

IN PRESS ---- PLEASE DO NOT QUOTE

Assessing and Improving Comprehension with Latent Semantic Analysis

Keith Millis

Joseph Magliano

Katja Wiemer-Hastings

Stacey Todaro

Northern Illinois University

Danielle S. McNamara

University of Memphis

Send correspondence to:

Keith Millis, kmillis@niu.edu

Dept of Psychology

Northern Illinois University

Dekalb, IL 60115

Abstract

We present research on using LSA to measure comprehension ability and to improve comprehension strategies. The research is based on the premise that LSA provides a measure of semantic similarity of think-aloud protocols and semantic benchmarks – groups of words that together represent different comprehension strategies. The think-aloud protocols (or in some cases, written protocols) are obtained after participants read particular sentences. The cosines between the protocols and the benchmarks have been shown to predict reading comprehension of both narrative and expository texts. Studies are described that examined the utility of different types of benchmarks and the use of cosines to deliver feedback on the quality of self-explanations. Our results indicate that LSA can be used successfully to predict comprehension, assess overall reading strategies, and possibly to improve reading comprehension.

Assessing and Improving Comprehension with Latent Semantic Analysis

Assessing and improving reading comprehension are crucial goals for educators and psychologists alike. Students across the nation have great difficulty in comprehending what they read. It is therefore important that educators are able to identify students at risk for failing, and to have some procedure available to help students improve their comprehension. Assessment and intervention go hand in hand. To know whether a student has trouble understanding a passage, one must identify whether the student had successfully performed any number of processes related to reading (e.g., word recognition, inferencing, making connections between the current and prior sentences, etc.). And to improve a student's comprehension, it is likely that the intervention address at least a subset of those same reading processes.

In this chapter, we summarize some research that we have conducted which uses LSA to assess reading comprehension. We also discuss research that we have done which uses LSA to give feedback to readers as they read with the goal that they become better readers.

Assessing Comprehension

Currently, educators usually rely on multiple-choice tests to assess their students' comprehension. In multiple-choice tests, the student reads a passage before answering multiple-choice questions about that passage. This time-honored tradition offers several advantages: easy administration and scoring, low

costs, and high reliability. Despite the advantages, there are certain limitations to multiple-choice tests. One is that they often do not accurately tap the cognitive representations and mechanisms that underlie learning. For example, to understand implicit relations in connected discourse, the reader must generate bridging inferences that conceptually link explicit ideas. However, most multiple-choice tests provide only a summary statistic of overall comprehension, rather than scores reflecting bridging ability or other component processes of comprehension (except perhaps, a vocabulary score). More importantly, multiple choice tests measure comprehension after the student reads the text.

Consequently, the tests are limited in accurately measuring inferential activities and reading strategies that occur during reading (Carver, 1992; Farr, Pritchard, & Smitten, 1990; Hanna & Oaseter, 1980; Katz, Lautenschlager, Blackburn, & Harris, 1990). A third limitation is that these tests tap surface knowledge (e.g., particular words or phrases) gained from a text rather than the deep understanding that leads to long-term learning (e.g., Shapiro & McNamara, 2000). Indeed, students often adopt special test taking strategies, such as reading the questions first before searching the text for answers (Farr et al., 1990). As a consequence, the student may answer the question correctly, yet fail to understand the text at a deep level.

Our research makes use of Latent Semantic Analysis (LSA) to address three limitations of standard post-reading comprehension measures. First, LSA

allows on-line comprehension assessment; second, the measures are easily and automatically calculated; and third, the comprehension measures reveal deep comprehension processes (Magliano & Millis, 2003; Shapiro & McNamara, 2000). In our research, we have readers tell us what they are thinking about while they read a text. These thoughts are called verbal protocols. Verbal protocols together with LSA provide an alternative to standard measures such as multiple-choice tests for assessing reading comprehension. This alternative arises from the assumptions that (1) the verbal protocols produced while reading a sentence (or immediately thereafter) captures the thoughts of the reader during (and not after) comprehension, (2) the protocols reflect the nature of the person's understanding, and (3) LSA can classify the protocols on understanding by comparing the protocols to texts representing different types or levels of understanding.

Although the use of verbal protocols has led to lively debates within experimental psychology, (cf., Ericsson & Simon, 1993; Nisbett & Wilson, 1977) verbal protocols are now commonly assumed to reveal thoughts in short-term memory that are codeable in language (Ericsson & Simon, 1993; Pressley & Afferbach, 1995). Using LSA to classify verbal protocols on the extent to which the reader understands the text is advantageous because verbal protocols are tedious and time consuming to hand-code. Moreover, expert knowledge is needed to identify and classify the units of language of interest.

Overview of Approach

In the studies to be reported in this chapter, participants read texts one sentence at a time on a computer screen. After specific sentences, they are asked to either say aloud or type into the computer their thoughts regarding their “understanding of the sentence in regard to the text as a whole.” Admittedly, the instruction is a bit vague, but we did not want to affect their natural responses by giving detailed examples. They are also told that there is no right or wrong answer. Their response comprises the verbal protocol for that sentence. The computer then computes the semantic similarity between the protocol and several “semantic benchmarks” using LSA. Semantic benchmarks are fixed texts associated with that sentence that convey different types of reading strategies that might be exhibited at that sentence (Millis, Magliano, Wiemer-Hastings, & McNamara, 2001). They are functionally similar to “ideal answers” to questions in Graesser’s AutoTutor (see Graesser et al., this volume) in that the magnitude of the LSA cosines between the verbal protocol and each of the benchmarks are taken to be indicators of conceptual understanding.

There are various constraints upon which the architecture of this approach can be successful. One is the extent to which the content of the semantic benchmarks captures a person’s level of understanding. This is quite challenging because readers can say a wide variety of things to any given sentence. They might restate the sentence, elaborate the sentence, combine ideas across

sentences, reveal a personal anecdote, type in a meta-cognitive statement (e.g., “Oh, I thought so”), or exhibit a combination strategies. In addition, the success of using our LSA-based system to measure comprehension depends on the assumption that skilled and less-skilled comprehenders will produce recognizably different content in their protocols. For example, it would be ideal, if for a given sentence, skilled comprehenders always say X and less-skilled comprehenders always say Y, and that X and Y are semantically dissimilar from one another.

