language Institute, Department of Linguistics, University of Pittsburgh, Pittsburgh, PA 15213, USA*
juffs@pitt.edu, liw@pitt.edu

Abstract. The REAP tutoring system provides individualized and adaptive English as a Second Language vocabulary practice. REAP can automatically personalize instruction by providing practice readings about topics that match interests as well as domain-based, cognitive objectives. While most previous research on motivation in intelligent tutoring environments has focused on increasing extrinsic motivation, this study focused on increasing personal interest. Students were randomly split into control and treatment groups. The control-condition tutor chose texts to maximize domain-based goals such as the density of practice opportunities for target words. The treatment-condition tutor also preferred texts that matched personal interests. The results show positive effects of personalization, and also demonstrate the importance of negotiating between motivational and domain-based goals.

Keywords. Motivation, Personal Interest, Intelligent Tutoring System, English as a Second Language

INTRODUCTION

While tutoring systems for vocabulary practice can be successfully deployed in language courses (Heilman, Collins-Thompson, Callan, and Eskenazi, 2006), if designers ignore motivational factors then students may become unmotivated or discouraged.

Motivation can be defined as the desire to engage in a specific activity (Shiefele, 1999). It interacts with perceived self-efficacy, which is a student's belief that he or she can succeed at a specific activity (Bandura, 1997), to affect students' behavior. Motivation can be separated into *intrinsic* and *extrinsic* forms (Deci and Ryan, 1985). In terms of education, extrinsic motivation depends on outside forces such as praise from teachers or the fear of receiving poor grades, while intrinsic motivation is the desire to learn because the task or content is enjoyable, satisfying, or fun.

Extrinsic motivation and intrinsic motivation can have different effects on learning. Lepper (1988) discusses the differences between these forms of motivation. Extrinsically motivated students often choose the easiest path to achieving an extrinsic goal. They are also more likely to quit after an initial failure if they perceive a task to be difficult. In contrast, intrinsically motivated students are

ISSN 1560-4292/10/$27.50

more likely to take risks, choose difficult learning paths, persist in the face of difficulty, and apply effective learning strategies.

One of the important precursors to intrinsic motivation is interest. Recent literature divides interest into two forms: personal interest and situational interest (Schraw and Lehman, 2001). Personal interest, also referred to as individual interest or topic interest, is topic-specific and has long-lasting personal value. It is based on pre-existing knowledge, experiences, and emotions. For example, a person might be motivated to read an otherwise dry piece of text because it discusses a topic of personal interest (e.g., financial news).

In contrast, situational interest is context-specific, of short-term value, and is triggered by the environment rather than by the self. Situational interest arises due to various factors including features of the task, specific prior knowledge, and features of the text or content such as vividness, seductiveness, and coherence. For example, a student might read a book because it is well-written and engaging even though the topic is not particularly personally interesting (e.g., a mystery novel).

Both forms of interest can play a role in motivating students. A recent study by Brantmeier (2006) found that among advanced students of Spanish as a Second Language, self-reported measures of both personal interest and situational interest were related to reading comprehension as measured by sentence completion and multiple choice items. It is also important to note that other divisions of intrinsic motivation and interest exist in the literature (e.g., Malone and Lepper, 1987).

This paper describes personalization in the REAP tutoring system, which provides individualized and adaptive vocabulary practice for English as a Second Language vocabulary. REAP can automatically personalize instruction by providing practice readings about topics that match personal interests as well as domain-based, cognitive objectives. An experimental study compared a group receiving personalized readings to a control group whose readings were chosen based solely on domain-based goals such as the density of practice opportunities for target words. The results show positive effects of personalization, and also demonstrate the importance of negotiating between motivational and domain-based goals.

Motivational Issues in Tutoring Systems

Situational Interest, Perceived Self-Efficacy, and Extrinsic Motivation

Many instructional systems attempt to motivate students by focusing on either situational interest, extrinsic motivation, or self-efficacy. These tutoring systems motivate by providing positive and negative feedback, making content vivid and coherent, or changing the features of a task. They increase perceived self-efficacy by encouraging students, or by allowing them to succeed by providing easier problems and extra assistance.

In order to increase situational interest, computer-based instructional systems aim to organize and clearly present content information to students. They focus on improving the coherence and clarity of information. While coherence and clarity are strongly related to cognitive factors such as working memory load, they also affect situational interest. For example, in a review of interest in relation to reading, Hidi (2001) reported higher interest levels for students reading well organized texts.

Attractive multimedia environments can also increase situational interest. However, Clark and Mayer (2003) caution against adding irrelevant information such as background music that may distract learners. Such extraneous information is often labeled as “seductive details,” and in some studies has had negative effects on learning even while interest increased (Harp and Mayer, 1998).

Another widely used tool for engaging and motivating students is the pedagogical agent (e.g., Chan 1996; Johnson, Rickel, and Lester, 2000). Agents facilitate rich, natural, face-to-face interactions with students. They may increase situational interest by facilitating more natural social interactions compared to other types of user interfaces. They may also increase perceived self-efficacy by providing encouragement and praise. They may even invoke social norms to provide some level of extrinsic motivation. Natural social interactions can also be supported in other ways. For example, in a study of a tutoring system for chemistry, McLaren, Yaron, Lin, and Koedinger (2007) compared hints and directions that were written in a formal tone to those that were written in a more polite and conversational manner aimed at increasing engagement. However, they did not find significant learning gains.

Educational games and game-like features can also increase student motivation. Randel, Morris, Wetzel, and Whitehill (1992) provide a review of research on educational gaming. However, the circumstances under which games improve learning are not clearly understood in the currently available literature, and further research is warranted. For instance, while games can be both engaging and effective, they do not lead to improved learning in all domains (Klawe, 1998).

Instructional systems may have to monitor and adapt to the student's current state in order to provide effective motivational support. For instance, a system might provide easier problems in order to increase self-efficacy beliefs if it detected low student confidence. Conversely, a system might give more difficult problems in order to provide challenges and maintain interest if it detected high student confidence. However, motivational planning must be balanced by domain-based planning because the ultimate goal is not to please the student but to make progress through the curriculum. Del Soldato and du Boulay (1995) provide a detailed discussion of the interaction of domain-based goals and motivational goals. They describe a rule-based system for choosing the level of difficulty of problems, provision of assistance, use of praise and other strategies for affecting self-efficacy and motivation based on student performance and estimates of student motivational states. A preliminary evaluation of that system, however, led to inconclusive results.

Personal Interest

Matching personal interests can also be an effective method for motivating students in an instructional system. Human teachers and tutors often attempt to connect classroom material to students' personal interests (Fives and Manning, 2005). However, just as it is difficult for teachers to tailor a curriculum to match each student's skills or conceptual knowledge, it is also difficult to tailor a curriculum to match each student's personal interests. Teachers have neither sufficient time nor sufficient resources in most cases.

Intelligent tutoring systems can, however, provide instruction that is tailored to specific students in various ways. Many tutoring systems adapt to individual students' skill levels and knowledge, a prominent example being the Cognitive Tutor for Algebra (e.g., Koedinger, Anderson, Hadley, and Mark, 1997). Tutoring systems that personalize instruction to match personal interests, however, are less common—one notable example is described by Cordova and Lepper (1996).

In the work described in this paper, functionality for matching instructional materials to personal interests was added to an existing tutoring system called REAP[1] (Brown and Eskenazi, 2004). In this

[1] REAP is not an acronym.

paper, this act of providing instructional materials that match the personal interests of the student is referred to as *personalization*.

Prior work on tutoring systems has involved personal interest, but has often combined it with choice. For a broad review of the provision of choices and learner control in instructional systems, see (Kay, 2001). Given a choice of activities, students will often choose ones that are personally interesting. As such, it can be hard in some situations to distinguish the effects of choice and personal interest. For example, Beck (2007) reported improvements in learning outcomes in a reading tutor when children were given a choice of practice reading passages based on their titles. However, it is unclear from that study whether the improvements were due to choice or to the fact that students chose texts that were more interesting or otherwise better practice.

Cordova and Lepper (1996) reported positive effects of both personalization and choice within an educational game for children in the domain of arithmetic. Those studies found that both personalization and choice played important roles: students given a choice of personalized tasks outperformed students given tasks without choice and/or without personalization.

On the other hand, Flowerday, Schraw, and Stevens (2004) reported no statistically reliable associations with learning from text for either choice or personal interest in a lab study on reading engagement, attitude, and learning. All participants in that study read the same passages regardless of their specified personal interests. A tutoring system like REAP, however, can automatically match topics of readings to personal interests.

