International Journal of Artificial Intelligence in Education Volume# (YYYY) Number
IOS Press

Self-Assessment in the REAP Tutor: Knowledge, Interest, Motivation, & Learning

Kevin Dela Rosa, *Amazon.com, Seattle WA*

Maxine Eskenazi, *Language Technologies Institute, Carnegie Mellon, Pittsburgh PA*

Abstract. Self-assessment questionnaires have long been used in tutoring systems to help researchers measure and evaluate various aspects of a student's performance during learning activities. In this paper, we chronicle the efforts made in the REAP project, a language tutor developed to teach vocabulary to ESL students through reading activities, to understand the usefulness of self-assessment questionnaires for gauging knowledge, motivation, and interest. Additionally, we discuss the appropriate use of self-assessment questions and correlations we have found with learning and user behavior.

Keywords. Self-Assessment, Intelligent Tutoring Systems, Motivation Diagnosis, Motivation Modeling, Language Learning

INTRODUCTION

Students who use a tutoring system are generally evaluated based on the way they perform on its tasks. Assessing students and performance may be labor and time intensive depending on the tasks completed. The use of self-assessments, like questionnaires, is a simple yet powerful method of evaluating students. Self-assessments are simple for computer-assisted tutoring systems to present and evaluate, and generally take a relatively small amount of time to administer. While self-assessments are powerful, it is also important to use them correctly.

This paper chronicles how self-assessments have been used and progressively improved through studies associated with the REAP tutoring system, an English as a second language vocabulary tutor (Brown & Eskenazi, 2004; Collins-Thompson et al., 2004). Over the course of the REAP project, self-assessments have been used to help measure knowledge and motivation. Additionally, self-assessments have also been used to understand student's reading topic interests and have helped REAP provide personalized instruction in the form of readings tailored to a student's personal interests. In these studies we have learned about the usefulness of self-assessment in different aspects of computer-assisted language learning, and we describe the key findings and implications of those studies in this paper. Furthermore, we discuss the appropriate use of self-assessment questions and correlations between these questions with learning and user behavior.

OVERVIEW OF THE REAP TUTORING SYSTEM

The REAP tutoring system was used to conduct all of the studies described in this paper. REAP, which stands for **REA**der-specific **P**ractice, is a web-based language tutor developed at Carnegie Mellon that harvests documents from the internet for English as a second language vocabulary learning and reading comprehension (Brown & Eskenazi, 2004; Collins-Thompson et al., 2004). The REAP tutor has the ability to provide reader-specific passages by consulting profiles that model a reader's reading level, topic interests, and vocabulary goals. The vocabulary words taught through REAP comes from the Academic Word List (Coxhead, 2000). Also, REAP has been used as a testing platform for cognitive science studies (Kulkarni et al., 2008; Dela Rosa & Eskenazi, 2010; Dela Rosa & Eskenazi, 2011a).

REAP's interface has a number of features that enhance a student's learning experience. One key feature is that it provides users with the ability to listen to the spoken version of any word that appears in a document, making use of the Cepstral Text-to-Speech system (2001) to synthesize words on demand when clicked on by the students. Additionally, students can look up the definition of any of the words they encounter while reading the documents using an embedded electronic dictionary, whose definitions come from the Cambridge Advanced Learner's Dictionary (Walter, 2005). The system also automatically highlights focus words, i.e. the words targeted for vocabulary learning in a particular reading.

All of the REAP studies described in this paper were conducted with a subject population of English as a second language college students. These students participated in the studies as part of either an intermediate or advanced ESL reading course at the University of Pittsburgh's English Language Institute. The native languages of the participants varied over the years, but typically included large populations of Arabic, Chinese, Korean, and Spanish speaking students. The studies conducted with the REAP tutor generally include a pre-test, consisting of either self-assessment or cloze questions (a.k.a. fill-in the blank questions, usually multiple choice), or both, a series of reading passages given as either homework or as in-class activities followed by practice vocabulary questions on the target words, and a post-test for vocabulary assessment, with motivation and interest surveys given at various times during the study.

KNOWLEDGE ASSESSMENT

One of the most common uses of self-assessment is for assessing a student's knowledge, often times as an initial assessment of the student's skills and knowledge. The obvious advantage of using self-assessment for knowledge is a reduction in time needed to assess a student's knowledge, but there are many things that must be considered when using self-assessment for knowledge. In this section we provide a discussion on past research conducted in this domain and also detail studies findings from studies conducted with REAP related to the implications of using self-assessment instruments for assessing second language vocabulary learning student's knowledge.

Background

Self-assessment has long been used and studied in education, particularly in tutoring and self-directed learning settings. Self-assessment has several advantages. LeBlanc & Painchaud (1985) pointed out that self-assessments can take much less time to administer than other methods of assessment, and since students are being asked how they feel about performing a task, the method of testing and collecting data becomes much simpler. Moreover, self-assessments eliminate the need for safeguards

against cheating, allowing students to fill out surveys at their leisure, which in turn reduces the pressure to have tight testing schedules. Also, self-assessments can help make students more active, help them perceive their own progress, and encourage them to see the value in what they are learning (Harris, 1997).

But the use of self-assessment comes with many caveats. For instance, a meta-review by Boud & Falchikov (1989) on self-assessment in higher education found that mature and competent students are able to rate themselves identically as a teacher would, but some students are supercritical of their own deficiencies, particularly those students working in a new subject, which can lead students to underrate themselves. They also found that weaker, less mature students tend to overrate themselves, since they tend to either not be aware of their deficiencies or do not subscribe to the teacher's standards, leading them to be optimistic about their abilities.

Researchers in language learning have also studied self-assessment. For example, Malabonga et al. (2005) found that a majority of students (92%) were successful in using a self-assessment instrument to select test tasks at an appropriate starting difficulty level. Harrington & Carey (2009) found that vocabulary knowledge self-assessment with binary choices were approximately as effective as grammar placement tests. Brantmeier (2006) found that self-assessments on reading ability were not reliable in predicting performance on computer-based reading activities. More over, Cole et al. (2010) found that self-assessments could be used in eliciting topic knowledge.

