

IN PRESS --- PLEASE DO NOT QUOTE

**Evaluating Self-Explanations in iSTART: Comparing Word-based and LSA Algorithms**

Danielle S. McNamara, University of Memphis

Chutima Boonthum, Old Dominion University

Irwin Levinstein, Old Dominion University

Keith Millis, Northern Illinois University

Send correspondence to:

Danielle McNamara

[d.mcnamara@mail.psyc.memphis.edu](mailto:d.mcnamara@mail.psyc.memphis.edu)

**Abstract**

This chapter compares the effectiveness of LSA and non-latent word-based algorithms in assessing the quality of self-explanations in iSTART, an automated tutor for improving students' self-explanations of science texts. We have compared various methods of using LSA and word-based methods to guide the responses the tutor makes to the students concerning the quality of their explanations. This chapter examines the effectiveness of 7 systems using word-based algorithms, LSA, and combinations of both, that vary in the degree of manual preparation of the target text required. Their effectiveness is measured in terms of their match to human ratings of the explanations. Our results indicate that the most successful method is a combined system with no hand coding. This fully automated system will make it possible for us to more easily expand iSTART's repertoire to include a wide variety of practice texts.

### **Evaluating Self-Explanations in iSTART: Comparing Word-based and LSA Systems**

Interactive Strategy Training for Active Reading and Thinking (iSTART) is a web-based application that provides young adolescent to college-aged students with self-explanation and reading strategy training (McNamara, Levinstein, & Boonthum, 2004). Although untutored self-explanation -- that is, explaining the meaning of text to oneself -- has been shown to improve text comprehension (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Chi, de Leeuw, Chiu, & LaVancher, 1994), many readers explain text poorly and gain little from the process. iSTART is designed to improve students' ability to self-explain by teaching them to use reading strategies such as comprehension monitoring, making bridging inferences, and elaboration. In the final phase of training, students practice using reading strategies by typing self-explanations of sentences from science texts. The computational challenge is to provide appropriate feedback to the students concerning their self-explanations. To do so requires capturing some sense of both the meaning and quality of the self-explanation. LSA is an important component in that process. Indeed, an important contribution of LSA is that it allows researchers to automatically capture meaning in text (see also, E. Kintsch et al., this volume; Graesser et al., this volume).

Interpreting text is critical for intelligent tutoring systems, such as iSTART, that are designed to interact meaningfully with, and adapt to, the users' input. One question, however, regards the extent to which LSA enables or enhances the accuracy of self-explanation evaluation in iSTART. Thus, in this chapter, we compare various systems of self-explanation evaluation that differ in terms of whether the algorithms are word-based, incorporate LSA, or use a combination of algorithms. Because we want to increase the number of texts available for practice in iSTART, we sought to develop evaluation systems that required less human preparation of the included

texts, so an important characteristic of the systems discussed is the amount of ‘hand-coding’ required. This chapter describes iSTART and our evaluation of these feedback systems.

### **iSTART: The Intervention**

iSTART is modeled after a human-delivered reading strategy intervention called Self-Explanation Reading Training, or SERT (McNamara, 2004; McNamara & Scott, 1999; O’Reilly, Best, & McNamara, 2004). SERT training can be administered to a small group of students in about two hours. It consists of three phases, an introduction with definitions and examples of self-explanation and six reading strategies, a demonstration of the strategies by a student on a videotape, and practice using the strategies with science texts. The six reading strategies include (a) comprehension monitoring, being aware of understanding; (b) paraphrasing, or restating the text in different words; (c) elaboration, using prior knowledge or experiences to understand the text (i.e., domain-specific knowledge based inferences); (d) logic or common sense, using logic to understand the text (i.e., domain-general knowledge based inferences); (e) predictions, predicting what the text will say next; and (f) bridging, understanding the relation between separate sentences of the text. During the introduction phase, a description of the strategy and examples of self-explanations using the strategies are provided. Comprehension monitoring is presented as a strategy that should be used all of the time. Paraphrasing is presented as a basis or jumpstart for self-explanation, but not as means for self-explaining text because it does not go beyond the text. The remaining strategies are various forms of inferences (i.e., domain specific, domain-general, predictive, and bridging) that improve the students’ ability to make sense of difficult text.