Even so, it is not clear that effective personalization will improve learning. As mentioned above, Clark and Mayer (2003) found that adding extraneous information to increase situational interest may negatively affect learning by distracting students. The negative effects of extraneous information are also important to consider in relation to matching personal interests. A system might introduce extraneous details when matching personal interests. For example, personalizing a math problem by adding an extraneous reference to a sports star might distract students. The possibility of introducing extraneous details may be less relevant, however, when a tutoring system selects from pre-existing materials, as REAP does, rather than manipulating existing materials or filling in templates.

Thus, while previous work has found correlations between personal interest and either learning (e.g., Hidi, 2001) or comprehension (Brantmeier, 2006), it is unclear whether matching instructional materials to personal interests will cause better learning in an interactive learning environment. This paper addresses the following two general questions. First, does selecting instructional materials to match students' personal interests lead to better learning? Second, is it possible to automate the selection of personalized materials in a tutoring system?

THE REAP TUTORING SYSTEM

Student Interactions with REAP

The REAP tutor is an intelligent tutoring system for English as a Second Language vocabulary and reading practice (Brown and Eskenazi, 2004). It provides contextualized practice on individualized vocabulary lists by selecting reading passages from a large corpus of annotated Web documents.

Students begin working with the REAP tutor by performing a series of self-assessments about target vocabulary words. These words come from a list of possibly hundreds of target vocabulary words that REAP aims to teach students. REAP is flexible with regard to the target vocabulary list, and can use any list chosen by instructors or researchers. For each word in the target vocabulary, the student responds to the question, "Do you know the word 'X'?" by clicking "yes" or "no." This preliminary measure of vocabulary knowledge is simple but effective. The goal of this interaction is to provide estimates of knowledge of the many words that might be practiced. These knowledge estimates allow the REAP tutor to individualize instruction according to the student's needs. While these self-assessments are simple, they are also very fast compared to other types of questions. In a pilot study conducted during the Summer of 2006, students spent approximately 6 seconds on average per self-assessment question compared to 40 seconds on average on multiple-choice cloze (i.e., fill-in-the-blank) questions.

Self-assessments also provide fairly reliable estimates of word knowledge. In fact, comparisons of pre-test self-assessments to cloze and synonym question performance indicate that negative self-assessments are very reliable. That is, students who claim to not know a word almost always respond incorrectly to other types of vocabulary questions for that word. When a student does claim to know a word, this information is not as reliable. The form of self-assessments used in REAP is similar to the "Yes-No Test" used by second language teachers and researchers (Meara, 1992).

Once the student completes all of the self-assessments, which usually takes about 15 minutes, he or she works through a series of practice readings gathered from the Web that contain target vocabulary words. The REAP tutor selects readings that contain target vocabulary words that the student does not know, as estimated from self-assessments and performance on prior training tasks. It ranks texts from a large corpus of Web documents by considering various pedagogical factors such as the number of unknown target words, the reading grade level, and text length (Brown and Eskenazi, 2004). REAP then provides students with a choice of the four highest-ranked readings. It includes short excerpts of about fifty words to help the student choose a reading.

The REAP tutor then presents the practice reading that was selected by the student through an interface that is illustrated in Figure 1. The reading passages are Web pages shown in original form with all links disabled. They are approximately 1000 words long, which corresponds to about two pages of text. The primary reason for using the Web as a corpus of authentic materials is to have access to a very large source of practice texts. By using the Web, the tasks of authoring and editing texts are not necessary.

Each practice reading contains one to ten target words, depending on the student and the curriculum. These target words are highlighted in order to draw attention to them. Knight (1994) showed positive results for both incidental vocabulary acquisition and reading comprehension when target words were marked to focus student attention on them. Participants were native English speakers learning Spanish as a Second Language. The effects of highlighting are not entirely clear,

however: for example, a study by de Ridder (2002) did not replicate the positive effects of highlighting.

While reading, students can access dictionary definitions for any word. Target words are hyperlinked, and any other word can be looked up by entering the word in a text box at the bottom of the screen. In recent versions, REAP can also allow students to click on any word in a text, not just target words. The definitions used in the studies described in this paper came from the Cambridge Advanced Learner's Dictionary[2] , a dictionary with simplified definitions suitable for intermediate and advanced ESL students. REAP also logs all dictionary use.

The REAP tutor typically places no restrictions on the time spent on each reading. In practice, students spend anywhere from 5-30 minutes depending on the student and the length of the text. In a typical training session lasting 40 minutes, each student works through 2-3 readings.

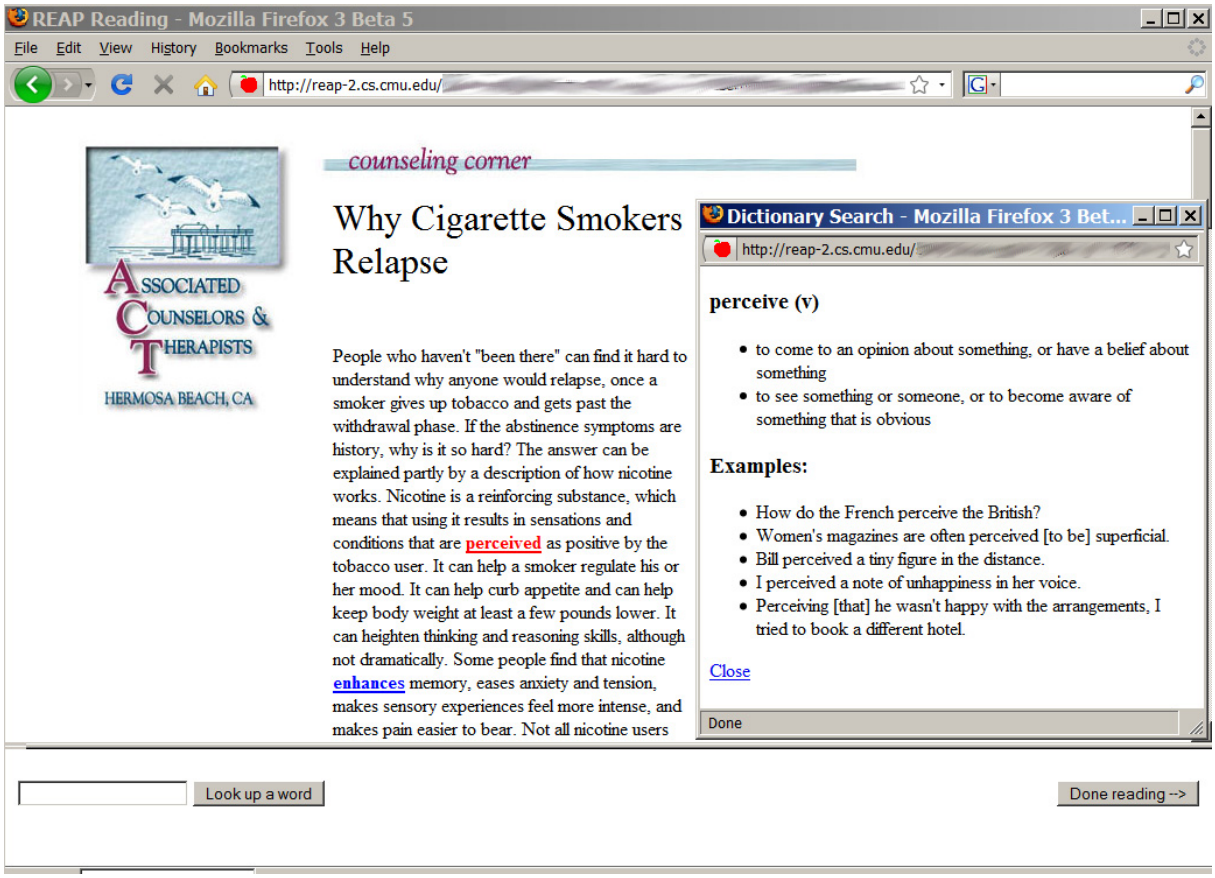


Fig. 1. A screenshot of REAP's interface for practice readings, showing a text classified into the Health category.

[2] http://dictionary.cambridge.org/

After each reading, the REAP tutor presents practice exercises for the target words. Cloze questions are the primary type of exercise in the REAP tutor. A cloze question is created from a sentence that uses the target word in an informative context. The target word is removed and replaced with a blank. The student fills in that blank by producing the appropriate word or by selecting it from a set of choices with distracters. Similar questions are commonly used in English as a Second Language courses, such as in the Vocabulary Levels Test developed by Laufer and Nation (1999). Currently, REAP uses a multiple-choice cloze format rather than a free-response format so that student responses can be graded automatically.

