Behavior Research Methods, Instruments, & Computers
2004, 36 (2), 213-221

Identifying reading strategies using latent semantic analysis: Comparing semantic benchmarks

KEITH MILLIS
Northern Illinois University, DeKalb, Illinois

HYUN-JEONG JOYCE KIM
Rhodes College, Memphis, Tennessee

STACEY TODARO, JOSEPH P. MAGLIANO, and KATJA WIEMER-HASTINGS
Northern Illinois University, DeKalb, Illinois

and

DANIELLE S. McNAMARA
University of Memphis, Memphis, Tennessee

We explored methods of using latent semantic analysis (LSA) to identify reading strategies in students' self-explanations that are collected as part of a Web-based reading trainer. In this study, college students self-explained scientific texts, one sentence at a time. LSA was used to measure the similarity between the self-explanations and *semantic benchmarks* (groups of words and sentences that together represent reading strategies). Three types of semantic benchmarks were compared: content words, exemplars, and strategies. Discriminant analyses were used to classify global and specific reading strategies using the LSA cosines. All benchmarks contributed to the classification of general reading strategies, but the exemplars did the best in distinguishing subtle semantic differences between reading strategies. Pragmatic and theoretical concerns of using LSA are discussed.

Active reading is necessary for understanding challenging text. One way that a reader can be active is to generate a *self-explanation*. A self-explanation is a written or oral verbal protocol produced by a reader in an attempt to understand the current sentence in the context of the passage. Generating self-explanations has been shown to increase the comprehension of unfamiliar scientific text (Chi, de Leeuw, Chiu, & LaVancher, 1994; Magliano, Trabasso, & Graesser, 1999; McNamara & Scott, 1999). Self-explanations comprise specific reading strategies, such as paraphrasing the current sentence, making bridges or conceptual links between the current sentence and prior text, activating relevant world knowledge, using logic and common sense, and recognizing when we do not understand (i.e., monitoring; McNamara & Scott, 1999).

McNamara and colleagues have been developing a Web-based program, called iSTART (Interactive Strategy Training for Active Reading and Thinking), that teaches high school students how to self-explain scientific text effectively (Levinstein, McNamara, Boonthum, Pillaraisetti, & Yadivalli, 2003; Magliano, Wiemer-Hastings, Millis, Muñoz, & McNamara, 2002; McNamara, 2003; Millis, Magliano, Wiemer-Hastings, & McNamara, 2001). iSTART includes a practice module during which the student produces a self-explanation after reading each sentence of scientific text. iSTART then classifies the self-explanation on the extent to which it reveals the use of active reading strategies. Once classified, iSTART gives feedback to students so that they can calibrate their self-explanations and, if need be, improve their ability to self-explain text.

Feedback in iSTART relies on the accurate classification of the self-explanation, and, therefore, classification is a critical feature of the practice module. We have been exploring and testing procedures in which latent semantic analysis (LSA) is used to classify self-explanations into categories of reading strategies. LSA is a statistically based technique that assesses the semantic similarity between any two units of language (words, sentences, paragraphs, texts; Landauer & Dumais, 1997). It is used in discourse psychology, educational psychology, and computer science. LSA provides a numerical value (a cosine) corresponding to the semantic similarity between any two texts. It has been used to grade essays (Foltz, Gilliam, & Kendall, 2000; Foltz, Laham, & Landauer, 1999), provide feedback in an automated tutor (e.g., Graesser et al., 2000), and gauge understanding of discourse (Shapiro & McNamara, 2000). The procedure common to these applications has been the comparison of verbal input with a standard text. We refer to the standard text as a *semantic benchmark*. For example,

Correspondence concerning this article should be addressed to K. Millis, Department of Psychology, Northern Illinois University, DeKalb, IL 60115 (e-mail: kmillis@niu.edu).

to grade the goodness of an answer to a question, LSA can compare the answer (the verbal input) to an “ideal” answer serving as the semantic benchmark. A high LSA cosine would indicate a good answer.

The present study systematically compared different types of benchmarks that LSA could use to classify reading strategies expressed by self-explanations. But before we describe the benchmarks, let us first describe general and specific reading strategies that they were designed to identify. By *general reading strategy*, we refer to the extent to which an individual is actively expanding on the current sentence in his or her self-explanation. There are three types of general reading strategies that a reader may adopt for any given sentence. One is a *sentence-focused* strategy, whereby the person largely focuses on the current sentence and adds little or no world knowledge or prior text to its representation. Another is a *local* strategy, through which the reader activates and integrates some information from the previous sentence into the self-explanation. The third general strategy is called a *global* strategy. A person using a global strategy activates prior text and his or her world knowledge in order to integrate them with the theme of the text.