Although SERT has been found to be effective both in terms of improving comprehension scores and course performance (McNamara, 2004; McNamara & Scott, 1999;

O'Reilly, Best, & McNamara, 2004), there are scale-up problems in administering SERT with human tutors to students in many classrooms. First, it is expensive to provide and train the tutors. Second, the delivery of training may vary from one class to the next, despite best efforts to maintain consistency. Third, human tutors cannot be made accessible to all students who need it. Finally, the training is delivered to students in groups and therefore cannot be tailored to the individual needs of the learner.

iSTART alleviates these shortcomings while at the same time providing comprehension gains equivalent to SERT (Magliano et al., 2004; O'Reilly, Sinclair, & McNamara, 2004; O'Reilly, Best, & McNamara, 2004). Since iSTART is web-based, it has the potential to make the training available to any school in the country with internet access. Furthermore, as an automated tutoring system, it can deal with students individually, which affords them self-paced learning. Automated tutors can also keep track of students' performance and provide adaptive scaffolding. Web-based systems can also be easily accessed outside the classroom, which potentially increases the time that students can spend on training. Finally, systems that incorporate pedagogical agents and automated linguistic analysis can engage the student in interactive dialog and thereby induce an active learning environment (e.g., Bransford, Brown, & Cocking, 2000; Graesser, Hu, & Person, 2001; Graesser, Hu, & McNamara, in press; Graesser et al., this volume; Louwerse, Graesser, & Olney, 2002).

In iSTART, pedagogical agents instruct trainees in the use of self-explanation and other active reading strategies to comprehend the meaning of text while they read. It consists of three modules, including an introduction, a demonstration, and a practice module. The introduction module presents reading strategy concepts in the form of a classroom discussion among three animated characters (an instructor and two students) that interact with each other, provide

information, pose questions, and give explanations and examples of the reading strategies. These characters are full-body figures with heads that are slightly exaggerated to make mouth movements and facial expressions more visible. The characters speak using a text-to-speech synthesizer. Their speech is audible and appears as text in a speech bubble. They possess a repertoire of gestures and can move about the screen. The interactions between the characters vicariously simulate the active processing necessary to learn the strategies. After each section of the introduction, the students complete brief multiple-choice quizzes to assess their learning of the strategies. The quizzes are designed as learning tools to guide the student to a better understanding of each SERT strategy by providing hints, prompts, and explanations for incorrect choices.

In the demonstration module, one animated character (Merlin) guides the trainees in analyzing the explanations produced by a second animated character (Genie). Genie (representing a student) reads aloud each sentence from a science text and produces a self-explanation that appears both in spoken form and as text in a box on the web page. Merlin continues by asking the student using the program to indicate which strategies Genie employed in producing his self-explanation. Merlin follows up by asking the student to identify and locate in the text box the various reading strategies contained in Genie's self-explanation. Merlin may follow up further by asking the trainee to identify the sentence referenced from the text when Genie uses a bridging strategy. Merlin then gives Genie verbal feedback on the quality of his self-explanation. This feedback mimics the interchanges that the student will encounter in the practice module. For example, sometimes Merlin states that the self-explanation is too short, prompting Genie to add to his self-explanation. Of course, Merlin also gives the trainee feedback, for example, applauding a correct identification.

In the practice module, Merlin plays coach to the trainees and provides feedback to them while they practice self-explanation using the repertoire of reading strategies. The goal is to help the student acquire the skills necessary to integrate knowledge from various sources to understand a target sentence. This knowledge may come from something they have read in the passage (e.g., the previous sentence), general knowledge, or domain knowledge. For each sentence, Merlin reads the sentence aloud before asking the student to self-explain it by typing a self-explanation. Merlin gives feedback, sometimes asking the student to modify unsatisfactory self-explanations. Once the self-explanation is satisfactory, Merlin asks the student to identify what strategies were used and where they were used in the explanation. As in the demonstration module, Merlin sometimes follows up by asking the student to identify an earlier sentence to which the self-explanation referred. Then Merlin provides general feedback. During this phase, the agent's interactions with the trainee are moderated by the quality of the explanation. For example, more enthusiastic feedback is given for longer, more relevant explanations, whereas increased interactions and support are provided for shorter, less relevant explanations.

The feedback that a student receives depends on algorithms that evaluate the characteristics and quality of the student's self-explanation. Clearly, the feedback is more appropriate when the algorithms successfully interpret the student's input. The focus of this chapter is on the algorithms used to evaluate the self-explanations and thus guide feedback during training.

### **iSTART: Description of Feedback Systems**

iSTART was intended from the beginning to employ LSA to determine appropriate feedback. The goal was to develop benchmarks for each of the SERT strategies relative to each of the sentences in the practice texts and to use LSA to measure the similarity of a trainee's

explanation to each of the benchmarks. Each benchmark is simply a collection of words, in this case, words chosen to represent each of the strategies (e.g., words that represent a bridge to a prior sentence). However, while work toward this goal was progressing, we also developed a preliminary “word-based” system to provide feedback in our first version of iSTART (see McNamara et al., 2004). The original word-based system included several hand-coded components. This is referred to as *WBI-Assoc* in Table 1, which stands *for word-based one, with associated words*. For example, for each sentence in the text, the “important words” were identified by a human expert and a length criterion for the explanation was manually estimated. Important words were generally content words that were deemed important to the meaning of the sentence. For each important word, an association list of synonyms and related terms was created. The lists were created by examining dictionaries and existing protocols as well as by human judgments of what words were likely to occur in a self-explanation as associations to each important word. In essence, the association list was meant as a stand-in for LSA until the LSA components were completed.