Theories of discourse comprehension have posited two assumptions that have guided us on what to look for in the verbal protocols that would discriminate between skilled and less-skilled comprehenders. One is that better readers are more likely than poor readers to reactivate text that is causally-related to the current sentence. That is, better readers actively construct a text representation that is guided by the causal structure of the text. When they read a sentence, they attach the ideas from the current sentence to causal antecedents in the prior text representation (Graesser, Singer, & Trabasso, 1994). The second is that better readers are more likely than poor readers to establish inferences based upon their world knowledge which link the current text representation to the theme of the text (Trabasso & Magliano, 1996). Better readers use their world knowledge to actively understand the text.

We have used these theoretical considerations to guide the construction of the semantic benchmarks. In most of the research to be described in this chapter

(except when otherwise noted), we compared verbal protocols to three semantic benchmarks: the current sentence, prior causally-relevant sentences, and world knowledge. The current sentence benchmark contained content words from the current sentence: nouns, verbs, adjectives, and adverbs. The prior causally relevant sentence benchmark contained the content words from sentences that were identified by a hand-conducted causal analysis to be causal antecedents to the current sentence (see Trabasso, van den Broek, & Suh, 1989). Theoretically, these benchmarks represent the ideas that skilled readers should be reactivating as they read the current sentence. The world knowledge benchmarks represented elaborations associated with each sentence. The words comprising this benchmark were empirically derived from a question-answering task and were not in either of the other two benchmarks. Although relevant knowledge could be identified by LSA from similar passages in a background corpus, we have not as yet done so.

Can LSA identify the source of the information contained in a verbal protocol? That is, can the cosines between a verbal protocol and the benchmarks reveal whether the verbal protocol contained information from the current sentence or from causal antecedents? A study by Magliano, Wiemer-Hastings, Millis, Muñoz, and McNamara (2002) addressed this question by having students supply self-explanations after reading sentences of expository texts suitable for college freshman. Self-explanations (McNamara et al., this volume; McNamara,

2004; McNamara, Levinstein, & Boonthum, 2004) are verbal protocols that express an attempt by the reader to explain the current sentence in the context of the passage by using various reading strategies, such as paraphrasing, bridging, elaborating, predicting, using logic and common sense. The self-explanations from Magliano et al. (2002) were coded on the extent that they elaborated the current sentence, revealing the use of the strategies. Self-explanations that merely restated or paraphrase the current sentence were coded as “sentence-focused.” Sentence-focused explanations do not provide any new information beyond the current sentence. Self-explanations that included a concept from an immediately prior text sentence were coded as “local-focused.” 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. Table 1 gives examples of self-explanations to sentence 13 of the text “Stages of Thunderstorm Development” which is presented in the Appendix, the benchmarks associated with that sentence, and the LSA cosines between them.

Table 1

As expected, Magliano et al. (2002) found a correspondence between the self-explanations and the source of information contained in them. For example,

the left-hand side of Figure 1 displays the proportion of clauses within each type of self-explanation as classified by human judges. It also displays the source of the clauses in terms of whether the main content of the clause came from the current sentence or prior text and/or world knowledge. Explanations classified as sentence-focused consisted almost exclusively of content from the current sentence, and very little from prior text or world knowledge. In contrast, global-focused self-explanations contained more content from prior text and/or world knowledge than from the current sentence. But more importantly, the same pattern emerged on the cosines that were generated between the self-explanations and the semantic benchmarks (see the right-hand side of Figure 1). These data suggest that LSA can “identify” the source of information in verbal protocols, enabling it to serve as a proxy for human raters.

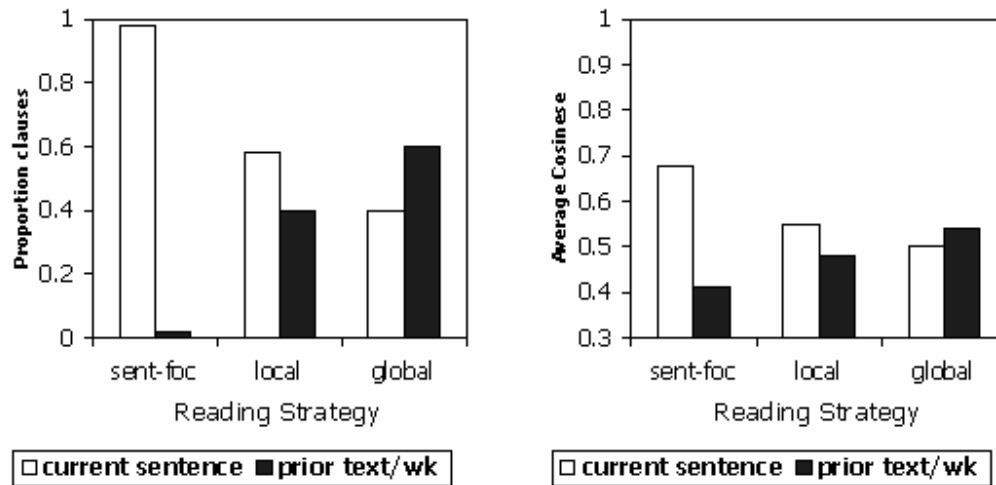


Figure 1

Do LSA Cosines Predict Comprehension?

Our LSA-based approach to measuring comprehension has produced very encouraging results. Before we describe these results, let us first outline what we mean by “comprehension.” The goal of comprehension is to generate a coherent representation of the discourse, and in particular, the situation and events depicted in the discourse. Discourse psychologists refer to the representation of the situations as the “situation model” of the text. Although the situation model is based on the explicit ideas mentioned in the discourse, it includes reader-generated inferences about temporal, causal, and spatial attributes of the entities, protagonists and events mentioned in the text. The situation model is

incrementally built across time, as each clause is read and as inferences pertaining to that clause are generated. In narratives, the situation model includes the location and goals of the protagonists. In expository texts, the situation model would ideally include the ideas underlying the exposition – what the author is describing. It is clearly out of our scope to use LSA to account for all of these processes. Our goal has been more restricted, namely to assess the extent to which the reader is focusing on (1) information from the text, and (2) world knowledge that is directly relevant to the ideas expressed by the current sentence.