An important issue in using self-assessment for knowledge is how reliable self-assessment questions are when dealing with second language vocabulary learners, compared to other types of questions. In the sections below we describe a study that compares the effectiveness of self-assessment and cloze questions in assessing a L2 vocabulary student's initial knowledge, describe the use of self-assessment for evaluating a student's aural knowledge of a word, and provide commentary on ways of augmenting self-assessment results to tease out guesswork and over estimation.

Self-Assessment in Second Language Vocabulary Instruction

In one of the first REAP studies on self-assessment Heilman & Eskenazi (2008) explored the usefulness of self-assessment in evaluating second language vocabulary knowledge. This study investigated how reliable it is to use self-assessments as part of an initial assessment of a student's knowledge of individual target words. The primary motivation for using self-assessments was to reduce the amount of time it took to assess a student, while still maintaining a reasonably accurate, though not perfect, assessment of the student's knowledge of the target vocabulary words.

Heilman & Eskenazi compared student's performance on a series of pre-test cloze questions, example shown in Figure 1, and responses to a pre-test self-assessment prompt, example shown in Figure 2, which asked a student to indicate whether they knew a word or not, to a series of follow-up cloze questions on the target words. With respect to response times, they found that it took students significantly longer to answer a cloze question than a self-assessment question, with an average response time of 38.8 and 6.1 seconds respectively, as shown in Figure 3. This average time difference means that if 100 questions were administered, one would expect to save approximately 54 minutes if self-assessment questions were used instead of cloze questions.

Initial analysis of the agreement between the pre-test and follow-up questions showed that responses to the cloze questions agreed 74.8% of the time (i.e. the student either answered both correctly or incorrectly) while the responses to the self-assessment agreed 56.4% of the time which is slightly better than random chance. Contingency tables, Table 1 and Table 2, were used to more

closely analyze the results, where the cells correspond to percentage of agreement between question pairs. Upon further inspection, they found that self-assessment questions were more accurate when students claimed to not know a word. For example, when students claimed a word as unknown they answered the follow-up questions incorrectly 91.7% of the time, while when students claimed to know a word they answered the follow-up question correctly only 29.0% of the time.

These results suggest that while self-assessment questions are much faster to administer than cloze questions, they are not generally as reliable as cloze questions, though self-assessment does appear to be reliable when student's claim that they do not know a word. More over, these results were used to develop a student learner model currently used in REAP whose initialization is based on self-assessment responses by assigning low probability to words students claim as unknown and moderate probability to words students claim to know.

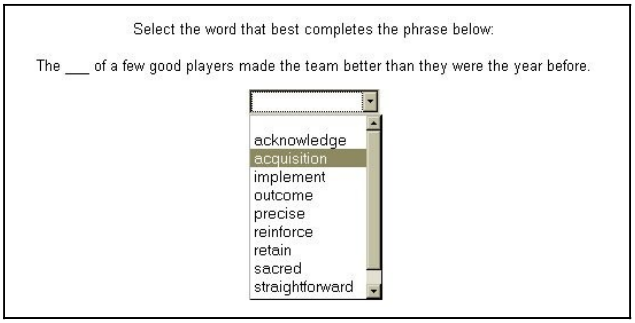


Fig. 1. Example cloze question in REAP. (From Heilman & Eskenazi, 2006).

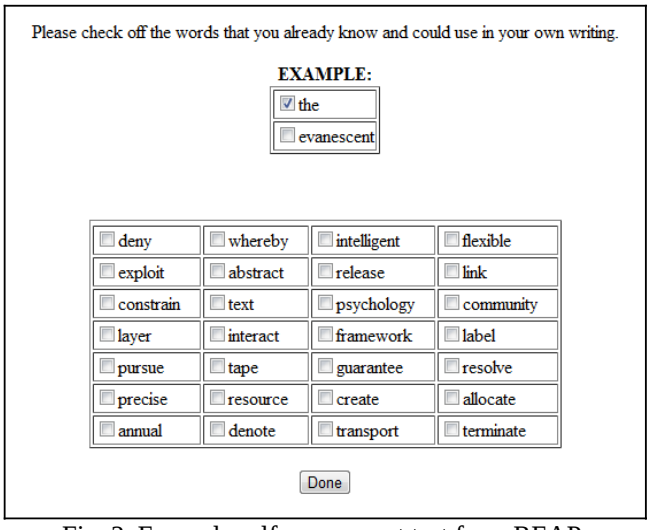


Fig. 2. Example self-assessment test from REAP.

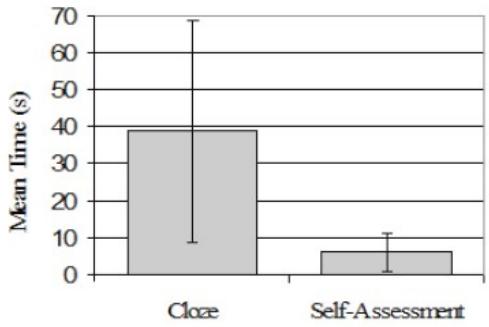


Fig. 3. Average time taken to answer questions by type. (From Heilman & Eskenazi, 2008).

Table 1
Contingency table for pairs of pre-test cloze and follow-up questions.
(Adapted from Heilman & Eskenazi, 2008)

	Cloze-Correct	Cloze-Incorrect	Total
Follow-up-Correct	16 (4.8%)	50 (15.2%)	66 (20.0%)
Follow-up-Incorrect	33 (10.0%)	231 (70.0%)	264 (80.0%)
TOTAL	49 (14.8%)	281 (85.2%)	330 (100%)

Table 2
Contingency table for pairs of pre-test self-assessment and follow-up questions.
(Adapted from Heilman & Eskenazi, 2008)

	Self-Assessment-Known	Self-Assessment-Unknown	Total
Follow-up-Correct	54 (16.3%)	12 (3.6%)	66 (19.9%)
Follow-up-Incorrect	132 (39.9%)	133 (40.2%)	265 (80.1%)
TOTAL	186 (56.2%)	145 (42.8%)	331 (100%)

Dela Rosa et al. (2010) conducted a related study in REAP with self-assessment questions similar to those used by Heilman & Eskenazi (2008) aimed at determining whether students knew the aural form of a word (the screenshot of the interface that was used is shown in Figure 4). Students played a sound clip of a word, produced via Cepstral Text-to-Speech (2001), by clicking on a green speaker icon, and were then asked to indicate whether they knew the word or not. In an effort to compensate for guesswork and overestimation in self-assessment, which was likely the case in the Heilman & Eskenazi study (2008), pseudo-words were introduced in the self-assessment pre-test. The formula that calculated student's performance on the test penalized a student's raw score if they claimed to know a pseudo-word (Milton & Hopkins, 2006), and essentially boils down to subtracting the percentage of pseudo-words that are claimed to be known from the self-assessment score of the actual words with care being taken to provide an appropriate number of pseudo-words in comparison to real words.

