Self-Assessment in Vocabulary Tutoring

Michael Heilman and Maxine Eskenazi

Language Technologies Institute, Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
http://www.lti.cs.cmu.edu

Abstract. To individualize instruction, a tutor must infer which knowledge components the student knows and does not know. Self-assessments, by which the student directly reports to the tutor whether a certain item is known, are a fast measure of student knowledge. We investigate their use to initialize a learner model used to individualize instruction ESL vocabulary. Experimental results indicate that self-assessments can be useful measures of knowledge for use in a tutoring system for vocabulary. Self-assessments appear to be particularly reliable when learners claim that words are not known.

Key words: self-assessment, language learning

1 Introduction

Students are typically evaluated by their performance on a given set of tasks. Depending on the tasks, these assessments can be time-consuming and labor-intensive to administer, either for a teacher or a computer. Computer-based instructional systems often face additional challenges in evaluating the correctness of responses. Self-assessment is an alternative method that takes little time to administer and presents less of a challenge for a computer to evaluate. This paper describes a study of self-assessments in the REAP tutor, a system that provides practice on English as a Second Language (ESL) vocabulary.

The REAP tutor [1] facilitates robust learning of target vocabulary words. The system currently focuses on providing intermediate and advanced ESL students with instruction on academic vocabulary words that are important for study at the post-secondary level in English-speaking countries. The tutor helps students to coordinate input from multiple forms of instruction about target words. These include passive implicit instruction in the context available around target words in the reading passages and short examples, passive explicit instruction in the form of dictionary definitions, and interactive instruction in the form of practice exercises.

The student receives individualized and adaptive instruction because REAP matches the difficulty of texts to the reading level of the student [2], and chooses texts that contain target words that he or she has not yet mastered. In a typical instructional cycle, the student starts with a practice reading, in which target words are highlighted to focus attention on them. The student can access dictionary definitions and short examples for any words while reading. After finishing

a reading, the student completes a series of practice exercises for the target words that appeared in the reading. After completing the exercises, the student moves on to the next reading, which usually contains a new set of target words.

1.1 Learner Modeling in REAP

To individualize instruction with respect to which target vocabulary words a particular student knows, the REAP tutor uses a learner model based on the knowledge tracing approach [3]. For each knowledge component the system maintains a Bayesian network in which there are two hidden states. One state corresponds to the knowledge component being known, and the other corresponds to it being unknown. In REAP, the knowledge component is a target word. There is an arc from the unknown state to the known state with a transitional probability, which represents the process of learning that knowledge component. Each hidden state is connected to a observed state that corresponds to the level of performance on a particular task. Observations of performance over time enable inferences about which state the student is in with respect to the particular knowledge component represented by the network.

An important point is that learner models such as knowledge tracing require some form of assessment to provide observations that enable inference about knowledge. In the REAP tutor, evaluations of student performance on the practice exercises following each reading are used to make inferences about the knowledge states for target words. The tutor uses these inferences to select future target words to practice.

In particular, the REAP tutor uses multiple-choice cloze, or fill-in-the-blank, exercises as the post-reading formative assessments. Other exercises, including multiple-choice synonym questions and multiple-choice definition questions have been tried. Sentence production tasks are also used as summative assessment tools, but not for knowledge tracing since they are manually graded by teachers. We claim that cloze exercises are appropriate measures of knowledge since they involve various types of word knowledge. Cloze exercises may simultaneously tap into knowledge of conceptual meaning, grammatical behavior, word associations, and collocations. In contrast, a multiple-choice definition question (e.g., Which of the following is a definition of concrete?) would focus almost exclusively on conceptual meaning. An example cloze question is shown in Figure 1. Further discussion of question types and automatic generation of questions is provided by Brown, Frishkoff, and Eskenazi [4].

1.2 Motivation for Using Self-Assessment

The primary motivation for self-assessments is to reduce the amount of time required for initial assessments-while still maintaining the accurate measures of knowledge needed to individualize vocabulary instruction. In teaching vocabulary, the REAP tutor assesses students at the beginning of a course to identify

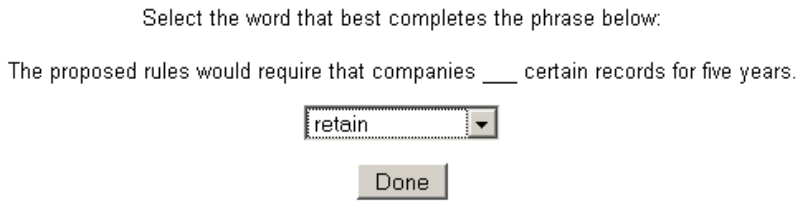

Fig. 1. Example cloze question

which words they most need instruction on. The tutor cannot assume that a particular set of words is known by all students due to the variety of backgrounds, skill levels, lengths of study, and other individual factors among students.