Table 1 shows examples of general reading strategies and self-explanations for Sentence 13 of a passage describing the development of thunderstorms (see the Appendix). *Sentence-focused* explanations (e.g., *The updraft cannot support the amount of precipitation after an hour*) are either paraphrases or rewordings of the sentence. All of the information in the self-explanation is based on the current sentence. *Local* explanations include aspects of the current sentence, as well as either a reference to the sentence immediately prior to it or an elaboration of a concept mentioned in the current sentence. In the first local example, the word *hailstones* had appeared in the preceding sentence, and the phrase *the system gets too heavy* elaborated the idea in the current sentence that the updraft cannot support the precipitation. *Global* explanations include a description of how the sentence ties in with the theme of the text (e.g., *Thunderstorms last only about an hour*), elaborations from world knowledge (e.g., *this must be when it starts raining*), and bridging or causal-based inferences (e.g., *since the rain is so heavy, the air cannot continue to rise*). Global self-explanations employ multiple reading strategies taught by iSTART.

The general strategies described above comprise specific strategies. There are three specific strategies of importance to the current study. The first is *paraphrasing*. Paraphrasing occurs when the reader rewords the current sentence as part of the self-explanation. Paraphrases typically occur in all three types of general reading strategies (see McNamara, 2003), but they are most salient in sentence-focused explanations. A *paraphrase only* refers to a self-explanation that contains only a paraphrase. The second strategy is *bridging*, in which the reader includes ideas mentioned earlier in the text. The mention of hailstones would count as a bridge between Sentences 11 and 12. The third strategy is *elaboration*. An elaboration occurs when the reader includes relevant information that had not been mentioned in the text. The knowledge that thunderstorms produce rain and hail would be an elaboration if these concepts had not yet been mentioned in the text. Sentence-focused self-explanations contain paraphrases, whereas global self-explanations could contain all three strategies.

As was mentioned above, in the present study we compared different types of semantic benchmarks that LSA could use to classify reading strategies. Recall that semantic benchmarks are the standard texts that LSA uses to compare verbal input. Researchers have used different types of benchmarks, but there is little research comparing their relative effectiveness. Foltz et al. (2000) compared two benchmark approaches in a study in which student essays were graded with LSA. In their componential approach, the benchmarks were sentences from the textbook that supplied the essay topic. Each benchmark sentence conveyed a different subtopic that was to be covered in the essays. The final essay grade was based on the number of subtopics (benchmarks) that resulted in

Table 1
Examples of Sentence-Focused, Local, and Global Self-Explanations to Sentence 13 of Stages of Thunderstorms

General Reading Strategy	Example Self-Explanation
Sentence focused	Within an hour, the cloud has way too much precipitation that it can handle. The updraft cannot support the amount of precipitation after an hour.
Local	Eventually, there is too much moisture and hailstones and the system gets too heavy. When the updraft can no longer hold the precipitation and hailstones, it releases them, causing rain or hail.
Global	The updrafts keep on adding to the amount of precipitation in the cloud. When this becomes too much, this must be when it starts raining. Thunderstorms only last for about an hour because since the rain is so heavy, the air cannot continue to rise.

Note—See the Appendix for the text of the Sentence 13.

an LSA cosine that exceeded a threshold value. In their holistic approach, prescored essays served as the benchmarks. With this method, a cosine was computed between the essay to be graded and each of the prescored essays. The essay's grade was the average of the grades of the benchmark essays, weighted by the cosines. Foltz et al. (2000) reported that both procedures correlated highly with those of human graders (r = .71–.89), but the holistic approach did better than the componential approach.

Types of Benchmarks

Because we wanted to identify reading strategies exhibited by self-explanations generated for each sentence text, it was necessary to construct benchmarks for each sentence. Three types of benchmarks were compared in the present study: *content words*, *exemplars*, and *strategies*. In Table 2, example benchmarks for the Sentence 13 of the Appendix are presented.

The content word benchmarks contained only content words—no syntax or function words. The content word benchmarks for a sentence represented different sources of information on which the reader could draw to construct a self-explanation for that sentence: the current sentence, the prior sentence, and world knowledge. The *current sentence* benchmark contained the nouns, verbs, adjectives, and adverbs from the current sentence, provided they were not semantically depleted (e.g., *is*, *are*). The *prior text* benchmarks contained content words from prior sentences judged to be causally related to the current sentence via a causal network analysis (Trabasso, van den Broek, & Suh, 1989). The *world knowledge* benchmark contained content words thought to reflect elaborative inferences that the reader could make from that sentence. We used these three benchmarks because readers draw on these sources of information during comprehension (see, e.g., Graesser, Singer, & Trabbasso, 1994). It should be noted that the content word method was similar to Foltz et al.'s (2000) componential procedure in that each benchmark represented particular pieces of information.