They found that this self-assessment with penalization of pseudo-words was an effective way of measuring student knowledge. Also, they found that the speech synthesis was good enough for both native speakers and non-native speakers to disambiguate between words they knew and words they didn't know, as made evident by the fact that both groups of speakers followed similar trends in the number of times they listened to words & pseudo-words in the pre-test, illustrated in Table 3 and Table 4.

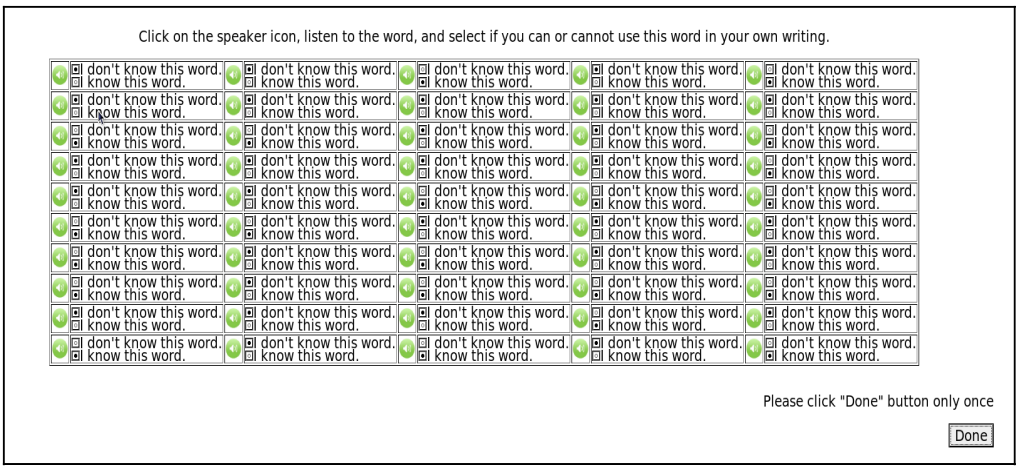


Fig. 4. Screenshot of Auditory Self-Assessment.

Table 3
Average number of times a word was listened to during self-assessment by Non-Native Speakers.
(Adapted from Dela Rosa et al., 2010)

	Known Words	Unknown Words	All Words
Actual Words	1.247	1.594	1.319
Psuedo-Words	1.625	1.830	1.824
All Words	1.253	1.693	1.403

Table 4
Average number of times a word was listened to during self-assessment by Native English Speakers.
(Adapted from Dela Rosa et al., 2010)

	Known Words	Unknown Words	All Words
Actual Words	1.073	---	1.073
Psuedo-Words	2.750	1.397	1.500
All Words	1.098	1.397	1.144

INTEREST ASSESSMENT & PERSONALIZATION

Surveys that gauge a student's interest in a task have often been used to evaluate people's satisfaction with a tutor. In REAP, we went a step further and made use of these assessments to help provide personalized readings in an attempt to help increase intrinsic motivation and subsequently learning. In this section we provide a discussion on previous work on improving intrinsic motivation and discuss the results of a study conducted in REAP that provided students with personalized instruction based on topical interests.

Background

Past research has shown that there is value in being intrinsically motivated in many settings, such as education and work environments, and much research has been done to find ways of increasing

intrinsic motivation. In a meta-analysis Deci et al. (1999) found that giving students rewards can be detrimental in that they undermine people from taking responsibility for motivating and regulating themselves. More over, a study by Schiefele (1991) found that interest is important for the depth of text comprehension, the use of learning strategies, and the quality of emotional experience while learning.

A key issue in intelligent tutoring is keeping students interested in the task at hand. One method that has been used to help increase interest and intrinsic motivation in learning has been the use of personalized instruction. For example, a study by Born et al. (1972) showed that psychology students in a personalized instruction courses performed better than students in a normal lecture section on their examinations, particularly when students had 'average' to 'poor' academic records. Additionally, Cordova & Lepper (1996) found contextualization, personalization, and choice to have a dramatic impact in elementary school children's motivation and engagement in learning. Moreover, Ku & Sullivan (2002) found that fourth grade math students had significantly greater pretest-to-posttest gains when provided with personalized instruction, and similar results were found for fifth and sixth graders (Anand & Ross 1987).

Many researchers have developed tutors and learning related tools that incorporate some sort of personalization (e.g., Gilbert & Han, 2002; Chen, 2008). For example, Martin & Rosa (2009) used an adaptive mobile learning environment that recommended different additional learning activities for computer science courses based on the students' learning style, previous activities, and context. Additionally, Graham (1999) introduced Reader's Help, a document reading environment that helps users read documents more efficiently by annotating and scoring documents based on the reader's topic interests.

In the following section, we describe a study conducted with REAP that made use of interest survey's to help provide personalized instruction.

Personalization Based on Topical Interests

At the end of one of the first REAP studies where the tutoring system was deployed in a classroom setting, the participating students were asked to assess the difficulty of the readings they encountered and how interested they were in the passages through an online survey. Heilman et al. (2006) found that while the readings that REAP was providing were at the appropriate difficulty level, they were not always engaging as made evident by the survey question relating to reading passage interest, shown in Figure 5. Moreover, when asked in the exit survey, students overwhelmingly indicated that they wanted to see more readings on topics they were interested, as shown in Figure 6. These findings led to a new research focus in the REAP project on interest and motivation, the former discussed in this section and the latter discussed in the following section of this paper.