One complicating factor for assessing vocabulary knowledge is that having knowledge of a particular word does not imply knowledge of another word. For instance, if a student knows "lawyer," he or she may not know "doctor." Of course, some words are related and so knowledge of one might provide evidence for knowledge of another. For instance, knowing "flask" might imply knowing "science". However, since the REAP tutor provides instruction on general purpose academic vocabulary words from the Academic Word List [5], such strong connections are rare among the target words in REAP. For example, "alter" and "cite" have little connection that might imply knowledge of one given evidence of knowledge of the other.

Another challenge for efficiently assessing initial knowledge is that there are a very large number of words which might be taught. For example, the Academic Word List used in REAP has 570 head words. Assessing knowledge of every word, or even a substantial fraction of them, with cloze questions or other performance-based methods would take a considerable amount of time which might be better devoted to instruction.

1.3 Previous Research on Self-Assessment

The idea of students being assessed by themselves rather than formal examinations has been a topic of research for a considerable period of time. Boud and Falchikov [6] provide a meta-review of research on self-assessment in higher education. A complex story unfolds, but one salient finding is that more experienced students tend to be more accurate self-assessors. Their findings also suggest that better students are more likely to underestimate their knowledge, while worse students overestimate knowledge.

Self-assessment has also been studied extensively in the intelligent tutoring systems community. Bull, Pain, and Brna [7] developed a system called Mr. Collins which involved the collaborative construction of an open learner model in the domain of Portuguese grammar. That system combined self-assessments with system-administered assessments to create a more accurate learner model than would be available using either source alone.

More recently, Mitrovic and Martin [8] investigated self-assessment in the context of tutors for SQL and database design. As in the review by Boud and Falchikov [6], they found that less able students were worse at self-assessments. Additionally, they found that self-assessments improved when students were able to access the system's estimates of their knowledge through an open learner model. They also viewed self-assessment as an important meta-cognitive skill and learning aid that could be supported by a tutoring system.

Language learning researchers have also examined self-assessments in detail. Ross [9], in a meta-review of the topic, claims that self-assessments are usually consistent with externally assessed variables. The research he discusses dealt mostly with general language skill areas such as speaking or reading, not individual knowledge components. Studies have found that self-assessment of language abilities is very accurate in some areas, but not in others. For instance, Malabonga, Kenyon, and Carpenter [10] found that 92% of students were able to choose an appropriate level of difficulty for an oral proficiency examination. On the other hand, Brantmeier [11] found that self-assessments of reading ability did not accurately predict performance on a subsequent computer-based test of that skill.

The self-assessment of vocabulary knowledge has been investigated as well, particularly in relation to the Yes/No Test developed by Meara [12]. In that test, students are asked to make a binary decision about whether or not they know a word. The test is designed to infer the overall size of a person's receptive vocabulary rather than knowledge of particular words. Mochida and Harrington [13] found that Yes/No test scores were strong predictors of performance on the Vocabulary Levels Test, a commonly used test developed by Nation that employs multiple-choice items [14].

This paper investigates a slightly different application of self-assessments than those discussed above. Rather than dealing with general language skills, it explores whether self-assessments can, in a tutoring system for vocabulary, effectively and efficiently assess knowledge of particular words.

1.4 Research Questions

This paper describes an investigation into the use of self-assessments as initial formative assessments of individual target word knowledge. Evidence from these initial assessments can be used with a learner model to make inferences about which target words a student most needs to receive instruction on. These inferences can then be used to individualize instruction immediately following the initial assessments, rather than adapting gradually during the tutoring process. We investigated whether self-assessments are faster assessments than cloze questions, and also whether the evidence provided by self-assessments and cloze questions is similar. Specifically, the paper addresses the following two questions:

- How long do self-assessments take on average compared to cloze questions, a performance-based measure of knowledge?
- How often do self-assessments agree with cloze question performance?

We hypothesized that self-assessments would be much a faster means of assessment than cloze questions. A moderate but not perfect level of agreement between the two types of assessment was expected.