Narrative text. Magliano and Millis (2003) tested whether LSA could be used to assess the comprehension of simple narrative text. In their study, college students “thought aloud” to selected sentences in very simple fairy tales, and read others silently. The students were grouped into skilled and less-skilled comprehenders based on their performance on the Nelson Denny test of comprehension. To measure how well they comprehended the stories, students in Study 1 answered questions about them and students in Study 2 recalled them. For each student, mean cosines were computed between each verbal protocol and the current sentence (i.e., the sentence to which the participants’ had thought aloud), and the prior causally relevant sentences (as determined by a hand-coded causal analysis). In this case, the text sentences served as semantic benchmarks.

The LSA cosines between the verbal protocols and the benchmarks predicted measures of comprehension when submitted to regression analyses.

Three predictors were included in the equations, including two LSA benchmarks (LSA current sentence, LSA prior causal sentence) and Nelson Denny scores. There were three regression equations, one predicting performance on the True-False questions, one predicting recall from the passages to which the participants thought aloud, and one predicting recall from the passages that were read silently. The standardized coefficients (beta weights) are shown in Table 2. The LSA cosines were predictive in all three equations. The negative slope for the LSA cosine for the current sentence indicates that greater similarity between the verbal protocols and the current sentence is associated with decreased comprehension. In contrast, the positive direction of the prior causal sentence predictor indicates that comprehension increased when the verbal protocols were similar to causally related antecedent sentences. Performance on the Nelson Denny test contributed significantly only to the prediction of recall of the silently read passages.

Table 2

There are three noteworthy points to these results. First, the direction of the LSA beta weights is consistent with theories of comprehension. Namely, a coherent representation is achieved when the reader attempts to conceptually link the current sentence with the causal structure of the text (Trabasso, van den Broek, & Suh, 1989). Second, the finding that the LSA cosines predicted recall for stories that were silently read indicates the cosines generalize to narrative text read in a more externally valid setting. That is, few competent readers externalize

their thoughts as they read a story. Generalization is especially important for a future comprehension test using this procedure. Third, the LSA-based cosines were more predictive of the comprehension measures than was the Nelson-Denny, thus highlighting the utility of this procedure to uncover comprehension processes.

Expository text. To test whether LSA could account for the comprehension of more difficult expository text, we had 75 undergraduate psychology students attending Northern Illinois University read one of two expository texts on the computer, and after reading each sentence, type in their thoughts at that moment in time. One passage described the origins of coal, and the other described medical problems associated with heart disease (modified from McNamara, Kintsch, Songer, & Kintsch, 1996). After completing both passages, they were given a comprehension test of the material. The comprehension test consisted of eight short answer questions. One-half of the questions were constructed to measure the explicit ideas mentioned in the text, whereas the other half were constructed to measure inferences thought to be included in the situation model. The score for each answer was the percentage of propositions that human raters had judged to be matches with the propositions needed for an ideal answer. The percentages for each question type were summed, creating a score for the textbase and for the situation model. The score for the textbase and situation model questions were 2.31 (SD = .77) and 1.46 (SD

= .61), respectively. The bi-variate correlation between the textbase and situation model scores was .43, indicating a shared variance of 18%. Students also took the Nelson Denny test of comprehension.

To examine whether the cosines predicted readers' comprehension of expository text, the cosines between what they had typed in after each sentence and the two benchmarks (current sentence, prior causal text), and their Nelson-Denny scores were entered into three regression equations, one for each type of comprehension score (textbase, situation model, total). The prior causal text contained content words from causal antecedent sentences that had to be at least two sentences prior to the current sentence. Thus, the prior text benchmark words represented "distal causes." The standardized regression coefficients (beta weights) are shown in Table 3. The results were very similar across the three equations and to the results obtained with narrative texts. The cosines for the distal causes were consistently significant. The positive slope for this variable indicates that the more semantically similar participants' self-explanations were to the distal causes for the current sentence, the more questions they answered correctly. The slope for the current sentence benchmark was statistically significant for performance on the textbase questions and the overall score. The negative slope for this variable indicated that the more semantically similar the self-explanation was to the current sentence, the more poorly participants did in answering the textbase questions, but not the situation model questions. This

result suggests that performance on the textbase questions was more closely tied to the reader's focus on the current sentence rather than on distal causes.

Table 3

Does the Type of Benchmark Matter?

The LSA benchmarks thus far discussed have represented information from current and prior causal sentences. They were chosen to capture the extent to which the reader links the current sentence representation to causal antecedents mentioned in the prior text. However, verbal protocols can reveal and be categorized on other processes as well. One is general reading strategies conveyed by self-explanations that were briefly described earlier in this chapter, namely sentence-focused, local, and global strategies (see Magliano et al. 2002; McNamara et al., this volume). These general reading strategies represent the extent to which the reader actively elaborates the current sentence with the prior text, the theme of the text, and world knowledge. Another process is to identify the use of more specific reading strategies, such as paraphrasing, bridging, and generating elaborations. Paraphrasing occurs when the self-explanation contains a restatement of the current sentence. Bridging occurs when the self-explanation contains a connection between the current sentence and a prior sentence. An elaboration occurs when the self-explanation contains an idea that was not mentioned in the text.

We have recently compared three different types of benchmarks on their utility for classifying verbal protocols on the general and specific strategies that they may reveal (Millis, Kim, Todaro, Magliano, Wiemer-Hastings, & McNamara, 2004). The first, we refer to as content words; these are the benchmarks that we have discussed earlier representing three sources of information: the current sentence, prior text, and world knowledge (see Table 1). The second, we call exemplars. The exemplars for a given sentence include three examples of sentence-focused, local-focused, and global-focused strategies that were previously collected and coded as good examples of the three categories. For example, the protocol “The updraft cannot support the amount of precipitation after an hour” was used as a sentence-focused benchmark for sentence 13 of the thunderstorm text. The third type is strategy benchmarks. The content of these benchmarks represent different types of reading strategies that could be displayed for that sentence. For example, the protocol “The surging of warm and moist air add to the height of the cloud” served as a bridging benchmark for sentence 13 in that it explains why the precipitation becomes too heavy for the updraft. There were up to three benchmarks representing paraphrases of the current sentence, bridges that could be made at that sentence, and elaborations which had been made at that sentence. The exemplar and strategy benchmarks were taken from hand-coded protocols previously collected.

