Behavior Research Methods, Instruments, & Computers
2003, 35 (1), 22-31

Use of latent semantic analysis for predicting psychological phenomena: Two issues and proposed solutions

MICHAEL B. W. WOLFE
Grand Valley State University, Allendale, Michigan

and

SUSAN R. GOLDMAN
University of Illinois, Chicago, Illinois

Latent semantic analysis (LSA) is a computational model of human knowledge representation that approximates semantic relatedness judgments. Two issues are discussed that researchers must attend to when evaluating the utility of LSA for predicting psychological phenomena. First, the role of semantic relatedness in the psychological process of interest must be understood. LSA indices of similarity should then be derived from this theoretical understanding. Second, the knowledge base (semantic space) from which similarity indices are generated must contain "knowledge" that is appropriate to the task at hand. Proposed solutions are illustrated with data from an experiment in which LSA-based indices were generated from theoretical analysis of the processes involved in understanding two conflicting accounts of a historical event. These indices predict the complexity of subsequent student reasoning about the event, as well as hand-coded predictions generated from think-aloud protocols collected when students were reading the accounts of the event.

Over the past 10 years, researchers have shown that computational approaches to meaning have substantial contributions to make to both theoretical and practical issues (e.g., Burgess, Livesay, & Lund, 1998; Landauer & Dumais, 1997). In this paper, we focus on one such computational approach, latent semantic analysis (LSA; Landauer & Dumais, 1997), and examine several issues relating to the evaluation of its utility in the prediction of psychological phenomena. In particular, we discuss the application of LSA to predicting adolescents' reasoning about a historical event.

As a theory of knowledge representation in memory, LSA is based on the relatively simple notion that the similarity in meaning of two words can be induced from their usage in written text. By using this principle, simulated representations can be created for words and texts in a variety of specific domains, as well as for general information, through a computational procedure that initially examines co-occurrence frequencies in some set of printed texts and then induces an optimal structure of semantic relatedness, using a computational procedure called *singular value decomposition*. It is beyond the scope of this article to provide a detailed explanation of the LSA computational procedure, and we refer readers to descriptions in Landauer and Dumais (1997) and in Landauer, Foltz, and Laham (1998). What is important in this context is that LSA is a model of knowledge representation that is based on the patterns of word usage in printed documents.

LSA has proven useful in a variety of comprehension and text-processing situations. These include rating the quality of essays and summaries (Foltz, Gilliam, & Kendall, 2000; E. Kintsch et al., 2000), differentiating among texts on the basis of internal coherence (Foltz, Kintsch, & Landauer, 1998), tracing information in students' history essays to the printed text source(s) from which it came (Foltz, Britt, & Perfetti, 1996), and optimizing the match between readers' conceptual understanding of a topic, based on prior knowledge assessments, and the conceptual difficulty of to-be-read texts on that topic (Wolfe et al., 1998). The reliabilities of the LSA-based assessments have been as good as the reliabilities of human raters asked to make the same judgments. For human raters, these are exceedingly labor-intensive tasks—if they are doable at all.

Some researchers have reacted to the promise of LSA by capitalizing on its practicality and have used it to reduce labor-intensive analyses of other language samples, such as think-aloud protocols (Magliano, Millis, Wiemer-Hastings, & McNamara, 2001) and answers to questions that appear in tutorial dialogues (Graesser et al., 2000).

Portions of this work were presented at the 2000 meeting of the American Educational Research Association. The research reported in this paper was supported in part by a grant from the Spencer Foundation to the second author. The opinions expressed are the sole responsibility of the authors. Correspondence concerning this article should be addressed to M. B. W. Wolfe, 2224 ASH, Psychology Department, Grand Valley State University, Allendale, MI 49401 (e-mail: wolfem@gvsu.edu).