2 Experimental Design

We conducted an in vivo study with the REAP Tutor to investigate the research questions regarding self-assessments. Forty-two adult English as a Second language students of a variety of native languages participated in the study as part of an upper-intermediate English Reading course in the Summer of 2006 at the University of Pittsburgh's English Language Institute. Although the REAP tutor was used for the duration of the semester, this paper deals only with data recorded during the pre-test of vocabulary knowledge and the subsequent first instance of instruction on particular words. There were ten words on which the tutor pre-tested the students and then later provided instruction. The target words were chosen at random from the set of words from the Academic Word List [5] that were not covered in the students' regular coursework. The words were the following: "acknowledge," "demonstrate," "controversy," "identical," "retain," "precise," "undergo," "sacred," "cease," and "outcome."

During the pre-test, the tutor assessed knowledge of all target words with both cloze and self-assessment questions. The questions were randomly ordered for each student. The self-assessment questions asked students to make a binary decision as to whether or not they knew a given word, as shown in Figure 2. For all cloze questions in this study, students had to choose from a word bank of twenty words that included the correct target word and a set of foils that were other target words used in REAP. The chance that a student would randomly guess the correct answer was thus five percent.

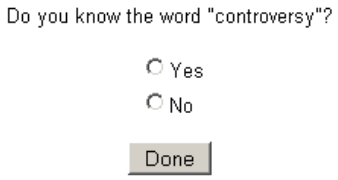

Fig. 2. Example Self-Assessment

The accuracy and reliability of different types of pre-test assessments were evaluated by a follow-up assessment administered just before instruction on the word. These follow-up assessments were cloze questions given as practice exercises. These questions differed from those on the pre-test because they involved different sentences into which the target words would fit. The tutor did not provide any instruction on a given target word between the pre-test and the follow-

up assessment on that word. Also, there were no hints available during these follow-up assessments. However, since they were part of the normal instruction, the tutor did provide feedback about the correctness of responses.

Follow-up assessments were not available for all word-student pairs. Some students spent longer than others on each practice reading. Since the total time on task was kept constant among the students, there was insufficient time for all students to cover all of the target words during the training sessions. As such, 331 self-assessment and follow-up question pairs were available rather than the possible 420. Also, some technical challenges related to recording responses over the web led to a very small portion of the data being unavailable, resulting in the slightly smaller dataset of 329 pairs of pre-test cloze and follow-up questions.

We assume that an agreement of the follow-up assessment with the pre-test assessment is an indication of accuracy of a pre-test assessment, whether self-assessment or cloze. The response time for all assessments was measured by the elapsed time between when the REAP tutor first displayed a question and when the student clicked on a button indicating that he or she had completed the question. The times of these events were recorded automatically by the REAP tutor.

3 Results

3.1 Comparison of Response Times

On the pre-test, students took 38.8 seconds ($N = 329$ questions, $SD = 29.9$) to complete a cloze question on average. In contrast, students spent only 6.1 seconds ($N = 331$, $SD = 5.2$) per self-assessment question on average. If knowledge of 100 target words was assessed, then using self-assessments instead of cloze questions would be expected to save approximately 54 minutes. The difference in mean response times is illustrated in Figure 3, with error bars indicating the sample standard deviations. A two-tailed Welch's t-test for independent samples with unequal variance indicates that the difference is statistically reliable ($p < 0.001$).

Pre-test cloze questions agreed with the follow-up cloze questions 74.8 percent of the time. That is, approximately three quarters of the time, the student either answered both correctly or both incorrectly. In contrast, self-assessments agreed with cloze questions only 56.4 percent of the time, which is barely above random chance.

Interesting patterns emerge, however, when the data are broken down into 2 x 2 contingency tables in order to more closely analyze the relationships between the variables. Table 1 is a contingency table for pre-test cloze and follow-up cloze questions, and Table 2 is a contingency table for pre-test self-assessments and follow-up cloze questions. The cells in each table are the percentages of question pairs in a particular case. Cells that correspond to agreement between question types are in bold typeface.