Students in this study read and provided self-explanations to each sentence of two expository texts. Cosines were computed between each self-explanation and the benchmarks. There were up to 21 cosines altogether – 3 for content words, 9 for exemplars that included 3 for each global strategy (sentence-focused, local, and global), and approximately 9 for specific strategies which included 3 for each particular strategy (i.e., paraphrases, bridges, and elaborations), although some sentences produced fewer than 3 bridges or elaborations. The mean cosine for benchmarks representing each strategy was computed. This resulted in nine cosines for each self-explanation; three for content words, three for exemplars, and three for strategies.

Human judges categorized the self-explanations on general and specific reading strategies. The inter-judge reliability on a subset of self-explanations was adequate, Kappa = .80. Using a randomly-chosen half of the self-explanations¹, Millis et al. (2004) predicted the type and presence of general and specific reading strategies from the cosines and the length of the self-explanation (log number of words). We computed a series of discriminant analyses for this task because discriminant analysis predicts group membership, and in this case, the type of general reading strategy (i.e., sentence-focus, local, global) and specific reading strategy (i.e., presence or absence of a paraphrase, bridge, or elaboration) is analogous to group membership. We then used signal detection to assess the

¹ The other half was used to test the coefficients obtained from the first half, but these data are not reported here because they do not directly address differences among benchmarks.

accuracy of prediction for each reading strategy. A d' was computed for each strategy based on hits and false-alarms. For example, a hit for a sentence-focused strategy occurred when both the discriminant analysis and human judges assigned the self-explanation to that category. A false-alarm for sentence-focused strategy occurred when the discriminant analysis categorized the self-explanation as a sentence-focus strategy, but the human judged it as either a local or a global strategy.

The d' 's for each strategy are shown in Figure 2. The magnitude of the d' 's ranged from .89 to 1.54 (chance performance equals zero). McNamara et al. (this volume) reports similar values, but has improved classification when word-based algorithms are added to the LSA cosines. Nevertheless, the magnitude is less important than the pattern across benchmarks since the goal is to compare types of benchmarks. The d' 's for general strategies in Figure 2 are the mean of the d' 's for local and global strategies. Across all types of benchmarks, the d' 's for sentence-focused ($\underline{M} = 1.50$) and global ($\underline{M} = 1.21$) general reading strategies were much higher than for local strategies ($\underline{M} = .63$), indicating that the approach is most suitable for identifying self-explanations that are either minimal or complete with multiple strategies.

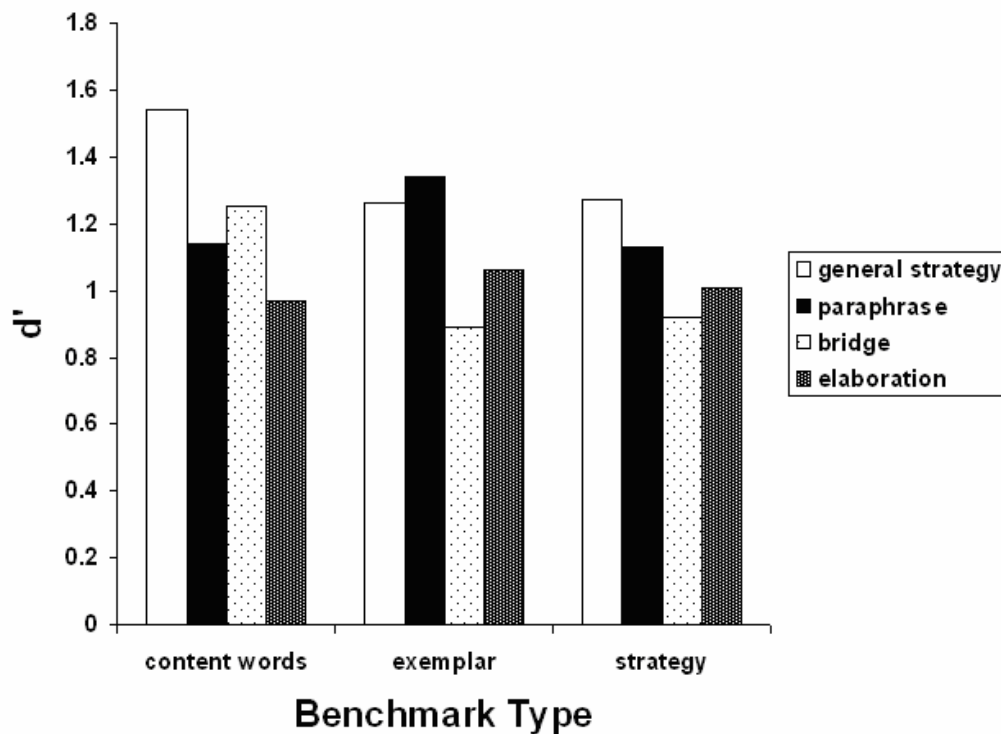


Figure 2

The d' 's indicate some differences among the different types of benchmarks in their utility for classifying reading strategies. In regard to identifying general strategies, the content words produced d' 's that were .28 higher than the exemplar and strategy approaches. Generally, the content words did better than exemplars and strategy benchmarks, except that the exemplars did slightly better than the content words in correctly identifying paraphrases and elaborations. From a practical consideration, this is advantageous because the content words' benchmarks were by far the easiest to construct.

Improving Comprehension: Can Feedback Affect General Strategy Use?

Given that we have been able to predict general strategies using LSA, it is important to consider whether it can be used to improve comprehension. A starting point is whether feedback based on the classification of self-explanations affects the use of the strategies. Feedback plays a crucial role in improving learning, inferencing, and in computer-assisted learning environments and intelligent tutoring systems (Winne, Graham, & Prock, 1993). The chapter by McNamara et al. (this volume) discusses a much more elaborate system than what is presented here, but both provide feedback based on typed-in self-explanations. In this study, we were interested in whether feedback based on an LSA-based classification of general reading strategies would be effective in producing a change in the quality of their self-explanations, as evaluated by the system. If feedback does increase the quality, then this would not only provide further support that LSA can be used to assess comprehension, but that LSA could be used to improve comprehension.