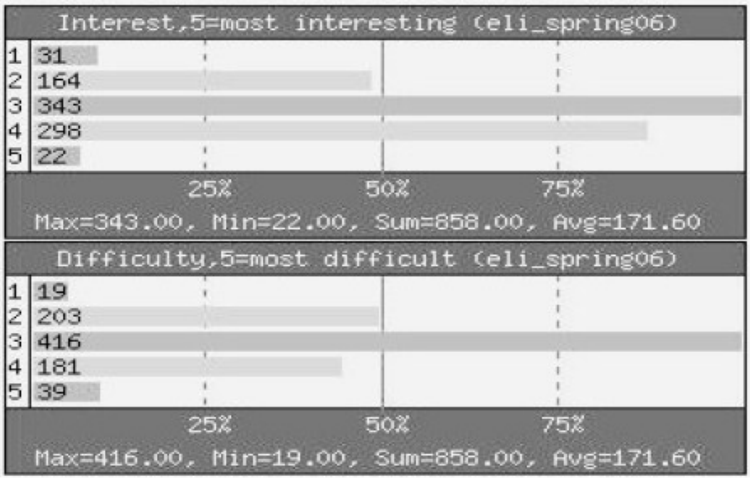


Fig. 5. Post-reading difficulty and interest feedback ratings by students from an early study. (From Heilman et al., 2006).

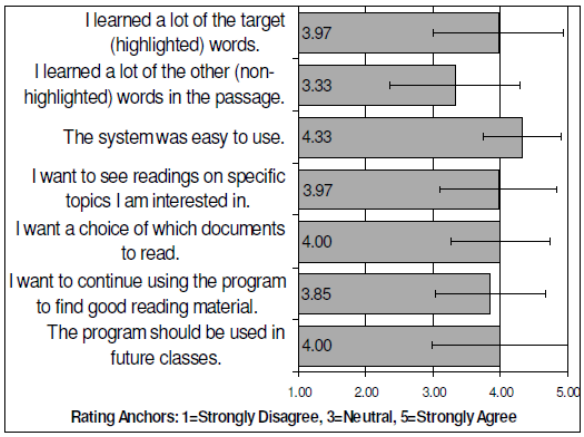


Fig. 6. Early opinions on REAP from an exit survey. (From Heilman et al., 2006).

Heilman et al. (2007) modified the REAP tutor to provide readings that matched student interests, which were gathered using the interest self-assessment survey shown in Figure 7, to see if these personalized readings would increase intrinsic motivation. This was motivated by past research that showed that intrinsic motivation lead to better learning. Heilman et al. found that they were able to classify readings into general topics using fairly simple machine learning techniques, which in turn let them provide readings tailored to a student's topical interests. In particular, they used support vector machine text classifiers with linear kernels trained on webpages from the Open Directory Project (2002), among which a subset of webpages from specific top-level categories (Arts, Business, Computers, Games, Health, Home, Recreation, Science, Society, and Sports), a total of 1000 pages per topic, was manually selected. In addition to the various constraints REAP uses to select readings, such as target words, reading length and difficulty level, Heilman et al. restricted the readings provided to the personalization treatment group of this study to the documents classified with topics that rate highly with the given student.

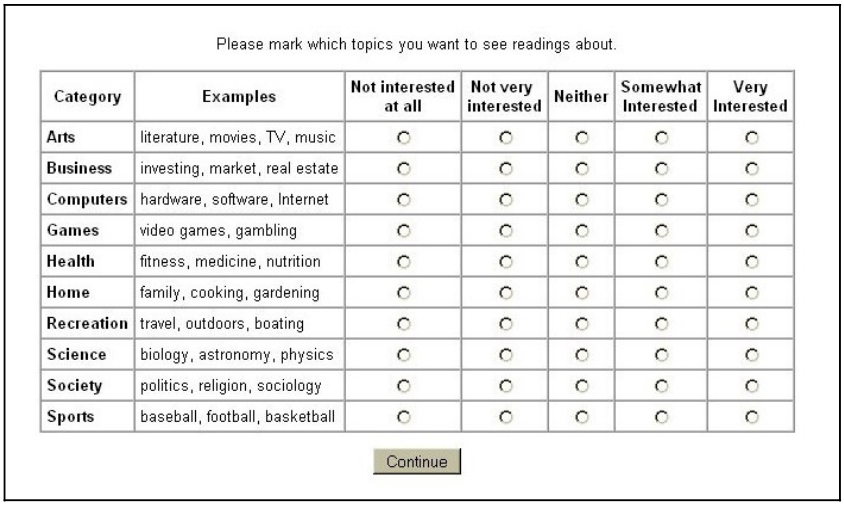


Fig. 7. Reading Interest Survey from REAP.

Interest questionnaires administered after the readings showed that REAP was able to improve interest, since students that were provided with readings personalized by topics were more frequently interested in the readings than students in the control condition, as shown in Table 5. Also, students that were provided with personalized readings performed slightly better than those in the control condition (though not statistically significantly), as made evident by the larger learning gains, with average gains of 35.5% and 27.1% for the treatment and control conditions respectively, illustrated in Figure 8.

Table 5
Breakdown of post-reading interest responses for students using REAP with and without personalized readings. (Adapted from Heilman et al., 2007)

Interest Rating	% of Scores (No Personalization)	% of Scores (Personalization)
1 (least)	6.9	4.1
2	15.0	19.1
3	**41.9**	32.0
4	30.0	**39.7**
5 (most)	6.2	5.2

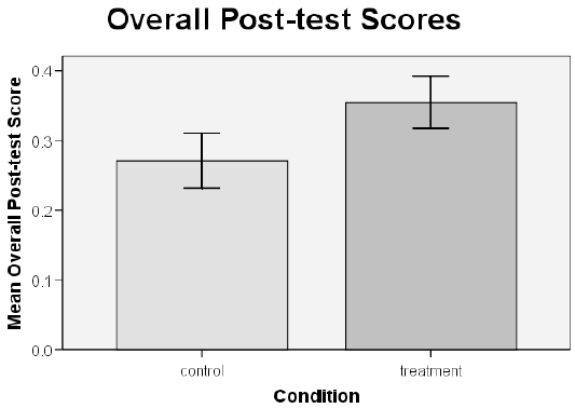


Fig. 8. Average post-test scores by condition, where the error bars show standard error (From Heilman et al., 2007).