The exemplar benchmarks for a sentence contained "good examples" of self-explanations previously obtained for that sentence. For each sentence, there were exemplars representing each type of general strategy. Our use of exemplar benchmarks closely resembled Foltz et al.'s (2000) holistic approach in that the input (self-explanation or essay) is compared with other precategorized examples of the same input type. The strategy benchmarks represented particular types of reading strategies, such as paraphrasing, bridging, and elaboration. Unlike the exemplars, the strategy benchmarks were not taken verbatim from prior self-explanations, but from phrases and sentences representing common strategies exhibited for that sentence. For both exemplar and strategy benchmarks, human judges identified and grouped together good examples and generated representative samples of commonly generated ideas, themes, and expressions. This will be discussed further in the Method section.

Previous research has made use of both content word and exemplar approaches. As was mentioned above, Foltz et al. (2000) used variants of both to grade essays. In our previous research, we have relied exclusively on content word benchmarks. We have been fairly successful in using LSA to classify self-explanations on general reading ability using content words. In a study by Magliano et al. (2002), LSA cosines between self-explanations and the current sentence benchmark decreased from sentence-focused to global strategies, whereas the cosines increased

Table 2
Examples of Content Word, Exemplar, and Strategy Semantic Benchmarks for Sentence 13 of "Stages of Thunderstorm Development"

		Content Word Benchmarks	
Category:	Current sentence	Prior text	World knowledge
Content:	hour, size, precipitation, amount, becomes, updraft, support	cloud release, develops, start, storm	fall, hold, down, heavy
		Exemplar Benchmarks	
Category:	Sentence focused:	Local	Global
Content:	The amount of the precipitation keeps the updraft from supporting it	When the hail becomes accumulated and heavy enough it cannot be supported in the cloud	So the process takes about an hour to accumulate sufficient precipitation that a wind with a speed of more than 60 mph would be needed to support it. So the storm takes about an hour tobuild before it starts "raining" or whatever.
		Strategy Benchmarks	
Category:	Paraphrase	Bridge	Elaboration
Content:	The precipitation becomes too much, too heavy, for the updraft to handle, support	The surging of warm and moist air add to the height of the cloud	It starts to rain when the clouds become too heavy

for the sum of prior text and world knowledge benchmarks.[1] Using a discriminant analysis, they were able to distinguish between self-explanations rated as sentence focused versus those rated as global ($R^2 = .28$) using this type of benchmark. Other researchers have relied on the exemplar approach. For example, "good answers" to questions within the automated tutor AutoTutor developed by Graesser et al. (2000) are coded as complete sentences describing the ideal answer. AutoTutor uses a threshold technique according to which, if the LSA cosine between an answer to a question and the benchmark exceeds a predetermined value, then the answer is treated as acceptable.

Which type of benchmark approach would enable the most accurate classification of reading strategies? According to a source hypothesis, classification should be maximized if the benchmarks represent the theoretical underpinnings of classification. According to our classification system, the general reading strategies draw differentially on three sources of information: the current sentence, prior text, and world knowledge (see Trabasso & Magliano, 1996). Sentence-focused self-explanations draw only on the current sentence; local self-explanations draw on the current sentence, but a little on prior text and world knowledge as well; global explanations draw on all three, but relatively more on prior text and world knowledge. Because the content word benchmarks were designed to represent the source(s) of information, the source hypothesis predicts that content word approach would lead to higher classification accuracy than the exemplar approach.

On the other hand, exemplars contain naturalistic examples of students' discourse. The expressions and words that students use to self-explain might be important in distinguishing subtle differences among strategies. As one can see from Table 1, the content of the self-explanations is very similar among the reading strategies. The local explanations might differ from either sentence-focused or global strategies by one or two content words. Representative samples of each strategy may provide enough contexts for classification to be maximized. Therefore, according to an *exemplar* hypothesis, the exemplar benchmarks should outperform the content word benchmarks.

The strategy benchmarks provided a compromise between content word and exemplar methods. The different strategies come from different sources: Paraphrases come from the current sentence, bridges come from prior sentences, and elaborations come from world knowledge. In this way, they are similar to the content words. However, they are sampled from previously collected self-explanations and are represented in the benchmark as sentences and phrases. In this sense, they are similar to the exemplar approach. We do not make any strong predictions regarding the strategy approach except that it should perform reasonably well at identifying strategies embedded in the self-explanations.