For the studies described in this paper, teachers manually authored the sentences used in the cloze questions. The researchers also checked these questions for quality control. Ideally, question generation would be automated, but recent research on generating cloze questions shows that a significant fraction of computer-generated questions are judged unusable by instructors (Liu, Wang, Gao, and Huang, 2005). For each cloze question, the REAP tutor randomly chose distracters from the set of target words that either appeared in the reading or had the same part of speech. In the Fall 2006 study described below, nine distracters and the correct answer were given as choices for each cloze question. After each cloze question, the REAP tutor provided immediate feedback by providing the correct answer. It also allowed the student to look up dictionary definitions while receiving feedback.

After the practice exercises, REAP asks a "reading-check" question to measure whether or not the student actually read the preceding passage. These questions are a simple measure of reader engagement and attention. The reading check questions are automatically generated for all the readings in REAP's database of practice readings. The tutor asks the student to select a set of words that appeared in the passage. The multiple-choice format provides four choices, each of which contains six words. The correct choice has six words which all appeared in the text, while the other choices contain some words which did not appear. The words from the text are automatically chosen to be salient and important words in that text, as measured by their frequency in the text normalized by frequency in general English. This method chooses rarer words such as "motivation" rather than more common words such as "word."

After the reading-check question, students rate the just-finished passages according to their interest and the perceived difficulty of the text. A scale from 1 to 5 is used, where 5 indicates the highest interest or difficulty. The student can also provide comments about the previous reading.

Following the interest and difficulty questionnaire, the student proceeds to another reading. REAP selects this reading based on an updated student model that takes into account which words the student has already practiced. In the Fall 2006 study described in this paper, the tutor favored target words which had been practiced fewer times when selecting subsequent reading passages. More recent versions of the REAP Tutor employ knowledge tracing (Corbett & Anderson, 1994) to update estimates of the student's knowledge of each word.

Selecting from a Large Corpus to Individualize Practice

As mentioned above, the REAP Tutor selects practice reading passages from a large corpus of Web texts. This corpus is created automatically prior to the beginning of a semester or a study. The first step in creating the corpus is the specification by teachers or researchers of a list of target words. The

system sends queries with subsets of the words on this list to a commercial search engine, Altavista,[3] through a special interface for researchers and developers. The search engine returns a list of Web pages that contain those target words. REAP then downloads, filters, and annotates these pages. REAP then uses filters to remove pages that are not suitable for reading practice due to a variety of reasons including profanity, discussion of inappropriate topics, and length beyond predefined thresholds (e.g., 100-2000 words). REAP uses another filter to ensure high "text quality," which is estimated by the proportion of text that is content rather than links, navigation menus, and other noisy information that is not useful reading material but commonly appears in Web pages. The text-quality filter is particularly important since many Web pages are marketplace sites aimed at selling products or navigation hubs that just contain lists of links. REAP is conservative in choosing passages for its corpus, and often retains less than 1% of the pages retrieved from the search engine. For instance, the corpus used in the study described below contained over 25,000 texts. REAP then annotates the readings in its corpus with reading grade level predictions (Collins-Thompson and Callan, 2005), and finally creates an inverted index using the Lemur Toolkit[4] for Language Modeling and Information Retrieval so that reading passages can be retrieved quickly by the tutor.

The REAP tutor creates this large corpus of readings in order to provide individualized practice. With or without personalization by topics, the tutor still individualizes instruction by selecting texts with the target words that a particular student does not know. Each student potentially knows a different subset of the target words in a specific study or course, and thus each student will usually see a unique set of texts.

The tutor chooses passages that have combinations of multiple unknown target words because presenting long passages with single target words would require too much time. However, finding combinations of words can be complicated by restrictions on reading difficulty level and other factors discussed in the next section. It might be possible to present target words one at a time in shorter texts, but high-quality short texts are typically hard to find on the Web. Using excerpts of long texts also presents a challenge because of the potential loss of important contextual information and authenticity. If presenting texts with combinations of words were not necessary, then a small set of pre-selected readings for each target word would suffice.

Factors for Selection of Practice Texts

When selecting the next set of readings to offer the student, the tutor calculates a real number from zero to one for the pedagogical value of each reading passage in its corpus. To calculate a reading's overall value, the tutor combines values for various factors including document length, reading difficulty level, the number of target words appearing in the reading, and the degree to which the topic of the reading matches personal interests. As a simple example, if the tutor assumes that five is an optimal number of target words per reading, then a reading with five words might be assigned a value of 1.0 and a reading with four or six words might be assigned a lower value such as 0.8, and so forth. The actual method for producing scores given optimal values is slightly more complex. Similar scores are calculated for document length and reading difficulty level based on optimal values for those factors. The tutor then calculates a weighted average of the values for all of the factors to produce a

[3] Other search engines could certainly be used, but Altavista (http://www.altavista.com/) provides a particularly useful interface for these purposes.
[4] http://www.lemurproject.org

final score for a given text. The tutor then sorts all texts by their scores to find the estimated best texts for the student at the current time.

Choosing an optimal reading is like finding an optimal point in a multidimensional space. Each pedagogical factor is a dimension in this space. Some readings may be optimal along one dimension but not others. For instance, a reading may have a good density of target words, but be too difficult for a student given his or her reading skill level. Potential interest is another dimension in this space. Unless potential interest is perfectly uncorrelated to the other pedagogical factors, then giving weight to personalization will necessarily affect those factors.

In practice, the truly optimal practice reading at a given time for a given student is likely never found by the tutor. There are a variety of reasons. First, even relatively common words (e.g., "television") are fairly rare in text, and the tutor is searching for texts containing multiple rare words. Second, the tutor searches for readings within a narrow range of reading difficulty that is appropriate to the overall reading skill of the student. Third, measures for some of the factors are imprecise. For instance, a reading difficulty measure may have a standard error of one or two grade levels. Other potentially important factors not currently measured by the REAP tutor are the correctness of grammatical structures and the value of the contexts around target words at allowing students to infer meaning.

Since the tutor has such difficulty optimizing for existing pedagogical concerns, adding another factor such as personalization can have substantial effects on reading selection. For example, if the weight for personalization is raised, then the weight for the number of target words in the reading task becomes proportionally lower. In this case, the tutor may select readings that are very likely to be interesting but contain fewer target words. Of course, the weight for personalization can be decreased, but then students may never see readings which match their interests. Tradeoffs between motivation and domain factors similar to those discussed by del Soldato and du Boulay (1995) should be dealt with in REAP.

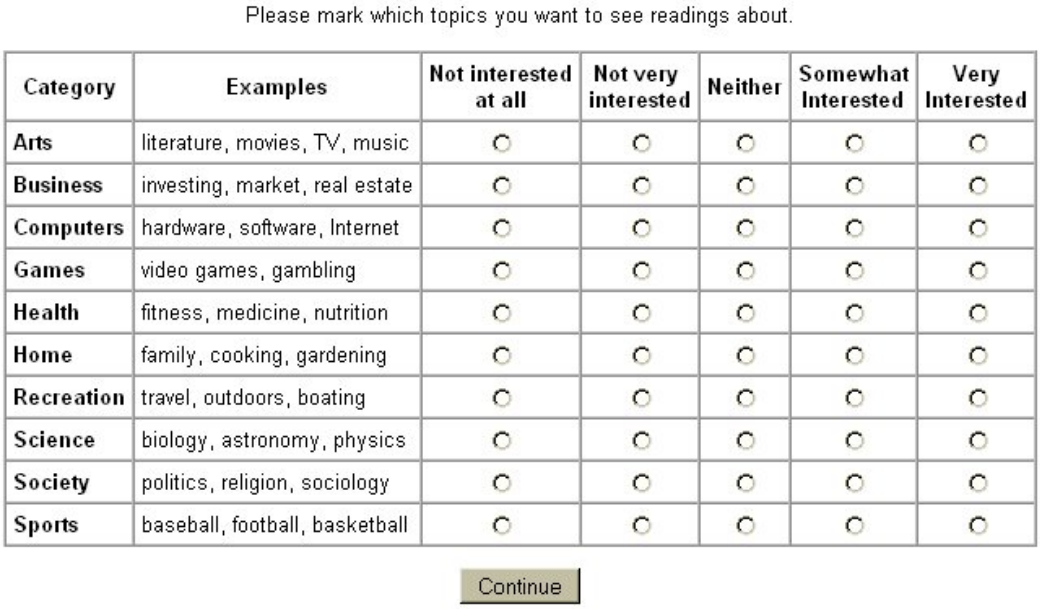
In the REAP tutor, the potential impact of introducing a new factor in the selection of readings is determined by the size and coverage of the corpus of available readings. For instance, with a small corpus of readings, the REAP tutor would frequently be unable to find a reading on a certain topic that contains more than one of a student's target words. With an infinitely large corpus distributed uniformly with respect to topic and vocabulary, the tutor would always be able to find an interesting reading that provides practice for multiple target words. Therefore, whenever the tutor has trouble satisfying personalization and other constraints, the problem can be mitigated by increasing the corpus of available readings, which will be discussed later.