MOTIVATION ASSESSMENT

Self-assessments are commonly used in motivation modeling and detection since they are simple to administer and are generally easy to interpret. Some issues that must be considered when using self-assessment for motivation include the types of questions to ask and how to effectively keep track of changes in motivation. In this section we provide a discussion on previous work on the use of self-assessment in motivation, as well as provide a discussion of the key findings of studies conducted in REAP on effective types of motivation questions and what can be done to track the evolution of motivation.

Background

Motivation modeling and detection and its relationship with user behavior have long been an interest of the educational computing community. Self-assessment questionnaires are a very common and straightforward approach to measuring motivation. One commonly used self-assessment instrument is the Motivated Strategies for Learning Questionnaire (MSLQ), which is an 81-item survey designed to measure college students' motivation orientations and their use of various learning strategies (Pintrich et al., 1991).

While self-assessment questionnaires are generally useful in detecting enduring motivational traits, some have been criticized, particularly questionnaires that are administered prior to interaction; a student's motivation is likely to change during an activity and it becomes important to use self-assessment along with other methods to adapt instruction and gather more transient information about their motivation (deVincente & Pain, 1998). With respect to understanding user behavior automatically, Baker (2007) found that machine learning models trained on student activity log data can be used to detect if a student is off task. Centintas et al. (2010) reached a similar conclusion using a regression model personalized for each student. Another study by Baker (2004) showed how a latent response model can be used to determine if a student is 'gaming' a tutoring system in such a way that leads to poor learning.

A study by Roll et al. (2011a) demonstrated that metacognitive models of help seeking could be used by an intelligent tutoring system to assess a student's moment-to-moment learning behavior

unobtrusively, and that metacognitive feedback can help improve a student's behavior in the tutoring system. In a related study, Aleven et al. (2010) described an automatic and unobtrusive method of action-by-action self-regulated learning assessment in a tutoring system, which resulted in lasting improvement in students' help-seeking behavior.

Others have studied methods of improving students' self-assessment skills, since improving metacognitive skills and self-regulation can improve a student's ability to learn independently (Roll et al. 2011b). A study by Roll et al. (2011b) described a tutoring system for self-assessment skills that makes use of established metacognitive principles to help students learn how to evaluate their own abilities.

An important issue in motivation self-assessment is the composition of the questionnaire questions. In the sections below we describe the construction of the motivation questionnaire used in the REAP tutoring system and some key findings we had over the years related to question detail level, frequency of questionnaire administration, and correlations between motivation, learning, and user behaviors.

Motivation Assessment in REAP

Pino et al. (2009) conducted a study with REAP on the use of self-assessment motivation questionnaires to analyze which constructs and specific questions have meaningful relations with learning, as well as analyze motivation over time and the opinion of students about motivation questionnaires. Dela Rosa & Eskenazi (2011b) conducted a follow up study that investigated how the level of detail in motivation questions influenced the effectiveness of these questions to measure motivation and also identified specific user actions that correlated well with the self-assessment questions and student performance on the tasks provided through the tutor. The remainder of this section describes the key findings of these two studies and their implications on motivation assessment and learning in an intelligent tutoring environment.

The initial motivation questionnaire used in REAP, shown in Table 6, was adapted from a survey used in math courses based on the MLSQ. The goal was to construct a questionnaire that had enough questions to cover most motivational constructs, and that was general enough to be reused in other computer-assisted language learning studies. Additionally, the questionnaire was designed to cover the constructs of motivation described by Pintrich & DeGroot (1990): self-efficacy, intrinsic value, test anxiety and self-regulation.

Table 6
Initial Motivation Questionnaire in REAP

Questions
I am sure I understood the ideas in the computer lab sessions.
I am sure I did an excellent job on the tasks assigned for the computer lab sessions.
I prefer work that is challenging so I can learn new things.
I think I will be able to use what I learned in the computer lab sessions in my other classes.
I think that what I learned in the computer lab sessions is useful for me to know.
I asked myself questions to make sure I knew the material I had been studying.

When work was hard I either gave up or studied only the easy parts.
I find that when the teacher was talking I thought of other things and didn't really listen to what was being said.
When I was reading a passage, I stopped once in a while and went over what I had read so far.
I checked that my answers made sense before I said I was done.
I did the computer lab activities carefully.
I found the computer lab activities difficult.

Level of Detail in a Motivation Question and Learning Correlations

In initial studies conducted with the motivation questionnaire described above, Pino et al. (2009) had difficulty finding correlations between the motivation questions and student learning. In fact only a subset of the questions, a few covering the self-regulation construct, showed a significant relationship with learning.

In response to these results, Dela Rosa & Eskenazi (2011b) conducted a study to see how the level of detail of a motivation question impacted the effectiveness of the questionnaires. In this study, the initial survey questions, which consisted of questions that covered many constructs of motivation while being general enough to be reused in studies on other subjects that are taught, were compared with more direct questions, shown in Table 7, which were more explicit items that focused on aspects directly related to the reading activities accomplished over the course of the study.

Table 7
Direct Motivation Questions added to the REAP Questionnaire

Direct Questions
I continued working on the computer lab activities outside the sessions.
I did put a lot of effort into computer lab activities.
I did well on the computer lab activities.
I preferred readings where I could listen to the words in the document.
Learning vocabulary in real documents is a worthwhile activity.

Dela Rosa & Eskenazi (2011b) found that the higher level general questions did not significantly correlate with the various learning measures in their study, consistent with the results by Pino et al. (2009), while the lower level direct questions did, as shown in Figure 9 and described in more detail in the next section. These results imply that general questions may be less effective at predicting the learning outcome of a student than direct motivation questions more closely focused on the tasks performed, and the level of detail of a motivation question should be considered when administering a motivation questionnaire for a study.