The present study addressed four research goals, each of which extended previous research. The first was to test the source and exemplar hypotheses in the context of identifying general reading strategies. This involved comparing the content word and the exemplar benchmarks. The second was to explore whether LSA could accurately identify specific strategies. Magliano et al. (2002) examined only general strategies, so the extent to which LSA can identify specific strategies, such as bridging inferences, is unknown. The third goal was to determine whether we can generalize from one set of results to another. Magliano et al. (2002) used the same set of self-explanations to generate coefficients (from either multiple regression or discriminant functions) and to test their predictive ability. This leaves unanswered the question of whether the data from one set will generalize to another set. This is a crucial issue, because iSTART will be classifying self-explanations generated by students via the Web using a classification procedure developed from the input of another group of participants. Our final goal was to see whether we could replicate Magliano et al. (2002) using an LSA space constructed specifically for iSTART (Kurby et al., 2003). Magliano et al. (2002) used the general reading space at the University of Colorado LSA Web site (http://lsa.colorado.edu), whereas this study used an LSA space that was based on science text.

METHOD

Participants

Thirty-six undergraduate students attending Northern Illinois University participated for course credit.[2]

Materials and Benchmark Construction

We used four texts that were adopted from high school textbooks on life science. One described thunderstorm development, one explained heart disease, one explained how coal is made, and the last described the food chain. The reading levels were suitable for college freshman and ranged in length from 20 to 34 sentences, for a total of 98 sentences. We constructed benchmarks for 5 sentences from each text, for a total of 20 sentences. The sentences were sampled from different parts of the text. Example benchmarks are shown in Table 2.

The content word benchmarks were taken from Magliano et al. (2002). The current sentence benchmark contained the nouns, verbs, adjectives, and adverbs of the current sentence. The prior text benchmark contained the content words from sentences identified as causal antecedents to the current sentence according to the procedure described by Trabasso et al. (1989). The words in the world knowledge benchmark were derived empirically from an independent group of participants who explained and elaborated upon the current sentence. Content words generated for each sentence by two or more participants were included in the world knowledge benchmark for that sentence. The three benchmarks for any given sentence did not have any words in common.

The exemplar and strategy benchmarks were constructed from self-explanations generated by another group of participants ($N = 40$; Magliano et al., 2002). The corpus consisted of 20 self-explanations for each of the 20 test sentences. In order to construct the exemplar benchmarks for each test sentence, each self-explanation in the corpus was first classified on general reading strategy by two independent human judges ($r = .79$). The two judges then identified self-explanations within each reading strategy that focused on similar information. For example, the two self-explanations *when the hail becomes accumulated and heavy enough, it cannot be supported in the cloud* and *the updraft cannot hold all the hailstones and precipitation, so like after an hour, it might just fall* were grouped together because both focused on the weight of the hail. Finally, for each test sentence the judges picked three representative self-explanations from each reading strategy (sentence focused, local, and global) in an attempt to pick examples from the different content groupings. Thus, an attempt was made to pick exemplars of each type of strategy that focused on different ideas. This was done to account for as many future responses as possible. The chosen self-explanations were the exemplar benchmarks.

For the strategy benchmarks, the two judges identified paraphrases, bridges, and elaborations embedded in the corpus. For an item to serve as a strategy benchmark, 2 or more students must have generated it as a main idea. The paraphrase strategy benchmarks contained phrases and words that the previous students had used to summarize information in the sentence. Bridges were words or clauses that had been reinstated from previous sentences. The example bridge in Table 2 is an amalgamation of two earlier sentences (10 and 11). Elaborations were words or clauses representing ideas not found in the text. For example, the elaboration in Table 2 (*It starts to rain when the clouds become too heavy*) is a causal consequence of the current sentence and was generated from world knowledge. An attempt was made to identify three examples of paraphrases, bridges, and elaborations for each test sentence. However, that was not always possible, because sentences differed in the extent to which they elicited particular strategies. Five sentences had only two paraphrase benchmarks; three had only two elaboration benchmarks. The most variability occurred for bridges: Two sentences did not have any bridge benchmarks at all, eight had only one, two had two, and three had all three.

Procedure

The participants in the present study first attended an hour-long session in which they learned about self-explanations and how they are used to increase comprehension. The session was based on McNamara and Scott (1999). The presenter (one of the authors) described and gave examples of the various reading strategies (paraphrasing, bridging, elaboration, comprehension monitoring, predicting, using logic, and common sense). The examples were also included in a booklet that was administered to each participant. It was emphasized that although paraphrasing was a reading strategy, it was only a start; good self-explanations usually include more than one strategy. The participants then watched a videotape of a young woman self-explaining a text, and the presenter led a discussion on the different reading strategies that she used. Finally, the participants worked in pairs, taking turns self-explaining a science text. When one member of the pair self-explained a sentence, the other was instructed to identify the strategies that he or she used.