Incorporating Personalization into a Tutoring System Curriculum

As an additional factor for selecting reading passages, the REAP system also included whether the topic of a reading matched the student's personal interests. The score for topic-interest matching of a passage is defined as the inner product (i.e., dot product) of vectors for student interests in each topic and the passage's membership in the topic categories. Interest values in each topic range from zero to one, and topic category membership values are positive and sum to one. Binary classifiers decide which topic the reading belongs to. Other topics are assigned small, non-zero weights so that readings on those topics have at least a small chance of appearing when the topic of readings cannot be matched to personal interests. For example, a text about a hospital merger might have a 0.5 value membership

in the *Business* and a 0.5 value for membership in *Health* category. If a student had a 0.9 value for interest in both categories, then the topic-interest score for that text and student would be 0.9.

Students take a brief survey before beginning to work with the REAP tutor in order to gather data about personal interests. A screenshot of the survey is shown in Figure 2. Students rate each of the ten general topic categories according to their interest on a scale from one to five. The interest survey results are mapped to values ranging from 0 to 1.


Please mark which topics you want to see readings about.

Category	Examples	Not interested at all	Not very interested	Neither	Somewhat Interested	Very Interested
Arts	literature, movies, TV, music	○	○	○	○	○
Business	investing, market, real estate	○	○	○	○	○
Computers	hardware, software, Internet	○	○	○	○	○
Games	video games, gambling	○	○	○	○	○
Health	fitness, medicine, nutrition	○	○	○	○	○
Home	family, cooking, gardening	○	○	○	○	○
Recreation	travel, outdoors, boating	○	○	○	○	○
Science	biology, astronomy, physics	○	○	○	○	○
Society	politics, religion, sociology	○	○	○	○	○
Sports	baseball, football, basketball	○	○	○	○	○

Continue

Fig. 2. Screenshot of personal interests survey.

Text Classification for Personalization of Reading Material

A set of binary topic classifiers was implemented to automatically classify each potential reading by its general topic. This component enables the REAP tutor to match up texts with student interests. It should be noted that the system does not assign topic labels based on a few keywords for each category, but rather uses a machine learning approach based on thousands of lexical features. By using sophisticated classification algorithms, the system is robust to the noise and variability of the Web pages used by REAP as practice texts.

A Linear Support Vector Machine text classifier (Vapnik, 1995) was trained on Web pages from the Open Directory Project (ODP[5]). These pages effectively have topic labels because they are organized into a multi-level hierarchy of topics. The following general topics were manually selected from the set of top-level categories in the ODP: Arts, Business, Computers, Games, Health, Home, Recreation, Science, Society, Sports. Web pages with these human-assigned topic labels from the ODP were used as gold-standard labels in the training data for the classifiers. These pages were first filtered so that their length was similar to the length of the readings used by the REAP tutor. SVM-

[5] http://dmoz.org

Light (Joachims, 1999) was used as the implementation of the Support Vector Machines. A Support Vector classifier was used because previous research has shown that the technique provides excellent performance on text classification tasks (Joachims, 2002). In preliminary tests, the linear kernel produced slightly better performance than a radial basis function kernel.

Various other techniques were considered, including the Naïve Bayes and *k*-Nearest Neighbor algorithms. A rigorous study by Yang and Liu (1999) showed that *k*-Nearest Neighbor and SVM performed significantly better than Naïve Bayes. Their results do not clearly indicate whether SVM or *k*-Nearest Neighbor performs better in terms of accuracy. However, efficiency concerns led to the choice of SVM in this work. *k*-Nearest Neighbor is a "lazy learning" algorithm that does not produce a model of the training data. Instead, it performs most of its computations at run-time, which would hinder the classification of the large number of texts in REAP's corpus. In contrast, the SVM algorithm is an "eager learning" algorithm that creates a discriminative model at training time and can then classify large numbers of new texts more quickly.

The process of feature selection for the classifiers was as follows. First, the feature set was defined to be the *tf.idf* values for all lexical unigram, or single word, features in the training corpus. *tf.idf* is a common term weighting measure (e.g., Salton & Buckley, 1988). Four hundred stop words, or common words such as "the," were removed from this list. Terms were not stemmed to dictionary forms because stemming does not usually improve performance in information retrieval tasks except when texts are very short (Krovetz, 1993). As such, variants of single root forms were treated as different features (e.g., "remove", "removal", "removed"). The features were ordered by their frequency in the training corpus, and only the most common 10,000 features were retained in the final set of features.

A text classifier was then trained for each topic category, using all Web pages of that topic as positive examples and all Web pages of other topics as negative examples. During evaluations, the system assigned a topic label to a text that corresponded to the classifier with the most positive output value. In the evaluations, the system assigned labels for only one topic because the training data only had labels for one topic. For annotating the texts in REAP's corpus of readings, however, the system assigned labels to a text for all the topics whose binary classifiers had positive values. Multiple labels were used in REAP's corpus since many texts cover multiple topics to some degree, as discussed below in relation to inter-annotator agreement.

Quantitative evaluations were conducted before deployment of the topic classification system. First, the binary classifiers for each topic category were evaluated according to precision, recall, and the *F1* measure (defined below), which are all standard measures for text classification (e.g., Joachims, 2002). In these evaluations, a system-assigned label was considered correct if it matched the gold-standard, human-assigned label from the ODP training data. Accuracy was not used as an evaluation metric since the skewed distributions of positive and negative classes would result in overly optimistic results. For example, a binary classifier for a single topic that always produced negative classifications would be correct about 90% of the time since only 10% of the training texts belonged to that particular topic. Precision, recall, and *F1* are more suitable measures for skewed distributions. *Precision* is defined as the proportion of texts assigned a label correctly out of all of the texts assigned that label. *Recall* is the proportion of texts assigned a label out of all the texts that should have been assigned that label according to gold standard labels. The *F1* measure, the harmonic mean of precision and recall, provides a measure that takes both precision and recall into account. For estimating the performance of each classifier, leave-one-out cross validation was performed using SVM-Light.

Table 1 shows the results of the first classifier evaluations. Precision values range from .70 to .89 and recall values range from .60 to .87 depending on the topic. Similarly, values for the *F1* statistic range from .68 to .86, with a mean value across categories of .76. The performance likely varies across categories due to differences in the degrees to which each category overlaps with others. For example, the Sports category likely performs well due to the fact that it has very indicative, frequently occurring lexical features such as "score" or "ball." These results indicate that the classifiers correctly chose the gold-standard label three quarters of the time. For comparison, random guessing would correctly choose the gold-standard label only one tenth of the time. Thus, while the topic classifications used by the REAP tutor to select readings were not perfect, they were correct most of the time.

Table 1
Precision, Recall, F1 measures by topic for binary classification of texts

Category	Precision	Recall	F1
Arts	.79	.74	.76
Business	.78	.60	.68
Computers	.80	.66	.72
Games	.86	.87	.86
Health	.85	.82	.83
Home	.75	.80	.77
Recreation	.73	.67	.70
Science	.70	.67	.68
Society	.66	.73	.70
Sports	.89	.82	.85
Average	.78	.74	.76

In a second quantitative evaluation, two human annotators labelled the practice texts used in the study conducted in Fall 2006 that will be described in the "Method" section. Each annotator assigned one of the ten general topic labels to each in a randomly-selected subset of 390 texts used in the study. Cohen's *kappa* was calculated to measure inter-annotator agreement. The *kappa* value between the human annotators was .44, indicating moderate agreement according to guidelines proposed by Landis and Koch (1977). The *kappa* values between the topic classification system and the two annotators were .32 and .57, indicating fair and moderate agreement, respectively.

The moderate agreement between the two human annotators demonstrates the difficulty of assigning texts to these ten general categories. The topics of many of the texts were somewhat ambiguous between two or more of the categories, according to the annotators. For instance, an article about a hospital merger could be reasonably labelled as either "Business" or "Health." Changing the granularity of the set of topics by either splitting or merging topics might simplify classification somewhat, but it also might make it more difficult for students to specify their interests. The possibility of using a set of approximately fifty topics was discussed with the instructors for the courses in which REAP was used in the Fall 2006 semester, but it was agreed that the set of ten topics was more reasonable.

Instruction on General-Purpose Vocabulary

With a corpus of texts annotated with general topic categories as described above, the REAP tutor can provide personalized texts that match students' interests. However, personalization does not mean that students learn only narrow-coverage words that relate to their topics of interest. Students with different interests practice similar sets of general-purpose vocabulary. For instance, in the study described in this paper, one student interested in the arts saw the word "endure" in a text describing an artist's early career struggles ("For an artist who has endured so many years of obscurity..."). Another student interested in business saw the same word used to describe economic hardship ("As California has endured a burst tech bubble, costly energy crisis and a staggering burden on its business community...").