Tracking the Evolution of Motivation: Frequency of Questioning & Indirect Indicators of Motivation in REAP

Another interesting result from Pino et al. (2009) involved student attitudes towards the frequency at which motivation questionnaires were administered. While questionnaires are generally useful in

detecting enduring motivational traits, many have been criticized, particularly those administered prior to interaction. Since a student's motivation is likely to change during an interaction, it becomes important to gather more transient information about a student's motivation, often with other methods (de Vincente & Pain, 1998).

As an initial attempt to gather more information about the students' evolving motivation in REAP, Pino et al. administered motivation questionnaires prior to the weekly reading activities in addition to the questionnaires administered during the pre and post tests. Students were not very enthusiastic about answering motivation questions on a regular basis. When asked whether '*the questions at the beginning of the computer lab sessions were too frequent'* during an exit survey, the students average reply was 5 (on a Likert scale of 1 to 7, with higher numbers indicating greater agreement), as shown in Table 8. Furthermore, students' annoyance with the frequent questionnaires became apparent, when students were given an opportunity to give the researchers feedback on the questionnaires, with the following comment typifying their disapproval of the procedure:

> *'Please,take this question off. I already answered this question all most 15 times. I don't want to answer again. It is really uncomfortable to answer the same question a lot of times.'*

Noting that frequent use of motivation questionnaires can be bothersome; Pino et al. (2009) concluded that it may be better to use more automatic measures of motivation that took advantage of user actions that were logged in the REAP tutor.

Table 8
Opinion on motivation questionnaires. (Adapted from Pino et al., 2009)

Question	Average Response (scale 1 to 7)
The questions at the beginning of the computer lab sessions were helpful to me	4.849
The questions at the beginning of the computer lab sessions bothered me	4.019
The questions at the beginning of the computer lab sessions were too frequent	5

Dela Rosa & Eskenazi (2011b) conducted a study that investigated which user actions in the REAP tutor correlated best with motivation and learning. Over the course of the study, user interactions with the tutor were recorded, including the word look up activity using the built-in electronic dictionary, word listening activity using the built in speech synthesis, and the average time spent on activity tasks such as reading and answering questions

Figure 9 shows significant correlations between motivational and learning factors, as well as correlations with explicit (survey questions) and implicit (recorded user actions) indicators of motivation. In the figure, A1-A9 refers to the following recorded user interactions which were hypothesized to indirectly correspond to motivation and possibly learning:

Word lookup activity, using our built-in electronic dictionary

A1: Total number of dictionary lookups
A2: Number of focus words looked up in the dictionary
A3: Number of dictionary lookups involving target words

Words listening activity, using our built-in speech synthesis

A4: Mean number of listens per word
A5: Total number of listens
A6: Number of words listened to

Average time spent on activity tasks

A7: Time spent reading the documents
A8: Time spent on practice questions

With respect to learning factors if figure 9, L1-L4 correspond to learning measures:

L1: Average post-reading practice question accuracy (for all questions appearing directly after reading the documents)
L2: Pre-test to post-test normalized gain
L3: Post-test accuracy
L4: Average difference between pre-test and post-test scores

Lastly with respect to the explicit indicators (survey questions) in figure 9, the columns that start with 'Q' correspond to direct questions given after readings, the columns starting with 'S' correspond to survey questions given during the pre-test and post-tests, the columns ending with '-General' correspond to general high-level survey questions, the columns ending with '-Direct' correspond to direct low-level survey questions, and the letters 'A', 'E', 'V', and 'O' found after the initial 'S' correspond to survey questions that relate to the motivation constructs of *anxiety* (deal with emotional reactions to a task), *self-efficacy* (deal with beliefs about a student's ability to perform a task), *intrinsic value* (deal with goals and beliefs about the importance and interest of a task), and *other*, respectively, where other typically consisted of questions related to self-regulation and learning strategies. A breakdown of each of specific question is shown in tables 9 and 10.

Table 9
Pre-test/Post-test Motivation Survey Questions. (Adapted from Dela Rosa & Eskenazi, 2011b).

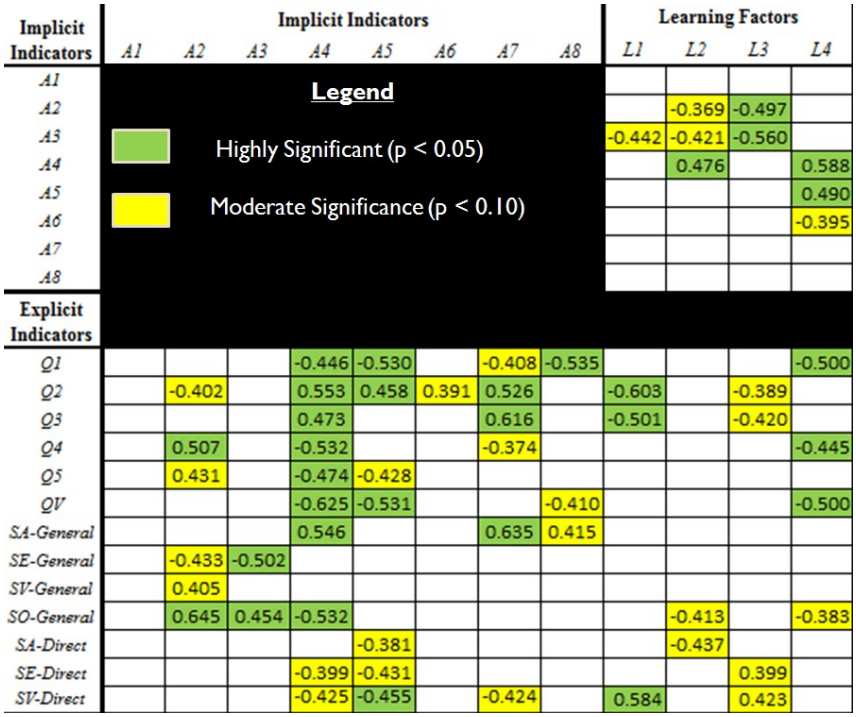
ID	Survey Question Prompt	Group	Type
S1	I am sure I understood the ideas in the computer lab sessions.	General	E
S2	I am sure I did an excellent job on the tasks assigned for the computer lab sessions.	General	E
S3	I prefer work that is challenging so I can learn new things.	General	A
S4	I think I will be able to use what I learned in the computer lab sessions in my other classes.	General	V
S5	I think that what I learned in the computer lab sessions is useful for me to know.	General	V
S6	I asked myself questions to make sure I knew the material I had been studying.	General	O
S7	When work was hard I either gave up or studied only the easy parts.	General	A
S8	I find that when the teacher was talking I thought of other things and didn't really listen to what was being said.	General	A
S9	When I was reading a passage, I stopped once in a while and went over what I had read	Genera	O