Within a week of the self-explanation training session, the participants self-explained every sentence of two of the four science texts. Texts and participants were counterbalanced so that an equal number of participants self-explained each text. The participants were tested individually at computers in separate rooms. The participants read each text, one sentence at a time, in a text box on the computer screen. Following each sentence, they were instructed to type a self-explanation in an answer box. When they were done typing the self-explanation, they clicked a Continue button that caused the next sentence of the text to appear. The presentation of the text was cumulative, so that they could scroll back to earlier parts of the text (but not to what they had previously written) if they wished. When they finished self-explaining both texts, they were debriefed and dismissed.

RESULTS

Overview of Strategy

We used a four-step procedure to answer the research questions. In Step 1, we coded the self-explanations on global and specific reading strategies and generated cosines between them and the benchmarks. This gave us an overall impression of the utility of the different benchmarks. In Step 2, we performed discriminant analyses on one half of the data set, predicting global and specific reading strategies from the LSA cosines with the various benchmarks. In Step 3, we used the functions from the discriminant analysis computed in Step 2 to predict the presence and type of reading strategies in the other half of the data. In Step 4, we assessed the accuracy of prediction using signal detection analyses. Steps 3 and 4 allowed a test of the source and exemplar hypotheses, whether we could identify specific reading strategies, and whether the results from one data set would generalize to another.

Step 1: Coding Reading Strategies and LSA

Two raters classified each self-explanation of the 20 selected sentences as based on a sentence-focused, a local, or a global strategy and judged whether it contained a bridge, an elaboration, a paraphrase, or no strategy other than a paraphrase (i.e., paraphrase only). Disagreements were resolved through discussion. At each test sentence, we calculated the LSA cosine between each self-explanation and each of the benchmarks associated with that sentence. For each sentence, there were 21 benchmarks: 3 for the content word (current sentence, prior sentence, and world knowledge), 9 for the exemplar (3 for each general reading strategy), and 9 for the strategy (3 for each specific reading strategy) benchmarks. Consequently, 21 cosines were computed for each self-explanation. For each self-explanation, we averaged the cosines across the benchmarks of a particular type: the three exemplars of sentence-focused, local, and global benchmarks, as well as the strategy benchmarks for paraphrases, bridges, and elaborations. This resulted in nine averaged cosines for each self-explanation. There was a total of 740 self-explanations.

The LSA space was constructed by sampling general as well as specific science texts, with an emphasis on biology topics (273 documents, 849,060 words; see Kurby et al., 2003). The space was empirically determined by Kurby et al. (2003) to be optimal for identifying strategies in verbal protocols obtained from participants reading science texts. The space was constructed from 20% general science documents (e.g., weather cycles) and 80% specific science documents (e.g., types of heavy storms). The space had 350 dimensions.

To test the suitability of the LSA space and whether we replicated Magliano et al. (2002), we computed and inspected the mean cosines between the benchmarks and general reading strategies (see Table 3). If the LSA space is suitable, there should be an association between the cosines for the benchmarks and the categories of items (i.e., types of self-explanations). Indeed, Table 3 shows such an association. For the content words, the cosines for the current sentence benchmark decreased from sentence-focused to global strategies, whereas they increased for prior text and world knowledge benchmarks. This is the pattern reported by Magliano et al. (2002), except that they had collapsed over world knowledge and prior text benchmarks. A similar pattern occurred for the exemplar and strategy approaches, although the cosines for the exemplar local benchmark did not differ among sentence-focused, local, and global reading strategies ($p < .80$ when the means were submitted to analysis of variance. For each type of benchmark, a significant benchmark type × reading strategy interaction occurred (Fs = 14.13–20.6, ps $< .001$). Together, these results replicated Magliano et al. (2002) with a different LSA space and suggest that the space will be suitable for the analyses reported below.

Step 2: Discriminant Analyses

The primary purpose of this step was to acquire discriminant functions from one set of data that could be applied to the classification of reading strategies in another set. Therefore, we randomly split the self-explanations into two sets: *original* and *test*. The original ($n = 341$) and test ($n = 389$) data sets were very comparable, which indicates their appropriateness for this approach. The percentages of sentence-focused, local, and global strategies in both sets were 20%, 45%, and 35%, respectively. The percentages of paraphrase only, paraphrase, bridging, and elaborations in the original data set were 20%, 65%, 34%, and 59%, respectively. In the test data set, the corresponding percentages were 18%, 60%, 29%, and 65%. In regard to the LSA cosines, we computed a t test on each benchmark, comparing their values across the two data sets. Only two statistically significant differences were found: The cosine for the content words for the current sentence was higher in the original set ($M = .48$, $SD = 25$) than in the test set [$M = .44$, $SD = .44$; $t(728) = 2.21$, $p < .05$], and the cosine for the content words for world knowledge was lower in the original set ($M = 11$, $SD = .11$) than in the test set [$M = .13$, $SD = 12$; $t(728) = 2.12$, $p < .05$].