The REAP system is flexible in that any target word list can be chosen according to the needs of instructors or researchers. In the study described below, the target words came from the Academic Word List (Coxhead, 2000). The Academic Word List contains 570 word families that are not in the set of the 2,000 most common words in English. These words occur with high and fairly uniform frequency in a variety of subject areas. Some examples of these valuable and broadly applicable words for English learners are the following: "accompany," "demonstrate," "involve," "logic," "status," "ultimate," and "voluntary." In order to avoid contamination, a subset of the Academic Word List was chosen which did not contain words which were covered in the course in which the REAP tutor was used.

METHOD

Participants

An experiment was conducted to test the effects of personalization on interest levels and learning of target vocabulary words in the REAP tutor. Forty-four students at the English Language Institute at the University of Pittsburgh participated in this experiment as part of an intermediate English as a Second Language Reading course in the Fall of 2006. The students were mostly college-age. A variety of nationalities were represented.

The students were randomly assigned to control (n = 22) or treatment (n = 22) conditions. Students were told neither the group to which they were assigned nor the nature of the manipulation of the experiment. The user interface was identical for the two conditions. The time on task available to students was also identical for the two conditions, though minor differences in actual time on task existed due to absenteeism. Students in both conditions worked in the same computer lab, and due to the room's setup, each student could see only his or her own computer screen.

General English Language Proficiency Test Scores as Proxy Pre-test Measure

A measure of general English proficiency was collected for all students. As part of their normal English Language Institute courses, all students took the Michigan Test of English Language Proficiency (MTELP), a general English Language proficiency test, prior to starting work with the REAP tutor. The MTELP assesses proficiency in vocabulary, grammar, and reading comprehension

to determine English as a Second Language proficiency. The test consists of 100 multiple-choice items. This general proficiency measure is included as a covariate in the statistical analysis of results discussed later.

In previous studies with REAP, it has been found that students of higher general proficiency levels learn more vocabulary and perform better on post-tests than those who are less proficient. This finding that better students advanced more quickly is not unique to the REAP tutor, but rather is common in second language learning. It has even been labeled the "Matthew effect" (Folse, 2004)—an allusion to a Biblical passage. In this domain, it refers to the tendency for more proficient students to learn vocabulary at a faster rate than less proficient learners. Including general English proficiency as a covariate controls (i.e., adjusts) for the variance in post-test scores due to differences in the incoming proficiency levels of students. By including these scores in the statistical analysis, students are effectively put on the same level with regard to incoming proficiency. MTELP scores were not reliably different between groups ($p > .10$, two-tailed t-test of independent samples, $t(31) = 1.55$, $p > .10$).

Attrition and mortality threats

Eleven students were dropped from the experiment for various reasons, corresponding to an overall attrition rate of 23%. Four students were dropped from the control condition (18%), and 7 from the treatment condition (35%). Complete data were available for 33 students, 18 in the control condition and 15 in the treatment condition. One student was dropped because he or she requested to be changed from his or her original condition. Another was dropped because he or she was not making an effort to read any of the readings. MTELP data were not available for three students. Most students, however, were dropped because they did not attend class on the day of the post-test. A student's work with the REAP tutor did not count for a grade other than a small portion allotted to attendance. Teachers verified that the overall attrition rate and rates of absenteeism were normal for the English Language Institute. A comparison of the MTELP scores that were available for students who were dropped to scores for students in the study showed that there was no indication of a reliable difference between conditions with respect to incoming English language proficiency (two-tailed t-test of independent samples, $t(42) = 1.09$, $p > 0.10$).

Measures and Procedures

Students completed the self-assessments during the first session. All students also completed the interest survey to inform the tutor about which topics they were interested in. Students began working through readings and practice exercises immediately after the interest survey. There were nine training sessions in total. Sessions were once a week and lasted approximately forty minutes.

For students in the control condition, the REAP tutor ignored student interest survey results and offered readings to students based only on domain goals—that is, optimizing the number of target words per text, reading difficulty level, and text length. For students in the treatment condition, the REAP tutor attempted to choose readings about topics of personal interest to students in addition to satisfying the domain goals.

The target vocabulary words that the REAP tutor presented to these students were chosen from the Academic Word List (Coxhead, 2000). There were 196 words, chosen because they were not part of the course curriculum. Most words were unfamiliar to the students, as they claimed *not* to know

142 of the words on average during self-assessments. The mean number of unfamiliar words was similar in the two groups—143.6 (SD = 31.6) in the control condition, and 137.3 (SD = 39.1) in the treatment condition. The tutor used the words identified as unknown through self-assessments to create an individualized target word list for each student.

Post-test cloze questions

After the nine training sessions, students took a post-test consisting of cloze questions and written sentence production tasks. The post-test questions tested knowledge of target words that were identified as unknown through self-assessments, which ensured that the baseline value for the post-test was approximately zero after correcting for guessing.

The post-test cloze questions were similar to the post-reading cloze questions. The only differences were that the sentences that the student was asked to complete were different, and that all the distracters were target words of the same part of speech. That is, the cloze portion of the post-test was isomorphic to the training task.

The post-test included forty cloze questions. For each student, these questions were divided roughly in half between practiced and unpracticed words. The unpracticed words provided a baseline measure of vocabulary growth against which to compare the treatment, which was of interest for other research purposes. During practice sessions, a student would work at his or her own pace through the readings and practice exercises. As a result, some students practiced more words than other students. Although time on task was controlled for, the number of texts read by students was allowed to vary. Therefore, the number of questions for practiced words on the post-test was different for each student.

The goal of the post-test is to estimate the overall number of words learned. Therefore, the primary outcome measure for the cloze portion of the post-test is the total number of correctly answered cloze questions after correcting for guessing. A standard formula was used for correcting for guessing (Diamond & Evans, 1973):

$$S = \frac{R - W}{k - 1} \tag{1}$$

In this formula, S is the final score, R is the number of correctly answered items, W is the number of incorrectly answered items, and k is the number of options available. This last parameter was 10 on this post-test.

Also of interest are the proportion of post-test cloze questions answered correctly and the number of target words practiced. Some students may have practiced fewer words, but learned those words with greater frequency.

Post-test sentence production transfer task

In addition to the cloze questions, students were asked to write sentences using ten of the words they practiced. This segment of the post-test was designed to measure transfer of word knowledge to a novel and more difficult task. A grading scheme, from 0 to 3, indicated that a first point should be assigned to a sentence if the target word was used in a grammatically correct way, a second point should be assigned if the word fit semantically but did not necessarily demonstrate knowledge of meaning, and a third point should be assigned if the word was used in a way that also demonstrated

knowledge of meaning. An example of correct semantic usage not clearly demonstrating knowledge of a word can be seen in the following sentence for "abandon:" "He *abandon* his work." Note that this would not receive a point for grammar. A third point was assigned if the sentence also clearly demonstrated knowledge of the target word. An example of a student-written sentence that received a full mark is the following: "you must never *abandon* your children on the street like that old lady did."

Sentences were graded by two course instructors and one curriculum supervisor. Each instructor graded the half of the sentences which were written by students in her section of the course. The curriculum supervisor graded the sentences for both sections of the course. In this way, each sentence received two grades, one from the curriculum supervisor and one from a regular instructor. Disagreement between the teachers and curriculum supervisor was resolved by averaging the scores. The correlation coefficient between grades assigned by instructors and the curriculum supervisor was 0.68. This statistic does not include blank responses which would trivially receive a grade of zero. This fairly low value reflects the difficulty of assessing vocabulary knowledge based on this type of written output, although clearer guidelines and better training might have improved inter-rater reliability.

Post-test sentence production performance data were analyzed in terms of the proportion of the maximum possible score. Since there were two graders who assigned up to three points to each of ten questions, the maximum score was 60. To calculate the proportion for a given student, the total number of points assigned from the two graders was divided by this number.

Methods of Analysis

Post-test results were analyzed using a univariate analysis of covariance (ANCOVA), with the given post-test outcome as the dependent variable, condition as the independent variable, and incoming MTELP scores for general proficiency as a covariate. Reading behaviors were analyzed *post hoc* using *t*-tests of partial correlations, controlling for MTELP scores. All statistical tests were two-tailed.

RESULTS

This section describes the results of the experimental study of personalization in the REAP tutor. It begins with the effects of personalization on post-test scores, followed by results pertaining to self-reported interest level, and an analysis of the effects of personalization on reading behaviours and other student interactions with REAP.