	so far.	l	
S10	I checked that my answers made sense before I said I was done.	General	O
S11	I did the computer lab activities carefully.	General	E
S12	I found the computer lab activities difficult.	General	A
S13	I continued working on the computer lab activities outside the sessions.	Direct	A
S14	I did put a lot of effort into computer lab activities.	Direct	A
S15	I did well on the computer lab activities.	Direct	E
S16	I preferred readings where I could listen to the words in the document.	Direct	V
S17	Learning vocabulary in real documents is a worthwhile activity.	Direct	V

Table 10
Post-reading Survey Motivation Questions. (Adapted from Dela Rosa & Eskenazi, 2011b).

ID	Survey Question Prompt	Type
Q1	Did you find the spoken versions of the word helpful while reading this document?	V
Q2	Do you find it easy to learn words when you read them in documents?	E
Q3	Did you find this document interesting?	V
Q4	Did you learn something from this document?	V
Q5	Does reading this document make you want to read more documents?	A

They found that the user interactions related to word listening and dictionary lookup related interactions had significant correlations with learning gains, which implies that these kinds of actions can be used as indirect indicators of motivation that can be helpful in predicting a student's motivational state. Interestingly, the absolute amount of time spent on task did not correlate well with learning, which may suggest that taking into account how the student used their time would be a better factor to consider.



Implicit Indicators	Implicit Indicators								Learning Factors			
	A1	A2	A3	A4	A5	A6	A7	A8	L1	L2	L3	L4
A1												
A2										-0.369	-0.497	
A3									-0.442	-0.421	-0.560	
A4										0.476		0.588
A5												0.490
A6												-0.395
A7												
A8												
Explicit Indicators												
Q1				-0.446	-0.530		-0.408	-0.535				-0.500
Q2		-0.402		0.553	0.458	0.391	0.526		-0.603		-0.389	
Q3				0.473			0.616		-0.501		-0.420	
Q4		0.507		-0.532			-0.374					-0.445
Q5		0.431		-0.474	-0.428							
QV				-0.625	-0.531			-0.410				-0.500
SA-General				0.546			0.635	0.415				
SE-General		-0.433	-0.502									
SV-General		0.405										
SO-General		0.645	0.454	-0.532						-0.413		-0.383
SA-Direct					-0.381					-0.437		
SE-Direct				-0.399	-0.431						0.399	
SV-Direct				-0.425	-0.455		-0.424		0.584		0.423	

Fig. 9. Significant correlations values between motivational & learning factors, and between implicit & explicit motivation indicators. Color signifies level of significance, with green representing strong statistical significance ($p < 0.05$), and yellow representing moderate significance ($p < 0.1$). Note that correlations among implicit indicators and correlations with low significance values were omitted. Also note that Q1-Q5 correspond to the students' average survey response values for all reading activities, and with respect to the implicit indicators, A1-A8 correspond to the students' average values of those indicator values over all reading activities.
(Adapted from Dela Rosa & Eskenazi, 2011b).

CONCLUSION

Self-assessment can be a powerful tool because it is simple and quick, but administrators of these questionnaires must take care while using them as they often need tweaking or support from other evidence or tools depending on the application. In the REAP project, we have used self-assessment to measure and evaluate various aspects of student performance in computer-assisted language learning environment, and tested our hypotheses in the classroom. Through the REAP project, we learned that students can accurately identify vocabulary words they do not know and that one must take caution when they indicate that they know a word through self-assessment. We learned that providing personalized readings can help improve interest in the readings and possibly help boost learning. Lastly, we discussed the importance of motivation questions in detail, the effects of providing questionnaires on a regular basis, and potential indirect indicators of motivation in a language tutor, such as word look up and word listening activity.

ACKNOWLEDGEMENTS

The authors would like to thank past researchers who have been associated with REAP, whose results have been discussed in this paper, such as Michael Heilman, Juan Pino, Alan Juffs, Kevyn Collins-Thompson, and Jamie Callan. This project is supported by the Cognitive Factors Thrust of the Pittsburgh Science of Learning Center, which is funded by the US National Science Foundation under grant number SBE-0836012. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF

REFERENCES

Aleven, V., Roll, I., McLaren, B. M., & Koedinger, K. R. (2010). Automated, Unobtrusive, Action-by-Action Assessment of Self-Regulation During Learning With an Intelligent Tutoring System. *Educational Psychologist*, 45/4, 224-233.

Anand, P. G., & Ross, S. M. (1987). Using Computer-Assisted Instruction to Personalize Arithmetic Materials for Elementary School Children. *Journal of Educational Psychology*, 79/1, 72-78.

Baker, R.S., Corbett, A.T., Koedinger, K.R. (2004). Detecting Student Misuse of Intelligent Tutoring Systems. In: Lester, J.C., Vicari, R.M., Paraguacu, F. (Eds.) *ITS 2004*. LNCS, 3220, 54-76. Springer, Heidelberg.

Baker, R.S. (2007). Modeling and understanding students' off-task behavior in intelligent tutoring systems. ACM *SIGCHI Conference on Human Factors in Computing Systems*, 1059–1068. ACM, New York.