In this paper, we report on our efforts to predict the complexity of adolescent students' reasoning about a historical event from LSA-based indices of how they processed texts about that event. Our work brings to the fore several critical issues that researchers need to attend to in efforts to use LSA for either practical or theoretical purposes. Failure to attend to these issues can result in either inaccurate claims about LSA—both positive and negative—or invalid assessments of a research question.

To preview our argument, there are two major issues. First, researchers need to pay careful attention to the role that semantic relatedness plays in the text-processing or comprehension task under investigation. Each of the successful applications mentioned previously is based on the semantic relatedness principle: Two samples of text are more similar, more coherent, or more related the higher the overlap in their meaning is. Should semantic relatedness predict task performance? If so, what should that prediction look like? Answers to these questions determine the appropriate LSA indices and computational algorithms that should be used. Second, the basis on which LSA determines semantic relatedness is heavily dependent on the "knowledge" (semantic space) LSA has been given, as determined by the document set from which the LSA space has been derived. General knowledge spaces are developed from document sets on a wide variety of topics, about which there is only a small amount of information per topic. General knowledge spaces are akin to a person's general knowledge of the world. Such spaces may not be sensitive to fine distinctions among technical terms within specific domains, much as a lay person would be unfamiliar with distinctions between specific psychological terms, such as retroactive inhibition and forgetting. On the other hand, knowledge spaces that are highly specific may fail to pick up semantic relatedness that is expressed in nontechnical language. Just as Goldilocks was faced with the "just the right size" dilemma, efforts to use LSA must provide evidence of "just the right space."

The Importance of the Argument

Failure to explore the task and space issues can lead to inappropriate applications of LSA and rejection of LSA as a theory of knowledge representation. Indeed, Glenberg and Robertson's (2000) rejection of LSA provides a prime example of inadequate attention to both issues. Glenberg and Robertson examined LSA's ability to predict people's sensibility judgments of sentence pairs in which common objects were used in uncommon ways. For example, consider the sentence pairs 1A and 2A.

> 1A. Marissa forgot to bring her pillow on her camping trip. As a substitute for her pillow, she filled up an old sweater with *leaves*.
>
> 2A. Marissa forgot to bring her pillow on her camping trip. As a substitute for her pillow, she filled up an old sweater with *water*.

Glenberg and Robertson used a general knowledge space to compute LSA similarity ratings for the two sentences in each pair. The similarity ratings for the sentences in 1A and 2A did not differ. However, peoples' sensibility judgments were significantly higher for 1A than for 2A. Glenberg and Robertson took this result as an indication that LSA is not capable of representing the knowledge that is needed to understand the difference between these sentences.

We think their conclusion is premature. First, what is the relationship between a sensibility judgment and semantic relatedness? One of the interesting aspects of the situations Glenberg and Robertson (2000) used is that they reflect creative problem solving: Given the pragmatic constraints of the situation (being on a camping trip without a pillow), using a sweater filled with leaves as a pillow is an ingenious solution to the problem and one that requires a metaphorical interpretation of a sweater filled with leaves (e.g., *A sweater filled with leaves is like a pillow*). W. Kintsch (2001) has shown that algorithms beyond those that compute the overall similarity of one sentence to another are needed for LSA to account for language use in which the interpretation of arguments is dictated by the context, as in metaphor comprehension, causal reasoning, and some similarity judgments. Indeed, the predicate algorithm that W. Kintsch used, in essence, takes into account the substitutability of leaves or water for pillow, interpretations that are dictated by their functional roles in the sentences. But generating LSA representations of the sentences as a whole misses these important attributes of the predicate and arguments. In short, the more appropriate LSA index of relatedness would be based on a different calculation than the one used by Glenberg and Robertson. The predication algorithm may be a computational analogy of the human reasoning involved in sensibility judgments, reasoning that extends beyond semantic relatedness judgments.

In addition to the problem of the derivation of a meaningful semantic relatedness index that reflects sensibility, there is a problem related to the semantic space that Glenberg and Robertson (2000) used. They used a general knowledge space that may not have had enough "experience" with leaves and water in appropriate contexts to make the distinctions needed to distinguish between the sensibility of the two sentence pairs (Burgess, 2000).