The difficulty of cloze tasks for the participants in this study is evident in that they correctly answered only 14.8% of pre-test cloze and 20.0% of follow-up

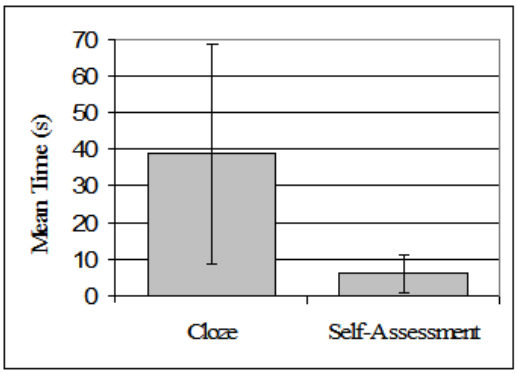

Fig. 3. Mean Response Times for Pre-test Cloze and Self-Assessment Questions

cloze questions, as indicated in the marginal totals in Table 1. Also, an incorrect answer on the pre-test cloze question for a word was a strong predictor of an incorrect answer on the follow-up question, and vice versa. When one of these two questions was incorrectly answered, the other was also incorrectly answered 82% of the time.

The self-assessments appear to be accurate when students claim that a word is unknown. When students claimed to not know a word, they answered follow-up cloze questions incorrectly 91.7% of the time, as indicated by the data in Table 2. When students claimed to know a word, however, they correctly answered the follow-up cloze question for that word only 29.0% of the time. These findings suggest that self-assessment of knowledge of particular vocabulary words are accurate when students claim not to know a word.

Table 1. Contingency Table for Pairs of Pre-test Cloze and Follow-up Questions.

	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.

	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(43.8%)	331(100%)

4 Implementation of Self-Assessments in the REAP Tutor

The findings from the study suggest that self-assessment can be effectively applied in the REAP tutor as an initial assessment of vocabulary knowledge. We modified the tutor in order to utilize self-assessments. The tutor's self-assessment module presents tables of target words to the student, as illustrated in Figure 4. The tutor presents checkboxes, initially unchecked, next to each word. It asks the student to put a checkmark next to all the words that he or she already knows.

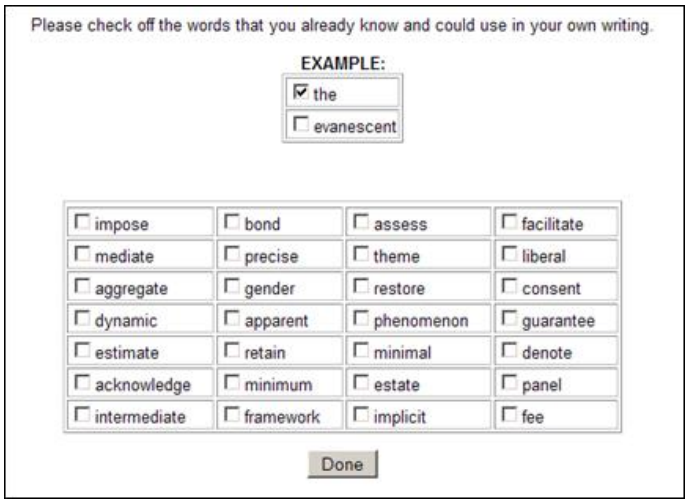

Fig. 4. Screenshot of Self-Assessment Portion of REAP

The tutor uses data from these self-assessments to initialize the learner model. When a student claims that a target word is unknown, the tutor assigns an initial probability of 0.05 for that word being known. When a student claims that a target word is known, the tutor assigns an initial probability of 0.30. These initial probabilities correspond roughly to the proportions of follow-up questions that were correctly answered in the study, as indicated in Table 2.

In other words, when a student claims to not know a word, the tutor assigns a very low probability to him or her knowing that word. However, when a student claims to know a word, the tutor makes less of a commitment by assigning a medium probability. The tutor thus requires further evidence beyond self-assessments before agreeing that the student knows a particular word. It should be noted that while the self-assessments are used to initialize the model, performance on practice cloze questions or other practice tasks is still used to update the learner model and adapt instruction over time.

5 Discussion

The results support several conclusions. Self-assessments of individual target vocabulary words can be much faster than other assessments such as cloze questions. Self-assessments of vocabulary are not, in general, as reliable as cloze questions at predicting future performance. However, self-assessments do appear to be reliable when students claim that they do not know particular words. Thus, it appears that data from self-assessments can be effectively used with a learner model to make inferences about word knowledge in order to individualize instruction at the beginning of a course.