A trainee’s explanation was analyzed by matching the words in the explanation against the words in the sentence and words in the corresponding association lists. A formula based on the length of the sentence, the length of the explanation, the length criterion, the number of matches to the important words, and the number of matches to the association lists produced a rating of 0 (inadequate), 1 (barely adequate), 2 (good), or 3 (very good) for the explanation.

#### Insert Table 1

The rating of 0 or inadequate was based on a series of filtering criteria that assessed whether the explanation was too short, too similar to the original sentence, or irrelevant. Length was assessed by a ratio of the number of words in the explanation to the number in the target

sentence, taking into consideration the length criterion. Similarity was assessed in terms of a ratio of the sentence and explanation lengths and the number of matching important words. If it was close in length with a high percentage of word overlap, the explanation was deemed too similar to the target sentence. Relevance was assessed from the number of matches to important words in the sentence and words in the association lists. If the explanation failed any of these three criteria, the trainee was given feedback corresponding to the problem and encouraged to revise the self-explanation.

The systems we consider in this chapter also include a metacognitive filter that searches the trainees' explanations for patterns indicating a description of the trainee's mental state such as "now I see ..." or "I don't understand this at all." While the main purpose of the filter is to enable the system to respond to such non-explanatory content when appropriate, we also used the same filter to remove "noise" such as "What this sentence is saying is ..." from the explanation before further processing. We have examined the effectiveness of the systems with and without the filter and found that they all perform slightly better with than without it. Thus, the systems in this chapter all include the metacognitive filter.

This first word-based system required a great deal of human effort per text, primarily because of the need to create an association list for each important word. However, because we envisioned a scaled-up system adaptable to many classrooms, we needed a system that required relatively little manual effort per text. One obstacle was the need to manually identify the important words in each sentence and to create the association lists. Therefore, we replaced the lists of important and associated words with a list of content words (nouns, verbs, adjectives, adverbs) from the sentence and the entire text. This algorithm is referred to as *WBI-TT* in Table 1, which stands for *word-based one, with total text*. The content words were identified using

algorithms from Coh-Metrix, an automated tool that yields various measures of cohesion, readability, other characteristics of language (Graesser, McNamara, Louwerse, & Cai, 2004; McNamara, Louwerse, & Cai, this volume; McNamara, Louwerse, & Graesser, 2002). The iSTART system then compares the words in the self-explanation to the content words from current sentence, prior sentences, and subsequent sentences in the target text, and does a word-based match to determine the number of content words in the self-explanation from each source in the text.

Some hand-coding remained in *WBI-TT* because the length criterion for an explanation was calculated based on the average length of explanations of that sentence collected from a separate pool of participants and on the importance of the sentence according to a manual text analysis. Besides being relatively subjective, this process was time consuming because it required an expert in discourse analysis as well as the collection of self-explanation protocols. Thus, the hand-coded length criteria were replaced with automated criteria based solely on the number of words and content words in the target sentence (i.e., *word-based two, with total text*, or *WB2-TT* in Table 1).

While the first version of iSTART was in use, the first version of the LSA-based system was created (*LSAI* in Table 1). The original goal of identifying particular strategies in an explanation had been replaced with the lesser ambition of rating the explanation as one of three levels (Millis et al., 2004; Millis et al., this volume). The highest level of explanation, called “global-focused,” integrates the sentence material in a deep understanding of the text. A “local-focused” explanation explores the sentence in the context of its immediate predecessors. Finally, a “sentence-focused” explanation goes little beyond paraphrasing. To assess the level of an

explanation, it is compared to four benchmarks or bags of words. The rating is based on a weighted sum of the four LSA cosines between the explanation and the four benchmarks.

The four benchmarks include: 1) the words in the title of the passage, 2) the words in the sentence, 3) prior words or sentences in the text that are causally related to the sentence, and 4) related words from verbal protocols that did not appear in the text. Two of the four benchmarks are created automatically: the words in the title and the words in the current sentence. However, two of the benchmarks require more effort to generate. The prior-text benchmark depends on a causal analysis of the conceptual structure of the text, relating each sentence to previous sentences. This analysis requires both time and expertise. The world-knowledge benchmark consists of words that appeared more than once in the previously collected explanations and did not appear in the other benchmarks.

In sum, *LSA1* required a good deal of manual coding to adapt the algorithm to a particular text. Thus, it too was not suitable for an iSTART program that would be readily adaptable to multiple practice texts. Therefore, we experimented with formulae that would simplify the data gathering requirements. Instead of the four benchmarks mentioned above, we discarded the world knowledge benchmark entirely and replaced the benchmark based on causal-analysis of prior-text with one that consisted of the previous two sentences (*LSA2* in Table 1). We could do this because the texts being considered were from science textbooks. Specifically, we took advantage of the highly linear argumentation in science texts and used the two immediately prior sentences as stand-ins for the set of causally related sentences. It should be noted that this approach may not succeed so well with other genres, such as fictional texts.