A Strategy for Deriving Appropriate Applications of LSA to Text-Processing Research

The key to applying LSA is to understand the psychological processes of interest and how semantic relatedness might impact those processes. For example, if students are instructed to read a text and summarize it, we would expect a high degree of similarity between the meaning of the text and the students' summaries. On the other hand, if instead of summaries, students were asked to write essays applying what they had read to a new situation, the essays would be expected to be less similar than the summaries to the original text. In addition, we would expect less similarity across students, and we might not expect the level of similarity to predict essay quality. In other words, LSA may not be appropriate for

analyses of reasoning phenomena (for related discussions, see Burgess, 2000, and W. Kintsch, 2001), or at least, simple similarity indices may not be. In the example to follow, we show how we derived LSA-based indices that *were* predictive of students' reasoning. LSA, as with people, knows only about words and concepts to which it has been exposed (Burgess, 2000; Landauer & Dumais, 1997). The ability of LSA to match human semantic relatedness judgments is dependent on LSA's having exposure to texts comparable to ones human judgment makers would have had exposure to. This issue can be especially important when applying LSA to specific content domains. In the example that follows, we present one method for establishing that the semantic space contains sufficient information to make sensible judgments in a particular domain.

Using LSA to Predict Adolescents' Reasoning: An Example

From a substantive point of view, we were interested in whether the complexity of reasoning about a historical event could be predicted by the processing students had done of two contrasting accounts of the causes of the event. We had been addressing this issue using hand-coded, labor-intensive analyses of students' think-aloud protocols to describe processing (Coté, Goldman, & Saul, 1998). We pursued LSA as a substitute for the hand coding. To do so we compared the predictive ability of indices derived from the hand coding of processing with LSA indices of processing that we derived through a logical analysis of how semantic relatedness and coherence of processing should be related. Both the hand coding and the LSA-derived indices reflected measures of the kinds of processing that the students did when they were reading the two texts.

Previous research on comprehension reveals a complex relationship between the processing students do during text comprehension, the types of mental representations that students construct, and performance on subsequent comprehension and reasoning tasks. In general, processing that leads to more connected and coherent relationships among different ideas in text leads to better learning. Coté et al. (1998) found that students vary on several dimensions of processing and that these variations have important consequences in terms of how the information is represented in memory. One important dimension is the extent to which on-line processing tends to be confined to just the information in the sentence a student is currently processing, as compared with utilizing prior knowledge or making connections to other information in the text they are reading. In the former case, students tend to paraphrase each sentence, often not making connections to other knowledge they have that might be relevant. This type of processing tends to produce representations that closely duplicate just the information in the text sentences. Because there is little connection among the sentences or to prior knowledge, these representations are more often fragmentary rather than coherent. In contrast, students who integrate text information with relevant prior knowledge demonstrate better understanding of the text than do students who pay attention to what is said in the text, but not to how it can be understood in the context of their prior knowledge. Also, when students create connections between information in a sentence they are reading and information from earlier in the text, these cross-text connections lead to more coherent representations, especially when the students attempt to explain the relations of different parts of the text to each other or to establish causal links.

Researchers have also found that students create more integrated text representations when they generate explanations of the text information (Chi, 2000; Chi, deLeeuw, Chiu, & LaVancher, 1994; Coté & Goldman, 1999; Coté et al., 1998). Self-explaining is a process by which students generate new knowledge during processing, by connecting text information with relevant prior knowledge, connecting information from different parts of the text, or a combination of the two. Coté and Goldman (1999; Coté et al., 1998) found that when prior knowledge or other text sentences were used in self-explanations, students' representations were particularly well integrated.