Methods

Undergraduate psychology students at Northern Illinois University ($N = 87$) were given a brief tutorial on self-explaining text. In this tutorial, they were told about how self-explanations can increase comprehension, and were given examples of self-explanations that varied on quality, and also examples of paraphrases, bridges, and elaborations. They were specifically told that although

paraphrasing provides a good first start to a self-explanation, effective self-explanations include multiple strategies. They were then instructed to evaluate a computerized tutor meant for students to practice generating self-explanations. The interface of the tutor was fairly simple. It consisted of three textboxes: one in which the text would appear one sentence at a time, one in which the participant was told to type his or her self-explanation, and one in which the tutor gave feedback to the student. Students in the feedback conditions (see below) were also asked to rate the appropriateness or the quality of the feedback using a 1 (inappropriate) to 6 (very appropriate) Likert-type scale. This task was meant to increase the plausibility of their task, namely to evaluate the tutor. Because of time restrictions, participants only self-explained one of two expository texts, “The Origins of Coal” or “Heart Problems.”

The participants were randomly assigned to one of four conditions. In a no-feedback condition, the participants did not receive any feedback on their self-explanations. In a general-only condition, they only received feedback consisting of one word (e.g., “OK”) based on the general quality of their self-explanation. In a full condition, they received general feedback and also feedback on using particular strategies, if they were identified. An example of full feedback would be: “Good. I'm pleased to see that you are bridging here.” In order to maximize the naturalness of the feedback, the computer randomly picked one out of six responses for each strategy. In a random condition, the computer gave full

feedback based on randomly generated values between 0 and 1.0. Examples of feedback are listed in Table 4.

Table 4

The type of feedback within the feedback conditions was based on the classification of the self-explanation. The cosines obtained with the content word benchmarks (current sentence, prior causal sentences, and world knowledge), along with the number of words in the self-explanation were entered into a weighted linear equation that produced a predicted general reading strategy (text focused, local, global). The weights were computed from a discriminant analysis that was conducted on an independent data set. The computer gave the feedback “Ok.,” “Good.,” “Excellent.” to classified text-focused, local, and global self-explanations, respectively. In the general feedback condition, this was the extent of the feedback. However, in the full feedback condition, additional feedback was given if the computer identified any specific strategy based on rules using the magnitude and pattern of the cosines. These threshold values were determined empirically from the prior independent data set. For example, the self-explanation was coded as a paraphrase if the cosine for current sentence was greater than .80, and if the cosines for prior text and world knowledge were both lower than .15. The self-explanation was coded as including a bridge if the cosine for prior text was greater than .40, and an elaboration if the cosine for world knowledge was greater than .37.

Results

Accuracy of Classifications. Two humans independently classified the self-explanations from sixteen participants as either being text-focused, local, or global on general quality. These were coded from 1 to 3, respectively. The reliability of the judges was adequate, $r = .79$ ($\alpha = .91$). The correlations between the raters and the computerized classifications were .70 and .75. Therefore, the classifications based on LSA agreed with the human judges roughly to the extent that they agreed with one another.

The Effect of Feedback on Overall Quality. For each participant, the computerized classifications of text-focused, local, and global self-explanations were assigned scores of 1, 2, and 3, respectively. A mean quality score was computed for each participant by averaging these scores. The averaged quality scores were analyzed using a 4 (feedback condition) by 2 (text) between-participants ANOVA. The main effect of condition was significant, $F(3, 79) = 3.67$, $p < .05$. The mean quality scores for no-feedback, random, general-only, and full were 1.72 ($SD = .63$), 1.85 ($SD = .50$), 2.17 ($SD = .65$), and 2.12 ($SD = .56$), respectively. Post-hoc comparisons indicated that the no-feedback conditions received significantly lower quality scores than the general-only and full conditions, but not the random condition. There was also a main effect of text, $F(1, 79) = 4.91$, $p < .05$. The mean for the heart and coal passages were 2.1

and 1.8, respectively. The condition by text interaction was not significant, $F(1, 79) = 2.41, p < .10$.

The Effect of Feedback on Using Effective Strategies. For each participant, we computed the percentage of self-explanations of each type. The percentages for bridging and elaborative inferences were combined because these two strategies were described to the participants as being effective strategies, whereas paraphrasing was described as merely a beginning to self-explaining. These values were submitted to a 2 (strategy type) by 4 (feedback condition) by 2 (text) mixed ANOVA with strategy as the within-participants factor. The only significant result was a strategy by condition interaction, $F(3, 79) = 3.72, p = .05$. The interaction is shown in Figure 3. As feedback became more complete, paraphrasing decreased whereas bridging and elaborating increased. Indeed, when there was no feedback, there was no statistical difference between the occurrences of paraphrasing and bridging or elaborative inferences ($p < .70$). In the presence of random feedback, the difference was marginally significant ($p = .06$); however, the difference was highly significant for general and full feedback (p 's $< .001$).

The Effect of Feedback on Appropriate Ratings. A mean appropriateness rating was computed for each participant in the feedback conditions. The means for the random, general, and full conditions were 4.7, 4.3, and 4.5, respectively. There were no significant effects when the means were submitted to ANOVA.

Summary of Feedback Study. In summary, feedback as determined by LSA cosines between the self-explanations and semantic benchmarks improved the quality of the self-explanations. Feedback in the general and full conditions elicited more effective general reading strategies than having no feedback. The same was true for specific reading strategies. Paraphrasing decreased whereas bridging and elaborations increased with feedback. Interestingly, there was no difference between the general and full conditions on general and specific reading strategy usage, suggesting that one word (e.g. “Good”) was just as effective as one word with additional information (e.g., “Good. I like it when you bring up prior text.”). Perhaps participants inferred the additional information in the general condition from the content of the feedback that they were given. Random feedback improved self-explanations somewhat and participants were seemingly unaware that the feedback was not appropriate. This finding is not too surprising given that random feedback will be correct by chance roughly one-third of the time. However, the result also indicates that the participants showed little meta-cognitive awareness of their strategy use with respect to the appropriateness of the feedback.