Effects of Personalization on Post-Test Measurements of Learning

As expected from previous studies, MTELP scores were positively associated with post-test scores across both conditions. That is, the initially more proficient students tended to perform better on the post-test. The Pearson Product Moment Correlation Coefficients between MTELP scores and the number of correctly answered cloze questions after correcting for guessing ($r = .530$, $p < .01$), the proportion correct of cloze questions for practiced words ($r = .555$, $p < .01$), and the proportion of the maximum possible score for sentence production ($r = .607$, $p < .01$) were all statistically significant. These associations justify the use of MTELP scores to control for incoming English proficiency in the subsequent analyses.

Means and standard deviations for learning outcomes are shown in Table 2. As shown in the first pair of columns, the treatment group correctly answered a slightly higher number of post-test cloze questions after correcting for guessing. However, this difference was not statistically reliable ($F(1, 31) = 2.12$, $p = .16$).

The treatment group practiced fewer words on average than did the control group. This statistically reliable result ($F(1, 31) = 9.48$, $p < .01$) is shown in the second pair of columns in Table 2. The reason is that for students in the treatment condition, the REAP tutor chose reading passages of likely interest with fewer target words. In contrast, for students in the control condition, the tutor gave relatively more weight to the number of target words in a reading because it did not have to include personalization as a factor for selection of practice materials. Other than the number of target words, no obvious differences between the two groups were observed with regards to the sets of words practiced. Neither were there any obvious differences in the amount of information about target words available from context. Finally, other important attributes of the reading passages, such as reading level and text length, were similar in the two conditions. For instance, passages for the control group averaged 780.56 words in length (*SD* = 166.16 words), while passages for the control group averaged 768.59 words (*SD* = 182.36 words).

Table 2
Post-test Results and Number of Practiced Target Words for Treatment and Control Groups
(* indicates a statistically significant difference between treatment and control groups)

	Correctly Answered Cloze Questions, Corrected for Guessing		Total Number of Practiced Words*		Proportion Correct of Cloze Questions for Practiced Words*		Proportion of Maximum Possible Score: Sentence Production*	
Condition	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Control ($n = 18$)	4.65	3.21	16.56	4.99	.34	.18	.27	.19
Treatment ($n = 15$)	5.04	2.72	11.73	3.41	.49	.20	.31	.19

While students in the treatment group practiced fewer words, they performed better on the words they did practice. After controlling for MTELP scores, the mean proportion of post-test cloze items that were correctly answered for practiced words was statistically reliably higher for the treatment group than the control group ($F(1, 31) = 17.58$, $p < .001$). Means and standard deviations are shown in the third pair of columns in Table 2.

The treatment group also performed statistically reliably better on the post-test transfer task. The mean proportion of maximum score for the sentence production items was statistically reliably higher for the treatment condition than the control condition after controlling for MTELP scores ($F(1, 31) = 4.82$, $p < .05$). Data are shown in the rightmost pair of columns in Table 2.

The higher post-test performance for practiced words does not appear to be associated with the smaller number of target words that were practiced. That is, students in the treatment condition were not learning more of the words they practiced simply because they were practicing fewer words. To test the relationship between number of words practiced and post-test performance, the number of

target words per text and total number of target words practiced were included as covariates along with MTELP score in two ANCOVAs with student condition as an independent variable, or fixed factor. The dependent variables in these analyses were the proportion correct of cloze questions for practiced words and the proportion of the maximum possible score for sentence productions.

No significant effects for number of practiced words were found after controlling for condition and MTELP scores. Neither the number of target words per text ($F(1,31) = 0.14$, $p > .10$) nor the total number of target words practiced ($F(1,31) = 0.15$, $p > .10$) was reliably associated with the proportion correct of cloze questions. Also, neither the number of target words per text ($F(1,31) = 0.22$, $p > .10$) nor the total number of target words practiced ($F(1,31) = 0.01$, $p > .10$) was reliably associated with the proportion of the maximum possible score for sentence productions.

Effects of Personalization on Interest

Students were also given an exit survey during their last week of practice with the tutor that asked them, among other questions, to indicate whether they agreed with the statement, "Most of the readings were interesting." The ratings were on a scale from one to five, with five indicating strong agreement and one indicating strong disagreement. Exit survey interest ratings by students in the treatment condition were not reliably higher than the ratings by students in the control condition (one-sided independent samples *t*-test, $t(1) = 0.65$, $p > .10$). The mean response for students who received personalized readings was 2.92, while it was 2.72 for students in the control condition. As such, contrary to expectations, the null hypothesis that self-reported interest was unassociated with the treatment cannot be rejected.

Students in the personalization condition received texts on topics of personal interest with higher frequency. The system's ability to match interests in the two conditions was measured by the proportion of readings whose system-assigned topic had been selected by the student as a topic of interest on the preliminary interest survey (Figure 2) by indicating either "somewhat interested" or "very interested." Students in the control condition specified an average of 5.45 topics as at least somewhat interesting on the survey ($SD = 1.96$), and received readings pertaining to those topics 47.8 percent of the time during practice sessions. Students in the treatment condition chose a similar number of interesting topics ($M = 5.50$, $SD = 2.38$), but received readings pertaining to those topics more often, 73.8 percent of the time.

Effects of Personalization on Reading Behaviors

A number of measures related to reading behaviors were also analyzed *post hoc*. The REAP tutor records various behavioral data, such as how long students spend on each reading passage, which words students look up in the electronic dictionary, and how students perform on post-reading exercises. Descriptive statistics for these data are shown in Table 3.

Most measures were not reliably different between groups. For instance, students in the treatment group did not look up significantly more or fewer words in the dictionary than students in the control group ($F(1, 31) = .764$, $p > .10$). Neither did they spend significantly more or less time on each passage ($F(1, 31) = 1.26$, $p > .10$).

Post-reading vocabulary practice exercise performance was also not reliably different between the groups. That is, practice cloze exercise performance was not different between groups even though post-test cloze performance for practiced words was. However, there was a reasonably high partial

correlation ($r(31) = .286$, $p = .11$) between mean practice exercise performance and condition when controlling for MTELP score.

Students in the treatment group did, however, perform reliably better on "reading check" questions than their counterparts in the control group. There was a statistically reliable partial correlation ($r(31) = .421$, $p < .05$) between reading check performance and condition after controlling for MTELP score. This suggests that perhaps personalization led students to read passages more closely because of increased interest. However, no strong conclusions should be made based on this *post hoc* analysis. As such, analysis of reading behaviors did not yield any strong evidence of the mechanisms by which personalization might improve learning.

Table 3
Measures for Reading Behaviors Including Dictionary Access and Time per Reading
(* indicates a statistically significant difference between treatment and control groups)

	Control Condition		Treatment (Personalization) Condition	
Measure	Mean	SD	Mean	SD
Time Per Reading	20.74	8.03	23.86	10.29
Number of Readings Completed	16.22	8.05	13.27	4.48
Number of Target Word Dictionary Lookups per Reading	1.72	.95	1.46	0.87
Number of Non-target Word Dictionary Lookups per Reading	7.08	4.41	7.35	4.26
Reading-Check Question Proportion Correct *	.47	.32	.62	.25
Practice Vocabulary Exercises Proportion Correct	.37	.21	.40	.21

FOLLOW-UP STUDY

In the Fall 2006 study described in the previous sections, the number of available texts necessary to provide practice readings that are both personalized and provided sufficient practice was underestimated. The effect of the small corpus of texts is evident in the low number of target words practiced by students receiving personalization. In a follow-up study in Spring 2007, during which all students received personalized readings, the coverage of target words and topics in the corpus was improved by more than doubling the size of the corpus. Additionally, improvements to the learner model in REAP, in particular the use of knowledge tracing, reduced the number of practice repetitions

of words for which students showed high practice exercise performance—which led to higher efficiency.

Due to the larger corpus and improved learner modelling in the Spring 2007 semester, students received personalized readings with more practice opportunities than in the original study. The distribution of students' proficiency levels, the total time on task, the methods, post-test instruments, and other procedures for the Spring 2007 study were similar to those of the Fall 2006 study. Moreover, the difficulty level of the target words and the passages were similar. Twenty-eight students participated in the study and completed the post-test.

Table 4 shows post-test results and number of practiced words for the Spring 2007 study alongside the corresponding data from Table 2 for the Fall 2006 study, for comparison. Students practiced 23.7 target words on average in Spring 2007 compared to 16.6 for students in the control condition of the Fall 2006. Thus, with a larger database of readings the REAP tutor was able to provide both personalization and multiple practice opportunities.

Students in the Spring 2007 study also correctly answered nearly the same proportion of post-test cloze questions (M = .47, SD = .23) for target words as the personalization treatment group in the Fall 2006 study (M = .49, SD = .20). The increase in practice opportunities and maintenance of post-test scores suggests that improving corpus coverage addressed the challenge of finding interesting passages that also provide practice opportunities for many target words. Thus, tutoring systems with sufficiently large databases of materials can personalize practice without sacrificing domain-based, or pedagogical, goals.