Finally, we developed two systems that were combinations of the word-based and LSA approaches (*LSA1/WB2-TT* and *LSA2/WB2-TT* in Table 1). In these combinatory systems, we

combine a weighted sum of the factors used in the fully automated word-based systems and those in LSA1 and LSA2. These combinations allowed us to examine the benefits of using the world knowledge benchmark (in LSA1) when LSA was combined with a fully automated word-based system.

### **iSTART: Evaluation of Feedback Systems**

We describe two experiments conducted to evaluate the performance of the systems. In Experiment 1, we compare the seven systems that vary as a function of approach (word-based, LSA, combination) and coding (manual, automatic) presented in Table 1. They are evaluated by being applied to a database of self-explanation protocols by college students that were evaluated by a human expert on a scale of 0-3. In Experiment 2, we compare the word-based, LSA, and combined systems using a database of explanations by middle-school students.

#### **Experiment 1**

**Self-Explanations.** The self-explanations were collected from college students who were provided with SERT training and then tested with two texts, Thunderstorm and Coal. Both texts consisted of 20 sentences. The Thunderstorm text was self-explained by 36 students and the Coal text was self-explained by 38 students. The self-explanations were coded by an expert according to the following 4-point scale: 0 = vague or irrelevant; 1 = sentence-focused (restatement or paraphrase of the sentence); 2 = local-focused (includes concepts from immediately previous sentences); 3 = global-focused (using prior knowledge).

The coding system was intended to reveal the extent to which the participant elaborated the current sentence. Sentence-focused explanations do not provide any new information beyond the current sentence. Local-focused explanations might include an elaboration of a concept mentioned in the current or immediately prior sentence, but there is no attempt to link the current

sentence to the theme of the text. Self-explanations that linked the sentence to the theme of the text with world knowledge were coded as “global-focused.” Global-focused explanations tend to use multiple reading strategies, and indicate the most active level of processing.

**Results.** Each of the seven systems produces an evaluation comparable to the human ratings on a 4-point scale. Hence, we calculated the correlations and percent agreement between the human and system evaluations (see Table 2). Additionally,  $d$  primes ( $d$ 's) were computed for each strategy type as a measure of how well the system could discriminate among the different strategies. The  $d$ 's were computed from hit and false-alarm rates. A hit would occur if the system assigned the same self-explanation to a category (e.g., global-focused strategy) as the human judges. A false-alarm would occur if the system assigned the self-explanation to a category (e.g., global-focused) different from the human judges (i.e., it was not a global-focused strategy).  $d$ 's are highest to the extent that hits are high and false-alarms are low. In this context,  $d$ 's refer to the correspondence between the human and system in standard deviation units. A  $d'$  of 0 indicates chance performance, whereas greater  $d$ 's indicate greater correspondence.

One thing to note is that there is general improvement according to all of the measures going from left to right. As might be expected, the systems with LSA fared far better than those without LSA, and the combined systems were the most successful. The word-based systems tended to perform worse as the criteria increased (from 0 to 3), but performed relatively well at identifying poor self-explanations and paraphrases. All of the systems, however, identified less successfully the sentence-focused (i.e., 2s) explanations. However, the  $d$ 's for the sentence focused explanations approach 1.0 as LSA is incorporated, particularly when LSA is combined with the word-based algorithms.

Apart from better performance with LSA than without, the performance is also more stable with LSA. Whereas the word-based systems did not perform equally well on the Thunderstorm and Coal texts, there is a high-level of agreement for the LSA-based formulas (i.e., the numbers are virtually identical in the two tables). This indicates that if we were to apply the word-based formulas to yet another text, we have less assurance of finding the same performance, whereas the LSA-based formulas are more likely to replicate across texts.