An important question is not conclusively answered by this study: does a fast period of self-assessment and immediately individualized instruction ultimately lead to better learning than having no initial assessment and instruction that gradually adapts to the student?[1] It seems that self-assessments would lead to better learning since they require only a very short amount of time that could otherwise be devoted to instruction. In more recent semesters in which REAP was used, students only spent about 10 minutes on the self-assessments before receiving individualized instruction from the tutor.

The appropriate granularity of self-assessment of vocabulary knowledge is also not certain. In this work, the tutor asked the student to make binary decisions about whether words were known or not. This approach is similar to that employed in the Yes-No test [12]. Self-assessments might also address different levels of knowledge. For example, the tutor might ask questions like, "Do you recognize this word?" or "Could you use this word in a sentence?" Asking more detailed questions might provide more accurate self-assessments. On the other hand, detailed self-assessments might also require additional time, and the tradeoffs between accuracy and time should be considered.

Individual differences among students performing self-assessments may also exist. For instance, prior research has found that novices often overestimate their knowledge and are less accurate at self-assessment. Whether these statements hold for assessing knowledge of individual words-and the possible implications for using self-assessment in a tutoring system for vocabulary-are undetermined.

Currently, the REAP tutor uses self-assessment only in initial assessments, but the tutor could also use self-assessment during instruction for identifying words to review, or co-constructing a learner model as in the Mr. Collins tutor [7]. If students were asked to perform self-assessments during instruction, then it might be useful to train or support their assessments as in the systems described by Mitrovic and Martin [8]. Training could also help students know when to seek help or access dictionary definitions.

[1] One might also compare self-assessments to an initial assessment with feedback, but the researchers claim that an assessment with feedback is similar to instruction that gradually adapts because students spend a considerable portion of the time processing the feedback.

6 Acknowledgements

The authors would like to acknowledge Jamie Callan, Kevyn Collins-Thompson, Jon Brown, James Sanders, Alan Juffs, and Lois Wilson for their work on the REAP project. This work was supported in part by the Dept. of Education through grant R305G03123; the Institute of Education Sciences, U.S. Department of Education, through Grant R305B040063; the National Science Foundation through grant 354420 to the Pittsburgh Science of Learning Center; and a National Science Foundation Graduate Research Fellowship 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

1. Brown, J. and Eskenazi, M.: Student, text and curriculum modeling for reader-specific document retrieval. Proceedings of the IASTED International Conference on Human-Computer Interaction. Phoenix, AZ (2005)
2. Collins-Thompson, K. and Callan, J.: Predicting reading difficulty with statistical reading models. Journal of the American Society for Information Science and Technology (2005)
3. Corbett, A. T. and Anderson, J. R.: Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction, 4(4). Springer Netherlands (1995)
4. Brown, J., Frishkoff, G. and Eskenazi., M.: Automatic question generation for vocabulary assessment. Proceedings of HLT/EMNLP. Vancouver, B.C. (2005)
5. Coxhead, A.: A New Academic Word List. TESOL Quarterly, 34(2), 213-238 (2000)
6. Boud, D. and Falchikov, N.: Quantitative studies of student self-assessment in higher education: a critical analysis of findings. Higher Education, 18(5), 529-549 (1989)
7. Bull, S., Pain, H., and Brna, P.: Mr. Collins: A collaboratively constructed, inspectable student model for intelligent computer assisted language learning. Instructional Science 23, 65-87 (1995)
8. Mitrovic, A. and Martin, B.: Evaluating the Effect of Open Student Models on Self-Assessment. International Journal of Artificial Intelligence in Education. 17, 121-144 (2007)
9. Ross, S.: Self-assessment in second language testing: a meta-analysis and analysis of experiential factors. Language Testing, 15 (1) 1-20 (1998)
10. Malabonga, V., Kenyon, D. M., and Carpenter, H.: Self-assessment, preparation and response time on a computerized oral proficiency test. Language Testing, 22 (1) 59-92 (2005)
11. Brantmeier, C.: Advanced L2 learners and reading placement: Self-assessment, CBT, and subsequent performance. System, 34, 15-35 (2006)
12. Meara, P. and Buxton, B.: An alternative to multiple choice vocabulary tests. Language Testing, 4, 142-45 (1987)
13. Mochida, A. and Meara, P.: The Yes/No test as a measure of receptive vocabulary knowledge. Language Testing, 23 (1) 73-98 (2006)
14. Nation, I.S.P.: Teaching and learning vocabulary. Newbury House (1990)