Table 4
Post-test Results Number of Practiced Target Words for Follow-up Study with Larger Database of Available Texts

	Correctly Answered Cloze Questions, Corrected for Guessing		Total Number of Practiced Words		Proportion Correct of Cloze Questions for Practiced Words		Proportion of Maximum Possible Score: Sentence Production	
Condition	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Fall 2006, no personalization (n = 18)	4.65	3.21	16.56	4.99	.34	.18	.27	.19
Fall 2006, with Personalization (n = 15)	5.04	2.72	11.73	3.41	.49	.20	.31	.19
Spring 2007, with Personalization (n = 28)	9.54	9.08	50.25	19.06	.47	.23	.27	.20

The mean proportion of maximum possible score for the sentence production tasks, however, was slightly lower in the Spring 2007 study compared to the treatment condition in Fall 2006. One

possible explanation is that students did not take the sentence production test as seriously in the Spring 2007 study due in part to the fact that performance on the post-test did not affect students' grades in the course. In fact, there were two students who left all sentence production responses blank, which did not happen at all in Fall 2006. Also, it should be noted that the sentence production tasks assessed knowledge of only 10 words even though students practiced 50.25 on average in the Spring 2007 study.

DISCUSSION

The results of the studies provide support for three main conclusions. First, it appears that personalization to match interests can lead to improved learning of the relevant knowledge components in a tutoring environment for vocabulary learning. Students in the treatment group correctly answered a higher proportion of questions on target words that were practiced in the REAP tutor.

Second, personalization can compromise domain-based goals. In the REAP tutor, an important domain-based goal is to give the student practice opportunities for many new target words. However, in the Fall 2006 study, students receiving personalization practiced fewer target words. The difficulty in achieving the domain-based goal of practicing many unknown words is due to the fact that the REAP tutor often could not find texts that included multiple target words and also matched personal interests. Of course, while sufficient practice is desirable, the density of target word practice should not overwhelm the student. For example, if every other content word in a text were unknown, a student would lack a basis on which to make inferences about target word meanings from context. However, showing too many target words is not much of an issue for the REAP tutor since it controls texts for reading level and only shows texts with small percentages of unknown words (i.e., < 5%).

Third, in the domain of second language vocabulary learning and reading practice, it is possible to partially automate the selection of personalized materials. Specifically, a tutoring system can use automatic text classifiers to differentiate between aspects of texts such as general topic area which are relevant to personal interests.

It appears that if the challenges of negotiating personalization and domain-based goals are met, then personalization can lead to improvements in overall learning. Students with personalization appeared to learn the words they practiced with greater frequency but practiced fewer target words, and as a result did not perform reliably differently than their controls on the overall post-test measure for cloze questions. The researchers attributed this lack of a difference to the fact that, in many cases, the tutor had to choose between interesting readings and those with more practice opportunities. However, the availability of readings that are both interesting and provide ample practice is a technical issue which can be solved in a straightforward manner by increasing the size and coverage of the corpus of available practice reading passages, as shown by the results from the Spring 2007 follow-up study.

Exit survey results did not indicate a reliable increase in self-reported interest between students working with the tutor that matched texts to their interests. Other, less subjective methods of measuring interest, such as time spent per text or number of definitions for non-target words accessed, might be more reliable indicators of interest than self-reporting at the end of a semester. However, positive feedback from teachers, indicates that REAP can effectively match practice texts to personal interests, and that students are more interested as a result.

Interestingly, as a result of using personalization in the Fall 2006 and Spring 2007 semesters, the ESL teachers at the University of Pittsburgh are convinced that REAP is much better with personalization. As such, they are now opposed to allowing any future experimental studies on personalization that involve a control condition without it. More complex experimental designs may be necessary.

Open Research Issues

Issues related to computerized vocabulary tutoring

One general observation about the results is that the post-test scores were fairly low overall. The developers of REAP identified a number of areas with room for improvement. One is that students need to practice a greater number of unknown target words per text, which was partially addressed by improving coverage of the corpus as described above. In addition, the number of possible target words was expanded to include all 570 head words of the Academic Word List (Coxhead, 2000) instead of the subset of approximately 200 words used in the study. Knowledge tracing was also deployed to more accurately estimate which words students should practice at any given time. The developers have also begun to focus on increasing the number and variety of practice vocabulary exercises that follow reading passages. In the personalization study, students received only one practice cloze exercise per target word after a reading. The developers have been working on incorporating other practice exercises such as synonym questions (Brown, Eskenazi, and Frishkoff, 2005), and also questions about word associates (Read, 1998) that are automatically generated using thesaurus extraction algorithms (Heilman and Eskenazi, 2007).

Issues related to improvements in personalization of reading passages

Also, more sophisticated techniques for personalization could lead to even greater increases in interest and intrinsic motivation. The current system used in REAP only classified texts into ten general topic categories, but finer-grain classification might lead to greater interest. For example, a student may be interested in chemistry but not physics. Topic selection by general categories would in that case not distinguish between interesting and boring texts. One issue with implementing finer-grain topic classification for readings is the method for determining student interests. Though a questionnaire can quickly measure personal interest in ten categories, a questionnaire for 100 or more finer-grain categories would probably not be feasible. One possibility is to ask the student directly about personal interest in general categories (e.g., Science), and then adjust interest estimates for finer-grain categories (e.g., Chemistry) based on self-reported interest after readings.

Additionally, it may be useful to incorporate confidence measures for the topic classifications in REAP. In experimental tests of classification accuracy, the classifiers chose a topic other than the gold standard approximately one quarter of the time. As noted, many of these choices may be due to texts pertaining to multiple topics at once. However, classification errors certainly do occur, and mitigating such errors is important. Confidence measures, available from many types of classifiers, may be useful to avoid incorporating erroneous classifications into tutorial decision-making. For instance, if the classifier is less confident about a particular text, then the tutoring system may want to give less weight to personal interest when deciding whether that text should be selected.

Issues related to negotiating motivational and domain-based goals

However, more sophisticated personalization might require better negotiation between domain goals and motivational goals. The system might need mechanisms for detecting levels of student motivation and altering the selection of practice materials accordingly, as suggested by del Soldato and du Boulay (1995). For example, more weight could be given to personalization when the system detected a low level of student interest. Even though personal interest was the focus of this work, while perceived self-efficacy and situational interest was the focus of del Soldato and du Boulay's work, similarities appear to exist with respect to the need for negotiation of domain-based and motivational goals. Namely, in both cases, if too much weight is given to either domain-based or motivational goals, then goals of the other type may be compromised. Tutoring systems must account for this conflict, possibly by assigning weights to prioritize more important goals, or by defining a set of negotiation rules.

The REAP tutor must weigh a variety of different domain and motivational factors in its selection of practice readings. These factors include reading difficulty level, text length, number of target word practice opportunities, estimates of target word knowledge—and, of course, personalization. To lend more or less importance to each factor, the developers chose weighting parameters manually, but ideally these parameters would be estimated using data including learning outcomes. However, large amounts of data are unavailable, especially at the start of a new semester with a new set of students, target words, and conditions. Data from previous semesters are commonly used to estimate parameters and refine models in intelligent tutoring systems (e.g., Cen, Koedinger, and Junker, 2007). However, ongoing development of the system, such as changes of the form of practice exercises, may render previous data unusable for parameter estimation. It would be valuable to develop robust methods for estimating such parameters using only sparse data from an ongoing semester or study.

Extensions to other domains

The results of this study agree with prior findings that personal interest is associated with learning. Matching personal interests appears to be more suitable for teaching skills rather than content (Fives and Manning, 2005). When content is the focus of instruction, increasing situational interest seems to be a more suitable means of motivation. As an extreme example, consider a case in which a history tutor has to teach a lesson on the early space program, and the student is mainly interested in music. The tutor probably could not alter the lesson to match personal interests. Instead, the tutor might increase situational interest in the lesson by using an interactive game, incorporating relevant multimedia, presenting material in a vivid and coherent manner, etc. However, for areas like reading or vocabulary, where the target knowledge components are more or less content-independent skills, personal interest is a useful motivational tool.

The approach to personalization described in this paper relies on the availability of a large database of practice materials, which in this case is a corpus of potential readings. Such large databases of practice materials may not always be available in other domains. However, Aleahmad, Aleven, and Kraut (2008) describe how to create such a database using a community authoring approach. Aleahmad and colleagues created a web site that asked instructors and other visitors to write personalized math problems for hypothetical students. Visitors to the site could also rate, make corrections to, and download existing problems for their own use. Their findings suggest that creating a large database of practice materials for personalization is feasible for domains other than vocabulary

learning. Thus, it would be valuable for future research to explore the extent to which personal interest can be used for skills in other domains such as computer programming or mathematics.