Insert Table 2

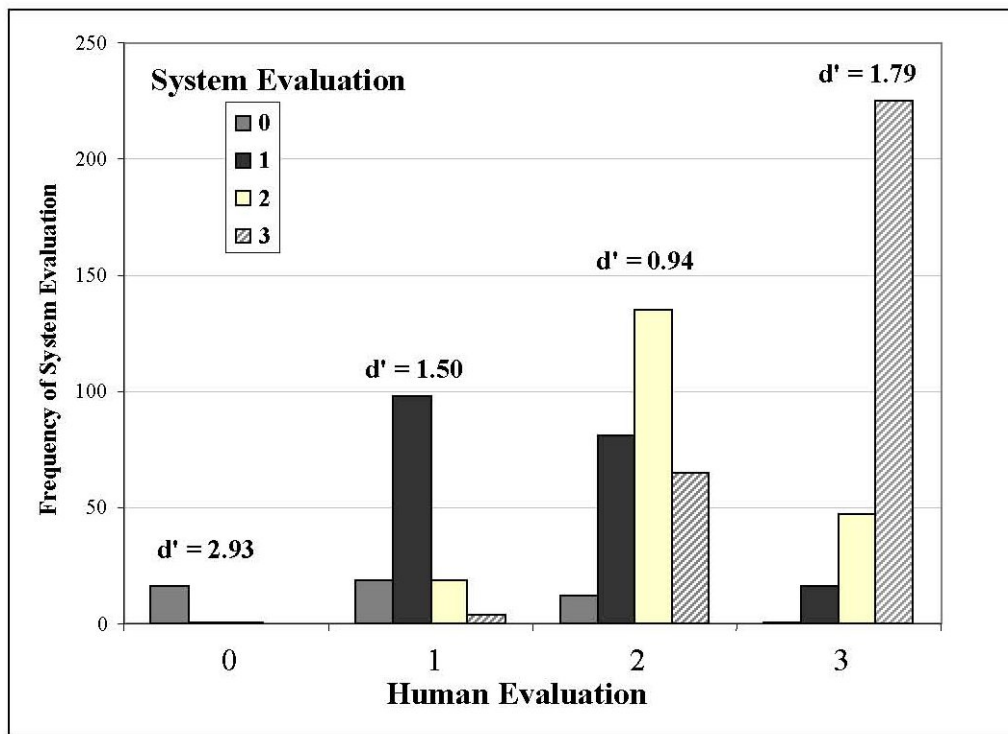


Figure 1

Figure 1 provides a closer look at the data for the combined, automated system, LSA2/WB2-TT. As the  $d'$ s indicated, the system's performance is quite good for explanations that were given human ratings of 0, 1, or 3. Thus, the system successfully identifies poor explanations, paraphrases, and very good explanations. It is less successful for identifying explanations that consist of paraphrases in addition to some information from the previous

sentence or from world knowledge. As one might expect, some are classified as paraphrases and some as global by the system. Although not perfect, we consider this result a success because so few were misclassified as poor explanations.

## **Experiment 2**

**Self-Explanations.** The self-explanations were collected from 45 middle-school students (entering 8<sup>th</sup> and 9<sup>th</sup> grades) who were provided with iSTART training and then tested with two texts, Thunderstorm and Coal. The texts were shortened versions of the texts used in Experiment 1, consisting of 13 and 12 sentences, respectively. This chapter presents only the data from the Thunderstorm text.

The self-explanations from this text were categorized as paraphrases, irrelevant elaborations, text-based elaborations, or knowledge-based elaborations. Paraphrases did not go beyond the meaning of the target sentence. Irrelevant elaborations may have been related to the sentence superficially or tangentially, but were not related to the overall meaning of the text and did not add to the meaning of the text. For example, indicating having seen a thunderstorm would be categorized as an irrelevant elaboration if the visual aspects of thunderstorms were not pertinent to understanding the sentence. Text-based elaborations included bridging inferences that made links to information presented in the text prior to the sentence. Knowledge-based elaborations included the use of prior knowledge to add meaning to the sentence. This latter category is analogous to, but not the same as, the global-focused category in Experiment 1.

**Results.** In contrast to the human coding system used in Experiment 1, the coding system applied to this data was not intended to map directly onto the iSTART evaluation systems. In this case, the codes are categorical and do not necessarily translate to a 0-3 quality range. However, our interest here was to examine which systems best distinguished between the

four types of strategies. One important goal, for example, is to be able to distinguish between paraphrases and elaborations (see also, Millis et al., this volume). If a system makes such a distinction, then the evaluation scores should yield a significant difference between the categories.

We focus our analyses on three systems (WB1-TT, LSA2, and LSA2/WB2-TT) which are essentially the best representatives of the word-based, LSA, and combined systems. A 3 x 4 GLM mixed ANOVA was conducted on the scores for each explanation, including the within-items factor of system (word-based, LSA, combined) and the between-items factor of strategy (paraphrase, irrelevant elaboration, text-based elaboration, knowledge-based elaboration). There was a main effect of system,  $F(2,1072)=5.3$ ,  $MSe=0.301$ ,  $p<.01$  ( $M_{WB}=2.06$ ,  $SE_{WB}=0.05$ ;  $M_{LSA}=2.21$ ,  $SE_{LSA}=0.04$ ;  $M_{CO}=2.08$ ,  $SE_{CO}=0.05$ ), showing significantly higher scores using the LSA system than the word-based or combined systems. A main effect of strategy,  $F(1,536)=70.1$ ,  $MSe=0.285$ ,  $p<.001$ , indicated significant differences in scores as a function of the four categories ( $M_P=1.62$ ,  $SE_P=0.03$ ;  $M_{IE}=2.04$ ,  $SE_{IE}=0.08$ ;  $M_{TE}=2.29$ ,  $SE_{TE}=0.04$ ;  $M_{KE}=2.52$ ,  $SE_{KE}=0.11$ ). Post-hoc Tukey HSD tests indicated significant differences between all categories except text-based and knowledge-based elaborations.