Summary

The use of LSA to assess comprehension offers several advantages over standard multiple-choice tests. First, the verbal protocols that LSA categorizes are snapshots of the comprehender’s mind during comprehension. Second, LSA

analyses of verbal protocols reveal the use of different types of reading strategies, such as paraphrasing, bridging, and elaborating. Third, using these classifications to provide feedback to the reader improves the quality of the self-explanations. Fourth, producing a verbal protocol in the manner that we have does not have the “feel” of taking a traditional pen-and-paper test, partly because the student is told that there is no right or wrong answers. This might reduce lowered performance due to test anxiety and “stereotype threat” (e.g., Steele & Aronson, 1995).

We should point out that we have had encouraging results with LSA despite the fact that others in this volume have reported difficulty when using short texts. With short texts, replacing one word might lead to very different cosines. In our case, the texts are benchmarks and self-explanations. We can only speculate as to why we have been relatively successful. One reason might be that most of our research is correlational in which many items are averaged, thereby reducing the effect of outliers. Classifying self-explanations into discrete reading strategy categories has been more challenging. For example, McNamara et al. (this volume) reports that various word-based algorithms must be added to LSA cosine in order to achieve a satisfactory level of classification. The feedback study described above also used a classification procedure, but in that study, the classification of the self-explanations was probably just good enough so that the feedback content could change the quality of the self-explanations.

In conclusion, our results indicate that LSA can be used successfully to predict comprehension, assess overall reading strategies, and identify specific reading strategies in verbal protocols. These results provide encouraging directions for assessing deep level comprehension using an on-line and automatic measure. Moreover, the LSA-based system can be used to generate feedback to the reader, which in turn improves strategy use. Thus, LSA based measures can be used to not only assess comprehension, but also to improve it.

References

- Carver, R. P. (1992). What do standardized tests of reading comprehension measure in terms of efficiency, accuracy, and rate? Reading Research Quarterly, 27, 346-359.
- Ericsson, K.A. & Simon, H. A. (1993). Protocol analysis: verbal reports as data. Cambridge, MA.: MIT Press.
- Far, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. Journal of Educational Measurement, 27, 209-226.
- Graesser, A.C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. Psychological Review, 101, 371-395.
- Hanna, G.S., & Oaster, T.R. (1980). A new tool for measuring and understanding individual differences in the component processes of reading comprehension. Journal of Educational Psychology, 93, 103-128.
- Katz, S., Lautenschlager, G., Blackburn, A., & Harris, F. (1990). Answering reading comprehension items without passages on the SAT. Psychological Sciences, 1, 122-127.
- Magliano, J. P & Millis, K. K. (2003). Assessing reading skill with a think-aloud procedure and latent semantic analysis. Cognition and Instruction, 21, 251-284.

- Magliano, P., J., Wiemer-Hastings, K., Millis, K. K., Muñoz, B.D., McNamara, D. (2002). Using latent semantic analysis to assess reader strategies. Behavior Research Methods, Instruments, & Computers, 34, 181-188.
- McNamara, D.S. (2004). SERT: Self-explanation reading training. Discourse Processes, 1-38.
- McNamara, D.S., Kintsch, E., Songer, N.B., & Kintsch, W. (1996). Are good texts always better? Text coherence, background knowledge, and levels of understanding in learning from text. Cognition and Instruction, 14, 1-43.
- 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.
- Millis, K.K., Kim, H.J., Todaro, S. Magliano, J., Wiemer-Hastings, K., McNamara, D. (2004). Identifying reading strategies using latent semantic analysis: Comparing semantic benchmarks. Behavior Research Methods, Instruments, & Computers, 36, 213-231
- Millis, K.K., Magliano, J.P., Wiemer-Hastings, K., & McNamara, D. (2001). Using LSA in a computer-based test of reading comprehension. In J.D. Moore, C. Luckhardt-Redfield, & W.L. Johnson (Eds.), Artificial intelligence in education: AI-ED in the wired and wireless future: Vol. 68. Frontiers in artificial intelligence and applications (pp. 583-585). Amsterdam: IOS Press.

- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. Psychological Review, 84, 231-259.
- Pressley, M., & Afflerbach, (1995). Verbal protocols of reading: The nature of constructively responsive reading. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Shapiro, A.M., & McNamara, D.S. (2000). The use of latent semantic analysis as a tool for the quantitative assessment of understanding and knowledge. Journal of Educational Computing Research, 22, 1-36.
- Steele, C. M. & Aronson (1995). Stereotype threat and the intellectual test performance of African Americans. Journal of Personality & Social Psychology, 69, 797-811.
- Trabasso, T., & Magliano, J. P. (1996). How do children understand what they read and what can we do to help them? In M. Grades, P. van den Broek, & B. Taylor (Eds.), The first R: A right of all children. NY: Columbia University Press.
- Trabasso, T., van den Broek, P., & Suh, S. (1989). Logical necessity and transitivity of causal relations in the representation of stories. Discourse Processes, 12, 1-25.
- Winne, P.H, Graham, L., & Prock, L. (1993). A model of poor readers' text-based inferencing: Effects of explanatory feedback. Reading Research Quarterly, 28, 53-66.

Table 1. Example Self-explanations and semantic benchmarks to the sentence 13 of the text “Stages of Thunderstorm Development” (see the Appendix): “Usually, within an hour the amount and size of precipitation becomes too much for the updraft to support” taken from a text on thunderstorms. Cosines between the self-explanations and benchmarks are in the cells.

Self-explanation Category	Example Self-explanations	Current Sentence Benchmark	Prior text Benchmark	World Knowledge Benchmark
		Hour size precipitation amount becomes updraft support	Cloud release develops start storm	Fall hold down heavy
Sentence-focused:	The updraft cannot support the amount of precipitation after an hour.	.75	.08	-.06
Local-focused:	Eventually, there is too much moisture and hailstones and the system gets too heavy.	.18	.22	.24
Global-focused:	Thunderstorms only last for about an hour because since the rain is so heavy, the air cannot continue to rise.	.33	.41	.32

Table 2. Beta weights from regression analyses to predict comprehension of narrative texts including LSA cosines and the Nelson-Denny Test of Comprehension.