Using the self-explanations in the original data set, we performed discriminant analyses that predicted general and specific reading strategies. Three discriminant analyses were performed for each type of strategy: one that used the content word cosines, one that used the exemplar cosines, and one that used strategy cosines. We also included the natural logarithm of the number of words in the self-explanation to control for length. Each discriminant analysis produced discriminant functions that we used as weights in a linear combination to predict group membership in the test data set. Discriminant functions are analogous to regression equations. In this case, *group membership* refers to the presence of a particular reading strategy.

Each discriminant analysis was highly significant for predicting general reading strategies in the original data set. Two discriminant functions were computed for each analysis, although the first function in each accounted for the clear majority of variance [range: 95%–99%, χ^2s(8) = 118.1–140.0]. In order to predict specific reading strategies, two judges scored the presence of specific reading strategies in the self-explanations. When a strategy was present according to the two judges, it was coded as 1, and when it was absent it was coded as 0. As for the general reading strategies, all of the discriminant analyses of the original data set were significant: The average chi-squares for the content word, exemplar, and strategy approaches (collapsing over reading strategy) were 102.9, 114.9, and 62.77, respectively (ps $< .001$, $df = 4$). The average chi-squares for paraphrase only, paraphrase, bridge, and elaboration were 100, 63.4, 71.9, and 75.8, respectively.

Thus far, the analyses indicated that the cosines for each type of benchmark could be used to predict the type and presence of reading strategies. Therefore, the findings replicated Magliano et al. (2002), who used only the content word approach. However, Magliano et al. (2002) did not address the extent of the accuracy of the predictions or the generalizability of the results. These issues are raised in Steps 3 and 4.

Table 3
Mean LSA Cosines Between Benchmarks and General Reading Strategies

	General Reading Strategy		
Benchmark	Sentence Focused	Local	Global
Content words			
Current sentence	.51	.45	.45
Prior text	.18	.25	.33
World knowledge	.10	.12	.14
Exemplar			
Sentence focused	.47	.42	.40
Local	.42	.40	.41
Global	.37	.38	.44
Strategy			
Paraphrase	.45	.41	.39
Bridge	.22	.25	.34
Elaboration	.25	.28	.30

Steps 3 and 4: Accuracy of Prediction in Original and Test Data Sets

As was mentioned earlier, we used the discriminant functions generated from the original data set in Step 2 to predict reading strategies in the test data set. We used signal detection to assess the accuracy of prediction. We computed hit rates, false-alarm rates, and then d's as a measure of discrimination for each type of strategy. A hit would occur if the analysis assigned the self-explanation to a category (e.g., local) and this assignment were correct according to the human judges. A false alarm would occur if the analysis assigned the self-explanation to a category (e.g., a local self-explanation), but it were incorrect according to the human judges (i.e., it was not a local self-explanation). d' is a measure of sensitivity between the presence of the signal with noise and noise alone. d' is the distance between signal + noise and noise alone in standard deviation (SD) units. The d's are high to the extent that hits are high and false alarms are low. In this study, we are less concerned with the magnitude of the d's than with differences between benchmark approaches.

General reading strategies. For a comparison of the different benchmarks, the d's for general reading strategies are listed in Table 4. The hit and false-alarm rates are not shown for reasons of space, but the hit and false-alarm rates for sentence-focused, local, and global strategies (averaging over benchmark type and using the original data set) were .70 and .18, .39 and .18, and .69 and .25, respectively. The *d*′s were very similar for content word (*M* = 1.30), exemplar (*M* = 1.09), and strategy (*M* = .96) benchmarks, but the content word benchmark had the advantage. All showed the same pattern: substantially higher *d*′s for the sentence-focused (*M* = 1.50) and global (*M* = 1.21) self-explanations than for the local (*M* = .63) self-explanations.

In regard to generalization, the *d*′s showed the same pattern in the test data as in the original data, but were slightly lower. The average hit and false-alarm rates (collapsing over benchmark) for the sentence-focused, local, and global strategies were .64 and .24, .27 and .15, and .76 and .38, respectively. The average *d*′ for sentence-focused, local, and global self-explanations was 1.08, .53, and 1.10, respectively. Lower *d*′s were expected because the functions are applied to a new data set. The *d*′s declined on the average of one-fifth of an *SD*.