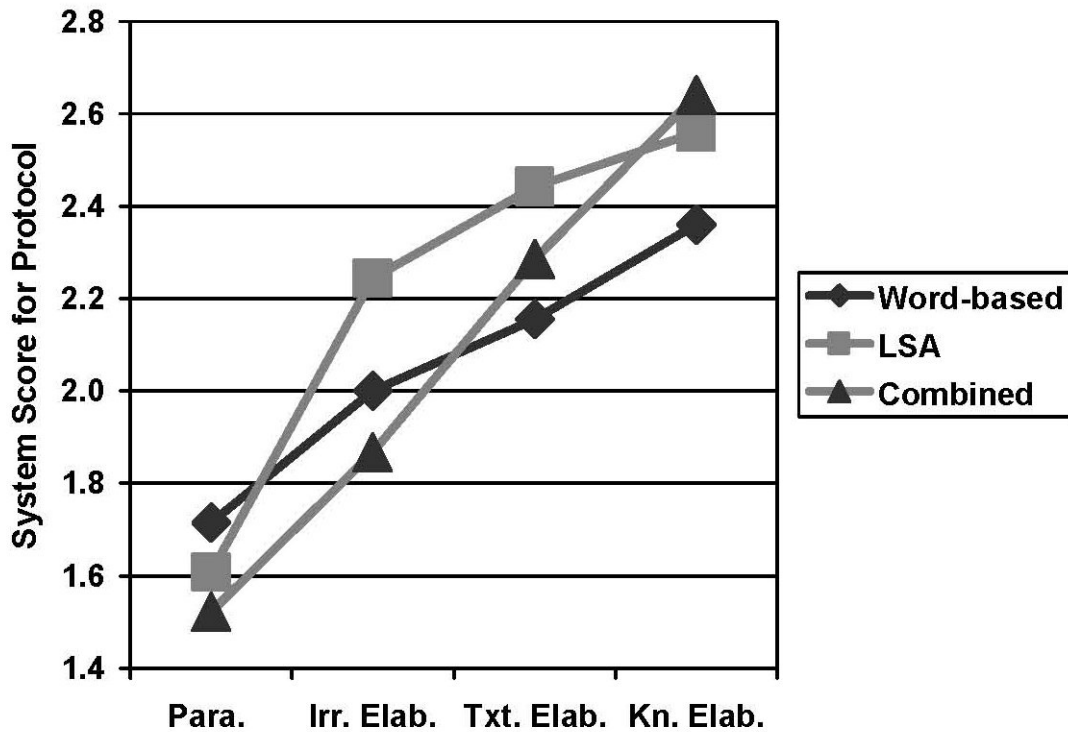


Figure 2

As shown in Figure 2, there was a reliable interaction of system and strategy,  $F(6,1072)=6.9$ ,  $MSe=0.301$ ,  $p<.001$ . Of most interest is observing which system best distinguishes between the four strategies. It is visually apparent that the combined system best accomplishes that goal because it led to the largest difference. To quantify that observation, we calculated the effect sizes (i.e., using Cohen's  $d$ ) of the differences between each strategy for each of the systems, which takes into account both the size of the difference and the amount of variance. The average effect sizes comparing each consecutive category (i.e., the smallest differences) were 0.31 ( $SD=0.10$ ) for the word-based system, 0.45 ( $SD=0.39$ ) for LSA, and 0.48 ( $SD=0.04$ ) for the combined system. Although, LSA showed comparable average effect sizes, there was greater variation in differences than observed for the combined system. The average

effect sizes comparing paraphrases to text-based elaborations and irrelevant elaborations to knowledge-based elaborations were 0.60 (SD=0.06) for the word-based system, 0.87 (SD=0.60) for LSA, and 0.99 (SD=0.01) for the combined system. Finally, comparing the difference between paraphrases and knowledge-based elaborations, the effect size was 1.16 for the word-based system, 1.47 for LSA, and 1.50 for the combined system. Thus, both LSA and the combined system show very good discrimination, but the combined system most consistently discriminated between the strategies. Better discrimination between strategies improves the system's ability to appropriately respond to the student.

### **Discussion**

Our results indicate that the systems that incorporated LSA more accurately classified self-explanations than did the word-based systems, and that the combined systems were most successful. Of these latter two systems, the fully automated system (i.e., LSA2/WB2-TT) performed best. In this system, manual coding, such as choosing associated words or benchmarks, is replaced by computational algorithms. This is indeed fortunate for iSTART because it indicates that a large variety of texts can be adapted to its practice module. Thus, one practical implication is that teachers may more readily specify texts for their students to use when learning how to self-explain and when practicing self-explanation.