Developers of e-Learning environments must take into account both motivational and cognitive factors. These two types of factors can often interact, as in the case of selecting texts that are of personal interest, at the right difficulty level, and contain a sufficient number of practice opportunities. The coherence of text is another case of such an interaction since it can both increase situational interest and to facilitate information processing. This work also highlighted the importance of balancing motivational and domain-based, cognitive goals.

ACKNOWLEDGMENTS

The authors acknowledge Jon Brown and James Sanders for their work on the REAP project. Also the authors would like to thank the teachers at the English Language Institute for using the REAP tutor in their classes. Thanks go to the anonymous reviewers as well as Robert G. M. Haussman for their comments. This work was supported by the Institute of Education Sciences of the Department of Education (Grants R305G03123 and R305B040063), the National Science Foundation through a Graduate Research Fellowship and through the Pittsburgh Science of Learning Center (Grant SBE-0354420), and a Siebel Scholarship awarded to the first author. Any opinions, findings, conclusions or recommendations expressed in this material are the authors', and do not necessarily reflect those of the sponsor.

REFERENCES

Aleahmad, T., Aleven, V., & Kraut, R. (2008). Open community authoring of targeted worked example problems. In *Proceedings of the Eleventh International Conference on Intelligent Tutoring Systems*. Montreal, Quebec, Canada.

Bandura, A. (1997). *Self-Efficacy: The Exercise of Control*. New York: W.H. Freeman.

Beck, J. (2007). Does learner control affect learning? In *Proceedings of the 13th International Conference on Artificial Intelligence in Education*. Los Angeles, CA.

Brantmeier, C. (2006). Toward a multicomponent model of interest and L2 reading: Sources of interest, perceived situational interest, and comprehension. *Reading in a Foreign Language, 18 (2)*.

Brown, J., Frishkoff, G., & M. Eskenazi. (2005). Automatic question generation for vocabulary assessment. In *Proceedings of HLT/EMNLP 2005*. Vancouver, B.C.

Brown, J. & Eskenazi, M. (2004). Retrieval of authentic documents for reader-specific lexical practice. In *Proceedings of InSTIL/ICALL Symposium 2004*. Venice, Italy.

Cen, H., Koedinger, & K., Junker, B. (2007). Is Over Practice Necessary? Improving Learning Efficiency with the Cognitive Tutor through Educational Data Mining. In *Proceedings of the 13th International Conference on Artificial Intelligence in Education.* Marina del Rey, CA.

Chan, T. (1996) Learning companion systems, social learning systems, and the global social learning club. *Journal of Artificial Intelligence in Education*, 7(2).

Clark, R. C. & Mayer, R. E. (2003). *e-Learning and the Science of Instruction*. Jossey-Bass/Pfeiffer.

Collins-Thompson, K. & Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, 56 (13).

Corbett, A. T. & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction,* 4(4). Springer Netherlands.

Cordova, D. I. & Lepper, M. R. (1996). Intrinsic Motivation and the Process of Learning: Beneficial Effects of Contextualization, Personalization, and Choice. *Journal of Educational Psychology*, 88(4), 715-730.

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

Deci, E. L. & Ryan, R. M. (1985). The relation of interest to the motivation of behavior: A self-determination theory perspective. In K. A. Renninger, S. Hidi, & A. Krapp (Eds.) *The role of interest in learning and development*, 43-70. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

del Soldato, T., & du Boulay, B. (1995). Implementation of motivational tactics in tutoring systems. *Journal of Artificial Intelligence in Education*. 6(4), 337-378. Association for the Advancement of Computing in Education.

de Ridder, I. (2002). Visible or Invisible Links: Does the Highlighting of Hyperlinks Affect Incidental Vocabulary Learning, Text Comprehension, and the Reading Process? *Language, Learning & Technology*, 6(1), 123-146.

Diamond, J., & Evans, W. (1973). The Correction for Guessing. *Review of Educational Research,* 43(2), 181-191.

Fives, H & Manning, D. K. (2005). Teachers' Strategies for Student Engagement: Comparing Research to Demonstrated Knowledge. Annual Meeting of the American Psychological Association*,* Washington DC.

Flowerday, T., Schraw, G., & Stevens , J. (2004). The Role of Choice and Interest in Reader Engagement. *The Journal of Experimental Education,* 72(2), 93-114.

Folse, K. (2004). *Vocabulary Myths: Applying Second Language Research to Classroom Teaching*. Ann Arbor, MI: University of Michigan Press.

Harp, S. F., & Mayer, R. E. (1998). How seductive details do their damage: A theory of cognitive interest in learning. *International Journal of Educational Psychology*, 90, 414-434.

Heilman, M., Collins-Thompson, K., Callan, J. & Eskenazi, M. (2006). Classroom success of an Intelligent Tutoring System for lexical practice and reading comprehension. In *Proceedings of the Ninth International Conference on Spoken Language Processing*. Pittsburgh, PA.

Heilman, M. & Eskenazi, M. (2007). Application of Automatic Thesaurus Extraction for Computer Generation of Vocabulary Questions. In *Proceedings of the SLaTE Workshop on Speech and Language Technology in Education.* Farmington, PA.

Hidi, S. (2001). Interest, Reading, and Learning: Theoretical and Practical Considerations. *Educational Psychology Review,* 13(3), 191-209. Springer Netherlands.

Joachims, T. (2002). *Learning to Classify Text Using Support Vector Machines*. Kluwer/Springer.

Joachims, T. (1999). Making large-Scale SVM Learning Practical. In B. Schölkopf and C. Burges & A. Smola (Eds.) *Advances in Kernel Methods – Support Vector Learning.* MIT-Press.

Johnson, W. L., Rickel, J. W., & Lester, J. C. (2000). Animated Pedagogical Agents: Face-to-Face Interaction in Interactive Learning Environments. *International Journal of Artificial Intelligence in Education,* 11, 47-78. IOS Press.

Kay, J. (2001). Learner Control. *User Modeling and User-Adapted Interaction*, 11, 11-127. Netherlands: Kluwer Academic Publishers.

Knight, S. (1994). Dictionary use while reading: The effects on comprehension and vocabulary acquisition for students of different abilities. *The Modern Language Journal*, 78, 285–299.

Klawe, M. (1998). When Does The Use Of Computer Games And Other Interactive Multimedia Software Help Students Learn Mathematics? *NCTM Standards 2000 Technology Conference,* Arlington, VA.

Koedinger, K., Anderson, J. R. Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, 8, 30-43.

Krovetz, R. (1993). Viewing morphology as an inference process. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 191–202.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.

Laufer, B. & Nation, P. (1999). A vocabulary size test of controlled productive ability. *Language Testing*, 16(1), 33-51.

Lepper, M. (1988). Motivational Considerations in the Study of Instruction. *Cognition and Instruction*, 5(4), 289-309.

Liu, C, Wang, C. Gao, Z. & Huang, S. (2005). Applications of Lexical Information for Algorithmically Composing Multiple-Choice Cloze Items. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP,* Ann Arbor, Michigan. Association for Computational Linguistics.

Malone, T. W. & Lepper, M. R. (1987). Making Learning Fun: A Taxonomy of Intrinsic Motivations for Learning. In R. E. Snow and M. J. Farr (Eds.) *Aptitude ,learning and Instruction*. NJ: Lawrence Erlbaum.

Meara, P. (1992). *EFL Vocabulary Tests*. Swansea: Center for Applied Language Studies.

McLaren, B. M., Lim, S., Gagnon, F., Yaron, D., & Koedinger, K. R. (2006). Studying the Effects of Personalized Language and Worked Examples in the Context of a Web-Based Intelligent Tutor. In M. Ikeda, K. Ashley, T. Chan (Eds.) *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*. Springer.

Randel, J. M., Morris, B. A., Wetzel, C. D., & Whitehill, B. V. (1992). The Effectiveness of Games for Educational Purposes: A Review of Recent Research. *Simulation & Gaming*, 23(3), 261-276.

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. J. Kunnan (Ed.) *Validation in language assessment*. Lawrence Erlbaum Associates.

Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24, 513-523.

Schraw, G. & Lehman, S. (2001). Situational Interest: A Review of the Literature and Directions for Future Research. *Educational Psychology Review.* 13(1), 23-52. Springer.

Shiefele, U. (1999). Interest and Learning from Text. *Scientific Studies of Reading*, 3(3), 257-279.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer.

Yang, Y. & Liu, X. (1999). A Re-Examination of Text Categorization Methods. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* 42-49.