Predictor Variable	Dependent Variable		
	True-False questions	Recall (think aloud passages)	Recall (silently read passages)
LSA Benchmark			
LSA current sentence	-.61*	-.36**	-.22*
LSA prior causal sentence	.36**	.58**	.47**
Nelson Denny	.02	.11	.23*
R ²	.38**	.45**	.38**

Note: * $p < .05$; ** $p < .01$

Table 3. Beta weights from regression analyses to predict comprehension of expository texts including LSA cosines and the Nelson-Denny Test of Comprehension.

Predictor Variable	Dependent Variable		
	Textbase	Situation Model	Total
LSA Benchmark			
LSA current sentence	-.61*	-.36	-.59*
LSA distal causes	.58**	.59**	.69**
Nelson Denny	.25*	.30**	.32**
R ²	.20**	.20**	.28**

Note: * $p < .05$; ** $p < .01$

Table 4. Computer feedback as a function of self-explanation category.

Self-Explanation Category	Computer Feedback
General quality:	
Text-focused	“OK”
Local	“Good”
Global	“Excellent”
Specific strategy:	
Sample Feedback	
Paraphrase	"I believe you are rephrasing the sentence for the most part." or "Are you primarily paraphrasing the sentence?" etc.
Bridge	"I like it when you connect your thoughts with a previous sentence." or "I'm pleased to see that you are bridging here." Etc.
Elaboration	"It looks to me that you are drawing upon what you

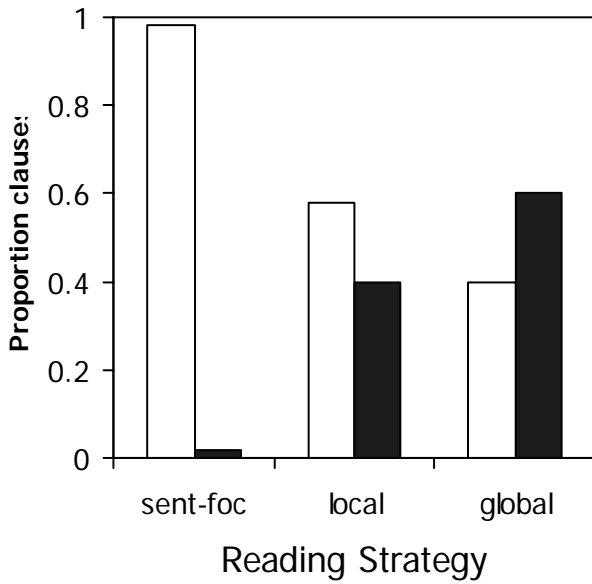
know." or "I like it when you tie in what you know into the self explanation." Etc.

Figure Caption

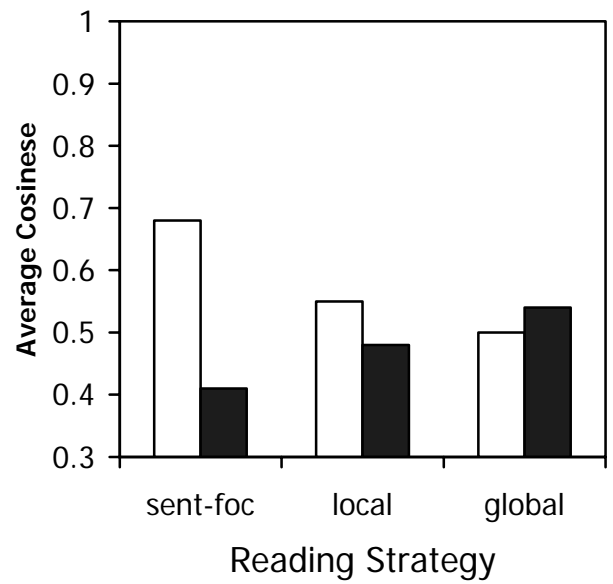
Figure 1. The percentage of clauses in verbal protocols judged by humans to contain information from either the current sentence and prior text or world knowledge (left) and LSA cosines between the verbal protocols and current sentence and prior text or world knowledge benchmarks (right) as a function of general reading strategy (from Magliano et al., 2002).

Figure 2. d 's as function of type of benchmark and strategy.

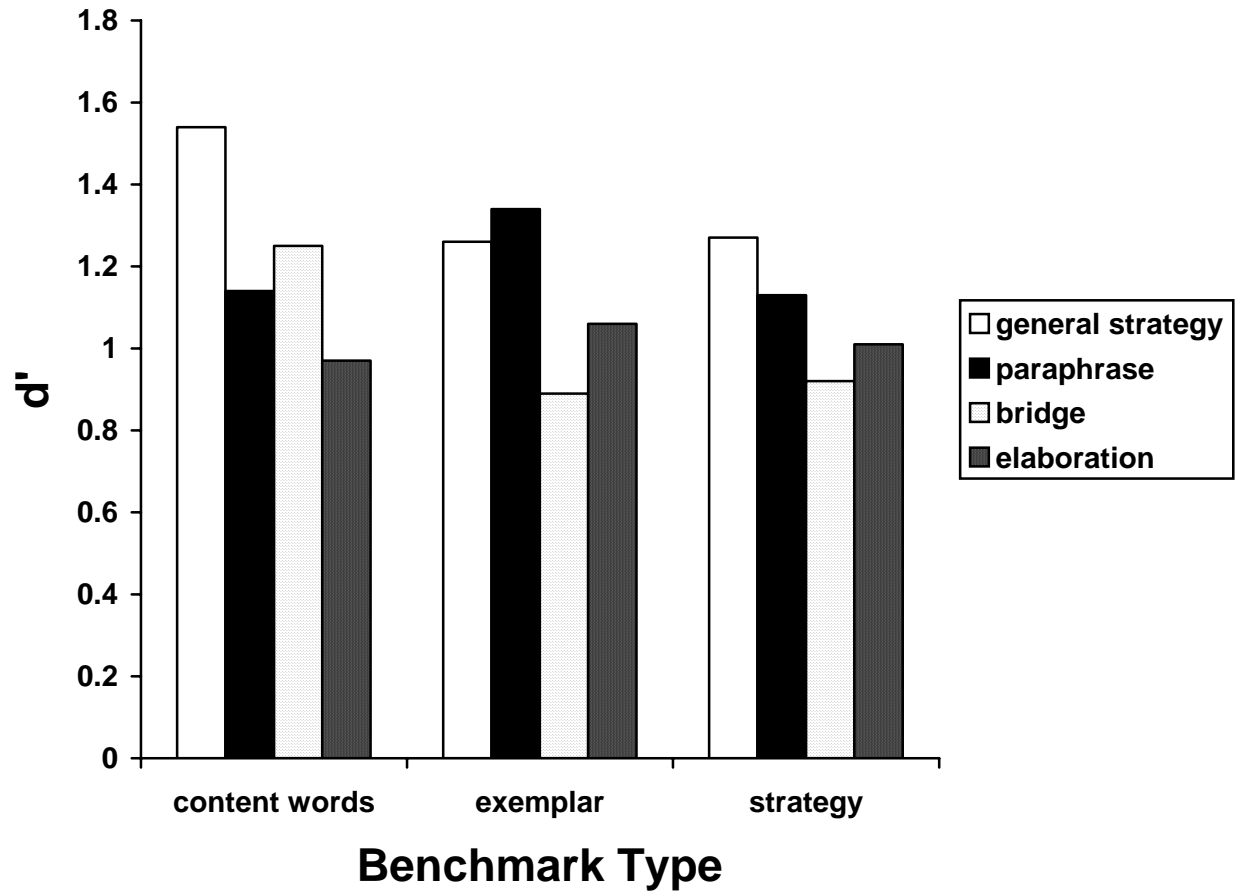
Figure 3. The percentage of paraphrases and bridges or elaborations in self-explanations classified by tutor under four feedback conditions.