A common method of tapping into on-line processing during reading is to have students provide think-aloud protocols during reading. Human raters then "hand" code the protocols with respect to the text. Indices derived from hand coding represent a fairly well-established and reliable means of characterizing processing (e.g., Chi, 1997; Coté et al., 1998; Trabasso & Magliano, 1996). The problem is that hand coding is a highly labor-intensive process. Our desire was to determine whether we could derive LSA indices that would require less labor-intensive analyses of the protocols but that would reflect the kind of processing known to be associated with more coherent and integrated representations. In the next sections, we first will describe the processing and reasoning data that were collected from adolescents, the hand coding of the protocols, and the relationship between the measures derived from hand coding and the subsequent reasoning data. We then will discuss our process for developing LSA indices for the processing of the texts and the relationship of these indices to reasoning performance.

Adolescents' Processing of and Reasoning About Historical Accounts

The processing and reasoning data that we consider here were collected from sixth-grade students who were asked to read two texts that provided opposing perspectives on the fall of the Roman empire. Full details of the study are available in Wolfe, Goldman, Mayfield, Meyerson, and Bloome (2000); here, we provide only a summary. The two perspectives agreed on some general information but disagreed as to the underlying cause of the barbarians' defeat of Rome. One text claimed that the Roman empire fell because the people became lazy, and the other text claimed that the empire fell because it became too big. Students thought out loud as they pro-

cessed the sentences of the texts. After reading the texts, the students answered several questions about their understanding of them and were then asked to provide their own explanation for the fall of the Roman empire: "If someone were to ask you why the Roman empire could not defend itself against the barbarian invasion, what would you say to that person?" From the think-aloud comments, we derived processing indices based on hand-coding methods and on LSA techniques that we devised as described below. These indices were used to predict the complexity of reasoning in the students' own explanations for the fall of Rome.

Hand-Coding the Processing

In order to assess student processing of the texts, we hand-coded the think-aloud data with respect to the content of the protocol statements, using the system of coding developed by Coté et al. (1998). These categories included paraphrases, in which the gist of the sentence was repeated without adding additional information, and elaborations, in which meaning was added to the information in the sentence. Other major categories that were coded included evaluative statements and statements indicating success or difficulty in comprehending the information. Elaborations were also coded as to the source of the information that was used and the type of reasoning that was displayed. The source of the information used in an elaboration could have been the students' prior knowledge, earlier parts of the text they were reading, or the first text they had read during the experiment (if they were reading the second text). With regard to the reasoning, of greatest interest were causal self-explanations, in which students created new knowledge by generating a causally based connection to some other information. The source of the information used in causal self-explanations could be any of the types mentioned above.

To evaluate student reasoning about the information in the texts, we scored the student explanations for why the Romans could not defend themselves against the barbarian invasion. The reasoning score was made up of three elements: the number of stated causes for the fall of Rome, the extent to which causes were elaborated or explained, and whether multiple causes were integrated with each other or merely presented as separate causes. The scores ranged from zero to five.

We used the hand codes of the think-aloud data to predict the reasoning the students engaged in after reading the two accounts of the fall of Rome. Simple correlations of each of the coding categories with the reasoning scores indicated that three of them were significant predictors of the reasoning: causal self-explanations [$r(43) = .37$, $p = .01$], elaborations that connected to the same text as the sentence being read [$r(43) = .31$, $p = .04$], and elaborations that connected to the information in the previous text that had been read [$r(43) = .38$, $p = .01$]. No other coding categories were significantly correlated with the reasoning scores. Multiple regression analyses indicated that, taken together, these three indices accounted for 21% of the variance in the reasoning scores. Table 1 shows the variance accounted for (R^2) by each of these indices of processing and the combined model. These findings are consistent with previous research on relationships between processing and understanding (Chi et al., 1994) or recall (Coté et al., 1998; Goldman, Coté, & Saul, 1995).