Born, D. G., Gledhill, S. M., & Davis, M. L. (1972). Examination performance in lecture-discussion and personalized instruction courses. *Journal of Applied Behavior Analysis*, 5, 33-43.

Boud, D., & Falchikov, N. (1989). Quantitative Studies of Student Self-Assessment in Higher Education: A Critical Analysis of Findings. *Journal of Higher Education*, 18/5, 529-549.

Brantmeier, C. (2006). Advanced L2 learners and reading placement: Self-assessment, CBT, and subsequent performance. *System*, 34, 15-35.

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

Cetintas, S., Si, L., Xin, Y.P., Hord, C. (2010). Automatic Detection of Off-Task Behaviors in Intelligent Tutoring Systems with Machine Learning Techniques. *IEEE Transactions on Learning Technology*, 3, 228-236.

Cepstral Text-to-Speech. (2001). http://cepstral.com

Chen, C. (2008). Intelligent web-based learning system with personalized learning path guidance. *Educational Computing*, 51/2, 787-814.

Cole, M. J., Zhang, X., Liu, J., Liu, C., Belkin, N. J., Bierig, R., & Gwizdka, J. (2010). Are Self-Assessments Reliable Indicators of Topic Knowledge?. *Annual Meeting of the American Society for Information Science and Technology.*

Collins-Thompson, K., & Callan, J. (2004). Choosing Reading Passages for Vocabulary Learning by Topic to Increase Intrinsic Motivation. *International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Cordova, D., & Lepper, M. (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., Koestner, R., & Ryan, R. M. (1999). A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation. *Psychological Bulletin*, 125/6, 627-668.

Dela Rosa, K., & Eskenazi, M. (2011a). Impact of Word Sense Disambiguation on Ordering Dictionary Definitions in Vocabulary Learning Tutors. *Florida Artificial Intelligence Research Society Conference*.

Dela Rosa, K., & Eskenazi, M. (2011b). Self-Assessment of Motivation: Explicit and Implicit Indicators in L2 Vocabulary Learning. *International Conference on Artificial Intelligence in Education*.

Dela Rosa, K., Parent, G., & Eskenazi, M. (2010). Multimodal learning of words: A study on the use of speech synthesis to reinforce written text in L2 language learning. *ISCA Workshop on Speech and Language Technology in Education*.

de Vincente, A., & Pain, H. (1998). Motivation Diagnosis in Intelligent Tutoring Systems. *International Conference on Intelligent Tutoring Systems*.

Graham, J. (1999). The Reader's Helper: A Personalized Document Reading Environment. *ACM SIGCHI Conference on Human Factors in Computing Systems*.

Gilbert, J. E., & Han, C. Y. (2002). Arthur: A Personalized Instructional System. *Journal of Computing in Higher Education*, 14/1, 113-129.

Harrington, M., & Carey, M. (2009). The on-line Yes/No test as a placement tool. *System*, 37/4, 614-626.

Harris, M. (1997). Self-assessment of language learning in formal settings. *ELT Journal*, 51/1, 12-20.

Heilman, M., Collins-Thompson, K., Callan, J., & Eskenazi, M. (2006). Classroom Success of an Intelligent Tutoring System for Lexical Practice and Reading Comprehension. *International Conference on Spoken Language Processing*.

Heilman, M., & Eskenazi, M. (2006). Authentic, Individualized Practice for English as a Second Language Vocabulary. Presented at *Interfaces of Intelligent Computer-Assisted Language Learning Workshop*. Ohio State University, Columbus, OH [Unpublished]

Heilman, M., & Eskenazi, M. (2008). Self-Assessment in Vocabulary Tutoring. *International Conference on Intelligent Tutoring Systems*.

Heilman, M., Juffs, A., & Eskenazi, M. (2007). Choosing Reading Passages for Vocabulary Learning by Topic to Increase Intrinsic Motivation. *International Conference on Artificial Intelligence in Education*.

Ku, H. Y., & Sullivan, H. (2002). Student Performance and Attitudes Using Personalized Mathematics Instruction. *Educational Technology Research and Development*, 50/1, 21-34.

Kulkarni, A., Heilman, M., Eskenazi, M., & Callan, J. (2008). Word Sense Disambiguation for Vocabulary

Learning. *International Conference on Intelligent Tutoring Systems.*

LeBlanc, R., %Painchaud, G. (1985). Self-assessment as a second language placement instrument. *TESOL Quarterly*, 19/4, 673-687.

Malabonga, V., Kenyon, D. M., & Carpenter, H. (2005). Self-assessment, preparation and response time on a computerized oral proficiency test. *Language Testing*, 22/1, 59-92.

Martin, E., & Rosa, M. C. (2009). Supporting the Development of Mobile Adaptive Learning Environments: A Case Study. *IEEE Transactions on Learning Technologies*, 2/1, 23-36.

Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners. *Canadian Modern Language Review*, 63/1: 127-147.

Open Directory Project. (2002). http://www.dmoz.org

Pino, J., Roll, I., & Eskenazi, M. (2009). Metacognition and Motivation in ESL. Presented at *Pittsburgh Science of Learning Center Advisory Board Meeting*. [Poster]

Pintrich, P.R., & De Groot, E.V. (1990). Motivational and Self-Regulated Learning Components of Classroom Academic Performance. *Journal of Educational Psychology*. 82, 33-40.

Pintrich, P.R., Smith, D.A.R., Garcia, T., & McKeachie, W. (1991). A manual for the use of the motivated strategies for learning questionnaire (MSLQ). *Report, Ann Arbor*.

Roll, I., Aleven, V., McLaren, B., & Koedinger, K. R. (2011a). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction*, 21, 267-280.

Roll, I., Aleven, V., McLaren, B., & Koedinger, K. R. (2011b). Metacognitive Practice Makes Perfect: Improving Students' Self-Assessment Skills with an Intelligent Tutoring System. *International Conference on Artificial Intelligence in Education*.

Schiefele, U. (1991). Interest, Learning, and Motivation. *Educational Psychologist*, 26/3, 299-323.

Walter, E., (Ed.) (2005). *Cambridge Advanced Learner's Dictionary*, 2nd Edition. Cambridge University Press.