Comparing benchmarks. To a large extent, the benchmarks produced similar results. The content word and strategy benchmarks showed the same pattern as each other, but the *d*′s were slightly higher for the former. The least amount of decline from the original to the test data sets occurred for the exemplar benchmarks. In this case, the *d*′s improved by roughly 1/2 *SD* for the local strategies, and less so for the global strategies. In regard to our hypotheses, the data provided stronger support for the exemplar hypothesis than for the source hypothesis. That is, the content word approach performed worse on the local self-explanations than on the sentence-focused and global self-explanations, whereas the exemplar approach did equally well on each.

Specific reading strategies. We wanted to test whether LSA can be used to identify specific reading strategies embedded in self-explanations. The *d*′s are shown in Table 5. In regard to the accuracy of prediction using the test data, the mean *d*′s, collapsing over strategy for the content word, exemplar, and strategy approaches, were 1.11, .93, and 1.05, respectively. Averaging over the type of benchmark, the *d*′s were higher for paraphrase only (*M* = 1.33) and for the presence of a paraphrase (*M* = 1.30) than for bridging (*M* = .85) and elaborative (*M* = .64) inferences. Thus, identifying paraphrases was easier than detecting other strategies. The content word approach generally did the best, and the exemplar approach did the poorest in classifying bridging and elaboration inferences.

Table 4
***d*′s as a Function of Benchmark Type, General Reading Strategies, and Data Set**

		Data Set	
Benchmark	General Strategy	Original	Test
Content words	Sentence focused	1.62	1.13
	Local	.82	.12
	Global	1.46	1.20
Exemplar	Sentence focused	1.43	1.20
	Local	.76	1.14
	Global	1.09	1.26
Strategy	Sentence focused	1.46	.91
	Local	.32	.24
	Global	1.09	1.01

Note—Original data set refers to the data used to construct the discriminant functions, and test data refers to test cases.

DISCUSSION

The use of LSA to gauge comprehension requires the comparison of verbal input to stored text—a semantic benchmark. Consequently, investigators who wish to exploit the strengths of LSA must decide what type of text to use in their benchmarks. In this study, we have compared the utility of different types of semantic benchmarks in classifying the content of students' self-explanations. Our results indicate some differences for the content word, exemplar, and strategy benchmark approaches. The content words and exemplar benchmarks did better than the strategy benchmarks on classifying general reading strategies. The exemplar approach did much better than the content word approach in correctly classifying local strategies. One reason for this might be that we used multiple exemplars for each reading strategy. If the input was similar to any one of them, the average cosine for that strategy was increased. Multiple exemplars might also be advantageous when the words representing them are semantically related. We had noted in the introduction that the semantic similarity between the local and either the sentence-focused or the global self-explanations was higher than the similarity between the sentence-focused and the global self-explanations. Thus, the local explanations were in the "semantic middle," between sentence-focused and global self-explanations. The exemplars could have done better than the content words in identifying the local strategies because they provided examples of this semantic "middle," which is something that the content words could not do. The advantage of the exemplar approach for the local strategies also replicated the superior performance of the holistic approach reported by Foltz et al. (2000). Therefore, the exemplar approach might be a good alternative when the categories are semantically similar.

Of course, the choice of benchmark is also based on practical considerations. Ours was partly guided by the goal that iSTART would be able to adopt any number of passages for self-explanation training. Ideally, a teacher could select a passage and put it through a program that would construct suitable benchmarks. We are not there yet. Constructing the benchmark approaches that we used here required some expertise and existing protocols. For the content words, the prior text benchmarks required a causal analysis of the text, and the world knowledge benchmarks required a list of related content words not mentioned in the text. The exemplar approach required "good" examples, and the strategy approach required a time-consuming content analysis. Of the three types, the content word benchmarks were the easiest to construct. This was fortunate, because they performed equally well or

Table 5
d's as a Function of Benchmark Type, Data Set, and Reading Strategy

		Reading Strategy			
Benchmark	Data Set	Paraphrase Only	Paraphrase	Bridge	Elaboration
Content words	Original	1.46	1.14	1.25	.97
	Test	1.36	1.27	1.07	.74
Exemplar	Original	1.32	1.34	.89	1.06
	Test	1.32	1.34	.55	.53
Strategy	Original	.87	1.13	.92	1.01
	Test	1.33	1.30	.93	.66

better than the other approaches, except in identifying local self-explanations. iSTART can partially circumvent this limitation by focusing feedback to cases when the student is clearly not using the full gamut of strategies (sentence focused) or when they are clearly doing so (global). We are currently exploring ways in which the construction of benchmarks could be computerized and streamlined.