These results, more completely reported in Wolfe et al. (2000), indicate that readers who actively engage in processing text information are more likely to construct representations of the information that support complex reasoning about the information. These active processes include elaborating on text information by making connections to relevant world knowledge and by generating

Table 1
Proportion of Variance in Reasoning Scores Accounted for (R^2) by Single and Multiple Predictor Models of Hand-Coded Data and Latent Semantic Analysis (LSA) Data

Indices	R^2
Hand Codes	
Individual factors	
Causal self-explanation	.13
Same text connection	.10
Previous text connection	.14
Multiple regression	
Causal self-explanation + same text connection + previous text connection	.21
LSA	
Individual factors	
Sentence.protocol (linear + quadratic)	.18
Opposite text	.16
Multiple regression	
Sentence.protocol (linear + quadratic) + opposite text	.26

Note—"Sentence.protocol" is the average cosine for each student between text sentences and their corresponding protocol statements. "Opposite text" is the cosine between the collection of protocol statements made in response to a text and the text to which the statements do not refer.

connections to other information in the set of texts from which the student is learning. In the next section, we will present an example of the use of LSA to account for these complex reasoning processes by using LSA analyses of the same think-aloud data to predict student reasoning. In applying LSA to the think-aloud data, our goal is to develop an LSA-based method to also tap into the underlying processes that support reasoning based on the information. Our goal, therefore, is not to account for the protocol codes themselves but, rather, to predict the same reasoning data that the protocol codes predict.

LSA Predictions of Reasoning From Multiple Texts

Developing an LSA-based method for accounting for the students' reasoning exemplifies efforts to deal with the two issues, discussed in the introduction, that arise when applying LSA to new tasks and topics. What LSA index reflects the kind of processing that the behavioral data reflect? Is the semantic space "just right" to provide reasonable estimates of semantic relatedness or similarity? First, we will address the "just right" dilemma, to establish and verify that the LSA space contained enough relevant text to be able to make reasonable semantic similarity judgments on the topic. Second, we will discuss the derivation of LSA-based indices that were appropriate to the task of accounting for the reasoning.

Establishing a valid LSA space. The goal in creating a semantic space for a specific application of LSA is simply to use a representative set of texts as input so that LSA may make semantic similarity judgments that are comparable to those a human would make in the same situation. In applications that do not involve specialized topic knowledge, researchers can utilize the general knowledge space available on the LSA Web site (http://lsa.colorado.edu). For more content-specific applications, however, the general knowledge space may not have enough texts about a particular topic to make sensible comparisons. In that case, it is necessary to construct a content-specific semantic space. It is important to realize that the use of a specific semantic space for specific content areas does not represent a theoretical or methodological shortcoming of LSA. It is simply an analogue of the situation with humans; if one is asked to make semantic judgments in a content area, that person is likely to read up on the topic, rather than perform the task with no particular expertise on the topic. For the fall of Rome topic, we constructed a semantic space, using age-appropriate materials related to ancient civilizations. Text documents were gathered from a sixth-grade textbook on ancient history. The specific civilizations covered were ancient Rome, Egypt, Greece, China, India, and the Middle East. In addition, texts from a CD-ROM encyclopedia on the same civilizations were collected. Each paragraph of text represented a separate document. A total of 1,534 documents with 8,352 unique words were used in creating the space.