One valuable contribution of the current study is that it systematically compared the utility of traditional word-based methods to LSA. Because the combined systems did better than either alone, it appears that both approaches have unique merits. One feature of the word-based system is that it looks for specific words in the self-explanation. If a key word is present, then the system knows that the student had used that word, regardless of the other words present in the self-explanation. On the other hand, an LSA-generated cosine between two texts, or a

benchmark and a self-explanation, will depend on the presence of other words in either one. For example, the cosine between the benchmark “rain” and the word “rain” is 1.00, but the cosine between “rain” and “When the rain is heavy, it falls” is slightly lower at 0.94. This suggests that if the evaluation of a self-explanation depends on the use of a particular word, as it might be at specific text locations, then a word-based approach might be more desirable. Of course, one of LSA’s strength is that it recognizes the latent structure of word meanings. For example, a word-based system looking for the word “rain” would not give any credit to the explanation “when it is heavy, it falls” because the word is absent from the utterance. However, the LSA-generated cosine between “rain” and “when it is heavy, it falls” is 0.61. Although this cosine is lower than when ‘rain’ is present in the explanation, it is well above 0 indicating some confidence that the explanation is addressing the concept of ‘rain.’ Therefore, if an evaluation depends on a cluster of concepts or ideas rather than particular words, then an LSA approach should fare better than a word-based approach.

In summary, both LSA and more traditional word-based algorithms have played important roles to the development of iSTART. For iSTART to effectively teach reading strategies, it must be able to deliver valid feedback on the quality of the self-explanations that a student types during practice. In order to deliver feedback, the system must understand, at least to some extent, what a student is saying in his or her self-explanation. Of course, automating natural language understanding has been extremely challenging, especially for non-restrictive content domains like self-explaining a text in which a student might say one of any number of things. This study, in addition to others in the volume (Graesser et al., this volume; E. Kintsch et al., this volume; Millis et al., this volume; Streeter et al., this volume), fortify the assumption that

LSA can be used to solve problems encountered in natural language understanding, and point to adaptive techniques that improve the meaning-seeking process.

### References

- Bransford, J., Brown, A., & Cocking, R., Eds. (2000). *How people learn: Brain, mind, experience, and school*. Washington, D.C.: National Academy Press. Online at: <http://www.nap.edu/html/howpeople1/>
- Chi, M. T. H., De Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*, 439-477.
- Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, R., & Glaser, R. (1989). Self-explanation: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*, 145-182.
- Graesser, A. C., Penumatsa, P., Ventura, M., Cai, Z., & Hu, X. (this volume). Using LSA in AutoTutor: Learning through mixed-initiative dialogue in natural language.
- Graesser, A. C., Hu, X., & McNamara, D. S. (in press). Computerized learning environments that incorporate research in discourse psychology, cognitive science, and computational linguistics. In A. F. Healy (Ed.), *Experimental Cognitive Psychology and its Applications: Festschrift in Honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer*. Washington, D.C.: American Psychological Association.
- Graesser, A. C., Hu, X., & Person, N. (2001). Teaching with the help of talking heads. In T. Okamoto, R. Hartley, Kinshuk, J. P. Klus (Eds.), *Proceedings IEEE International Conference on Advanced Learning Technology: Issues, Achievements and Challenges* (460-461).
- Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers*, *36*, 193-202.

- Kintsch, E., Caccamise, D., Dooley, S., Franzke, M., & Johnson, N. (this volume). Summary street: LSA-based software for comprehension and writing.
- Louwerse, M. M., Graesser, A. C., Olney, A., & the Tutoring Research Group. (2002). Good computational manners: Mixed-initiative dialog in conversational agents. In C. Miller (Ed.), *Etiquette for Human-Computer Work, Papers from the 2002 Fall Symposium, Technical Report FS-02-02*, 71-76.
- Magliano, J. P., Todaro, S., Millis, K. K., Wiemer-Hastings, K., Kim, H. J., & McNamara, D. S. (2004). *Changes in reading strategies as a function of reading training: A comparison of live and computerized training*. Submitted for publication.
- McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes*, 38, 1-30.
- McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy training for active reading and thinking. *Behavior Research Methods, Instruments, & Computers*, 36, 222-233.
- McNamara, D. S., Louwerse, M. M., & Graesser, A. C. (2002). *Coh-Matrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension*. Technical report, Institute for Intelligent Systems, University of Memphis, Memphis, TN.
- McNamara, D. S., & Scott, J. L. (1999). Training reading strategies. In M. Hahn & S. C. Stoness (Eds.), *Proceedings of the Twenty-first Annual Meeting of the Cognitive Science Society* (pp. 387-392). Hillsdale, NJ: Erlbaum.
- Millis, K. K., Kim, H. J., Todaro, S., Magliano, J. P., Wiemer-Hastings, K., & McNamara, D. S. (2004). Identifying reading strategies using latent semantic analysis: Comparing

semantic benchmarks. *Behavior Research Methods, Instruments, & Computers*, 36, 213–221.