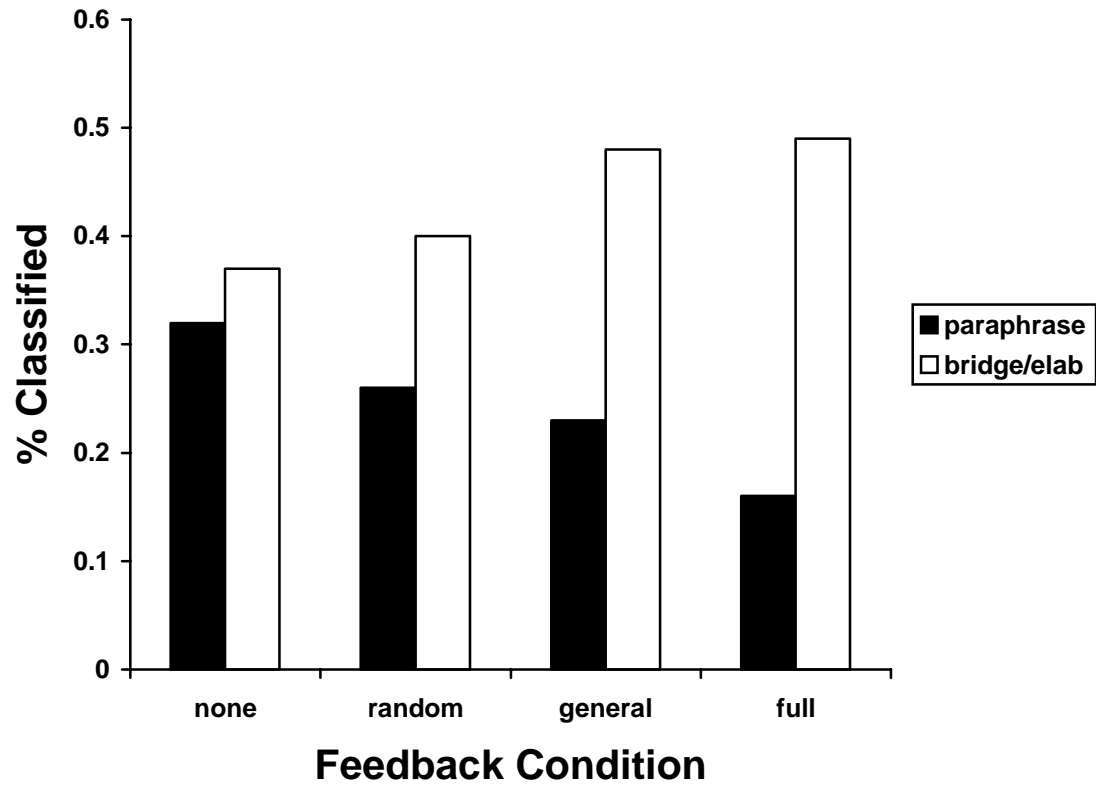


□ current sentence ■ prior text/wk



□ current sentence ■ prior text/wk





APPENDIX

Stages of Thunderstorm Development

1. All thunderstorms have a similar life history.
2. Thunderstorms start with the development of large cumulonimbus clouds.
3. The development of these clouds requires warm, moist air.
4. As this warm, moist air is lifted, it releases sufficient latent heat to provide the buoyancy necessary to maintain its upward flight.
5. This process is facilitated when there are high surface temperatures.
6. As such, thunderstorms are most common in the late afternoon and early evening.
7. However, surface temperature alone is not sufficient for the growth of towering cumulonimbus clouds.
8. Fueled by only surface temperatures, at best the cloud would be small and evaporate in 1-15 minutes.
9. The development of large cumulonimbus clouds requires a continual supply of warm, moist air.
10. Each new surge of warm, moist air rises higher than the last.
11. This process continually adds to the height of the cloud.
12. When these updrafts reach speeds up to 60 miles per hour, they are capable of supporting hailstones and great amount of precipitation.

13. Usually, within an hour the amount and size of precipitation becomes too much for the updraft to support.
14. One part of the cloud develops a downdraft.
15. Rain begins to fall.
16. These downdrafts can also cause gusty winds.
17. It is during this stage that lightening usually occurs.
18. Eventually downdrafts dominate throughout the cloud.
19. The cooling effect of falling precipitation coupled with the influx of colder air aloft mark the end of the thunderstorm activity.
20. Although the life span of a cumulonimbus cell is only about an hour, a storm can develop new cells as it moves.