The simplest way to establish the validity of a space is to take two text samples that, a priori, ought to be highly similar and test whether LSA does indeed provide cosines that indicate high similarity. At the same time, it is important to show that for two texts that, a priori, ought not be very similar, LSA provides cosines that indicate low similarity. In the present context, we reasoned that think-aloud statements that were coded as paraphrases of the text ought to be judged as highly similar to the presented text, whereas elaborations ought to be less similar. This is so because the meaning of a paraphrase of a sentence is similar or identical to the meaning of the sentence itself. For example, after the sentence *These places were very far away from Rome, the center of the empire*, was read, a student response that was coded as a paraphrase was "Rome was the center of the empire, and the places were far away from them." Comparing the text statement and the student's statement in the Ancient Civilization space produced a cosine of .84, indicating a high degree of semantic relatedness. This is just what we would expect if the semantic space was appropriate. But it is possible that the high cosine was the result of overlap on just a few of the words. What we also needed to demonstrate was that statements with a greater difference in overall meaning from the text statement were in fact assessed to have lower similarity, even if the statements shared some words in common. In the present context, a student's think-aloud response that was coded as an elaboration was less similar than the one coded as a paraphrase. Indeed, a response coded as an elaboration, "So they were afraid that something would happen because it was so far away or something," produced a cosine of .30. For this one sentence, the semantic space seemed to be "just right": LSA correctly assessed the semantic relatedness of the paraphrase statement to be higher than that of the elaboration. To determine that this was not an effect unique to the one sentence, we needed to look at all of the sentences and determine whether similarity was higher for paraphrases than for elaboration events in the corpus as a whole (all students, both texts). To do this, we calculated the text sentence to protocol statement similarity (cosine) for each pair and averaged across both texts. For students who produced a large number of paraphrases, the average cosine (similarity) should be higher than that for those who produced a low number of paraphrases.

The data in Figure 1 show that this is exactly the result we obtained. The correlation of average LSA similarity was quite high (see Table 2). Conversely for elaborations, we would expect a much weaker or no relationship between the frequency of elaborations and semantic similarity, because students may have been bringing a wide variety of information to bear that varies in semantic similarity to the text information. As a result, across the texts we would not expect a significant correlation between the number of elaborations and average similarity. Indeed, Figure 2 shows that there was no significant relationship.

These results show that, indeed, we have an LSA space that yields similarity values that make sense, given process differences in protocol statements. The results validate the use of our Ancient Civilization semantic

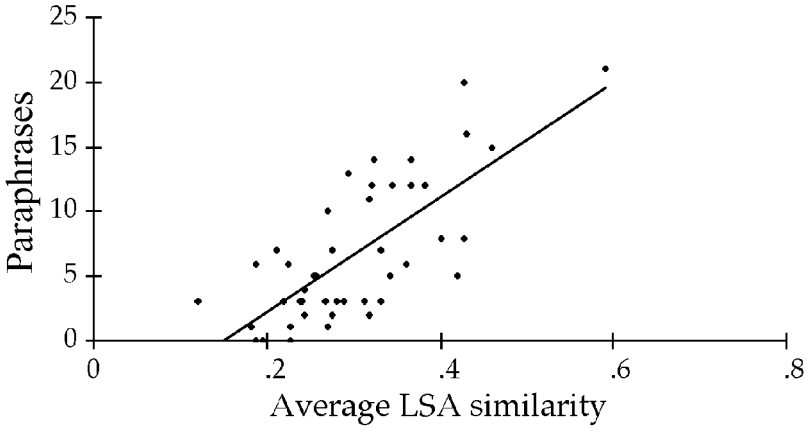

Figure 1. Average latent semantic analysis (LSA) similarity of each text sentence to its corresponding protocol statement correlated with number of paraphrases for each student.

space as being reasonable for distinguishing the meaning of think-aloud statements.

There was one final step in the "just right " validation process. We performed these same analyses, using a general knowledge space available on the LSA Web site (the TASA text corpus) and using a space constructed from documents dealing solely with the Roman empire. The results are presented in Table 2. The Ancient Civilization space provided the best discrimination among students with respect to the number of paraphrases; the three spaces produced similar data for the elaborations. Therefore, we concluded that the Ancient Civilization space was the most valid for our purposes. In a different content area (psychology), Shapiro and McNamara (2000) also concluded that a specific content space provided better discrimination than did the general knowledge space. Having validated the space, we turn to our primary goal of accounting for the reasoning students did when explaining the fall of Rome.