Millis, K. K., Magliano, J. P., Wiemer-Hastings, K., Todaro, S., & McNamara, D. S. (this volume). *Assessing comprehension with Latent Semantic Analysis*.

O'Reilly, T., Best, R., & McNamara, D. S. (2004). Self-Explanation reading training: Effects for low-knowledge readers. In K. Forbus, D. Gentner, T. Regier (Eds.), *Proceedings of the Twenty-sixth Annual Meeting of the Cognitive Science Society* (pp. 1053-1058). Mahwah, NJ: Erlbaum.

O'Reilly, T., Sinclair, G. P., & McNamara, D. S. (2004). Reading strategy training: Automated verses live. In K. Forbus, D. Gentner, T. Regier (Eds.), *Proceedings of the Twenty-sixth Annual Meeting of the Cognitive Science Society* (pp. 1059-1064). Mahwah, NJ: Erlbaum.

Streeter, L., Lochbaum, K., Psotka, J., & LaVoie, N. (this volume). Automated tools for collaborative learning environments.

### **Acknowledgements**

This project was supported by NSF (IERI Award number: 0241144). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF.

Table 1.

Descriptions of evaluation systems in terms of whether word-based or LSA-based factors are included and the amount of hand-coding used.

		<b>Word-Based</b>			
		-no word-based factors -	Hand-coded Important Words, Association List, Length Criteria	Automated Selection of Content Words, Total Text Replaces Association List	
				Hand-coded Length Criteria	Automated Length Criteria
<b>LSA-Based</b>	- no LSA-based factors -	-	<b>WB1-Assoc</b>	<b>WB1-TT</b>	<b>WB2-TT</b>
	Hand-coded Benchmarks	<b>LSA1</b>	-	-	<b>LSA1/WB2-TT</b>
	No Hand-coded Benchmarks	<b>LSA2</b>	-	-	<b>LSA2/WB2-TT</b>

Table 2.

Measures of agreement for the Thunderstorm and Coal texts between the seven system evaluations and the human ratings of the self-explanations.

<b>Thunderstorm Text</b>	<b>WB1-Assoc</b>	<b>WB1-TT</b>	<b>WB2-TT</b>	<b>LSA1</b>	<b>LSA2</b>	<b>LSA1/ WB2-TT</b>	<b>LSA2/ WB2-TT</b>
<b>Correlation</b>	0.47	0.52	0.43	0.60	0.61	0.65	0.64
<b>% Agreement</b>	48%	50%	27%	55%	57%	60%	62%
<b>d' of 0's</b>	2.21	2.26	0.97	2.13	2.19	2.00	2.21
<b>d' of 1's</b>	0.84	0.79	0.66	1.32	1.44	1.53	1.45
<b>d' of 2's</b>	0.23	0.36	-0.43	0.47	0.59	0.76	0.85
<b>d' of 3's</b>	1.38	1.52	1.41	1.46	1.48	1.53	1.65
<b>Avg d'</b>	1.17	1.23	0.65	1.34	1.43	1.46	1.54

<b>Coal Text</b>	<b>WB1-Assoc</b>	<b>WB1-TT</b>	<b>WB2-TT</b>	<b>LSA1</b>	<b>LSA2</b>	<b>LSA1/ WB2-TT</b>	<b>LSA2/ WB2-TT</b>
<b>Correlation</b>	0.51	0.47	0.41	0.66	0.67	0.70	0.71
<b>% Agreement</b>	41%	41%	29%	56%	57%	61%	64%
<b>d' of 0's</b>	4.67	4.73	1.65	2.52	2.99	2.52	2.93
<b>d' of 1's</b>	1.06	0.89	0.96	1.21	1.29	1.55	1.50
<b>d' of 2's</b>	0.09	0.13	-0.37	0.45	0.49	0.83	0.94
<b>d' of 3's</b>	-0.16	1.15	1.28	1.59	1.59	1.73	1.79
<b>Avg d'</b>	1.42	1.73	0.88	1.44	1.59	1.66	1.79

## Figure List

*Figure 1.* Correspondence between human evaluations of the self-explanations and the combined (LSA and word-based), automated system (i.e., LSA2/WB2-TT). Explanations were evaluated by humans as vague or irrelevant (0), sentence-focused (1), local-focused (2), or global (3).

*Figure 2.* Comparison of the scores generated by three systems (word-Based, LSA, and combined) as a function of experts' classification of the self-explanation.