The present results indicate that LSA might be used to identify reading strategies expressed by self-explanations. This is an important finding for us because the type of feedback iSTART employs is partially driven by the classification of reading strategies. Magliano et al. (2002) showed that LSA cosines with the content word benchmarks are associated with different general reading strategies, but they did not test whether their results would generalize. We found here that although the d's declined slightly when the classifying functions were applied to a new data set, they were instructive when compared across the benchmarks, which was our primary goal. We have raised the overall accuracy of prediction of the test sentences by adding a few more predictor variables. In addition to the content word benchmarks, we added the LSA cosine between the self-explanation and (1) the title of the passage (as a measure of thematic processing), (2) words representing logical thinking (e.g., *therefore*, *because*), and (3) the number of words in the text sentence. Under these conditions, the d's for sentence-focused, local, and global self-explanations on the new data set increased to 1.67, .77, and 1.78, respectively. It should be noted, however, that the same 20 sentences used to generate the functions in the original data set were tested in the new data set. This raises the question of whether the functions would generalize to other sentences. In order to answer this question, we collected self-explanations to the Thunderstorm text from another group of 35 participants (from the same population as those who participated in the study reported here) who had undergone self-explanation training. Using the above predictors, the d's for sentence-focused, local, and global self-explanations were 1.55, .58., and 1.20, respectively. The magnitude of these data indicates that generalization should be better for sentence-focused explanations than for global strategies.

Because of the increasing use of LSA in language research and in computerized tools, future research is needed to identify and compare procedures that maximize classification on the basis of LSA cosines. We know of two general types of procedures. We used a least squares procedure in which a weighted linear combination of cosines is computed that maximizes prediction. The second is a pattern-based approach, in which the designer specifies a pattern and magnitude of cosines for the benchmarks as signifying a particular response. For example, the pattern [current sentence $>$ .70, prior text $<$ .20, world knowledge $<$ .10] might represent a sentence-focused response. We have tested this procedure with the thresholds based on the means and *SD*s of the cosines, with little success. One problem is that if one is not careful, many cases are left unclassified. The least squares approach provided a better (and easier) fit of the necessary parameters.

REFERENCES

Chi, M. T. H., de Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, **18**, 439-477.

Foltz, P. W., Gilliam, S., & Kendall, S. A. (2000). Supporting content-based feedback in online writing evaluations with LSA. *Interactive Learning Environments*, **8**, 111-129.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The Intelligent Essay Assessor: Applications to educational technology [Special issue]. *Interactive Multimedia Educational Journal*, **1**(2).

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, **101**, 371-395.

Graesser, A. C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Person, N., & the Tutoring Research Group (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments*, **8**, 129-147.

Kurby, C. A., Wiemer-Hastings, K., Ganduri, N., Magliano, J. P., Millis, K. K., & McNamara, D. S. (2003). Computerizing reading training: Evaluation of a latent semantic analysis space for science text. *Behavior Research Methods, Instruments, & Computers*, **35**, 244-250.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, **104**, 211-240.

Levinstein, I., McNamara, D. S., Boonthum, C., Pillaraisetti, P., & Yadivalli, K. (2003). Web-based intervention for higher-order reading skills. In *Proceedings of the ED-MEDIA '03 conference* (pp. 835-841). Honolulu.

Magliano, J. P., Trabasso, T., & Graesser, A. C. (1999). Strategic processes during comprehension. *Journal of Educational Psychology*, **91**, 615-629.

Magliano, J. P., 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. (2003). *SERT: Self-explanation reading training*. Manuscript submitted for publication.

McNamara, D. S., & Scott, J. L. (1999). Training reading strategies. In M. Hahn & S. C. Sontess (Eds.), *Proceedings of the Twenty-first Annual Meeting of the Cognitive Science Society* (pp. 387-392). Mahwah, NJ: Erlbaum.

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.

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.

Trabasso, T., & Magliano, J. P. (1996). Conscious understanding during text comprehension. *Discourse Processes*, **21**, 255-288.

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.

NOTES

[1] Instead of using the labels *sentence-focused*, *local*, and *global strategies*, Magliano et al. (2002) used the labels *minimalist*, *sentence focused*, and *knowledge building*. The definitions remain the same across studies, but the labels were changed because the current terms were more apt than the previous ones.

[2] We did not assess the participants' prior knowledge of the text topics. The familiarity of a text topic would most likely impact the reading strategy employed by a reader. By not taking topic familiarity into account, it is possible that we have underestimated the accuracy of prediction.

APPENDIX
Text of "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 a great amount of precipitation.
13. Usually, within an hour the amount and size of precipitation become 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 lightning 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.

(Manuscript received November 3, 2003;
revision accepted for publication April 4, 2004.)