Using LSA to predict reasoning. Deriving the LSA indices that would be appropriate predictors of the reasoning scores requires a detailed understanding of the relationship between processing, representation(s) formed on the basis of the processing, and subsequent reasoning, which is presumably based (at least to some degree) on the representation formed during text processing. As was discussed in the introduction, previous research indicates that more coherent representations result from processing that connects ideas in text to one another and to prior knowledge, especially if those connections are elaborative and causal (Chi et al., 1994; Coté & Goldman, 1999; Coté et al., 1998). At the same time, paraphrases and associations to single sentences do not tend to produce coherent representations of text. More coherent representations of text are associated with better learning. Indeed, the hand-coded data bore out this relationship in the case of the adolescents' reasoning about the causes of the fall of Rome. A simple translation of this relationship into the LSA world would lead to the expectation that comparisons of text sentences with protocol statements should yield average cosines that would be negatively correlated with reasoning scores: The higher the similarity is, the more paraphrases (and fewer elaborations) there are; hence, the lower the reasoning scores are.

However, the situation is more complex than that, and the simple translation algorithm is flawed. At one extreme, we expect that students who tend to process sentences by primarily considering the information in the immediate sentence will not create representations that are integrated with prior knowledge or with the other text they read. This is the tendency for students who paraphrase each sentence, largely in isolation from the others. They may have a good understanding of the ideas conveyed in a single sentence, but not how those ideas relate to other sentences in the text they are reading or to other texts they have read. As a result, we predict that students with a high degree of relatedness between protocol comments and text sentences will produce fewer causes and be less likely to produce integrated explanations for the causes.

At the other extreme are students who provide protocol comments that are *very different* from the meaning of

Table 2
Correlations Between the Average Latent Semantic Analysis (LSA) Similarity of Each Text Sentence to Its Corresponding Protocol Statement and the Number of Paraphrase or Elaboration Statements for Each Student

Semantic Space	Paraphrases	Elaborations
Ancient Rome	.63*	−.02
Ancient Civilization	.74*	−.14
TASA–All (general)	.71*	−.22

Note—LSA cosines in the Ancient Rome space were generated using 100 dimensions, which provided the best correlations. *$p < .0001$.

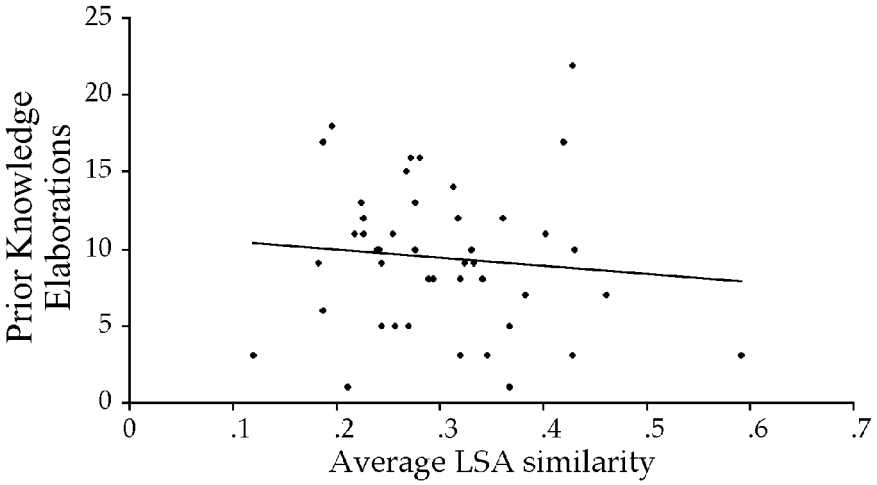

Figure 2. Average latent semantic analysis (LSA) similarity of each text sentence to its corresponding protocol statement correlated with number of prior knowledge elaborations for each student.

the text sentences: They bring in associated but largely irrelevant knowledge (e.g., personal experiences, such as "I went to Rome once"). These students make use of prior knowledge, but in ways that are not helpful in creating a good representation of the text as a whole; nor does this knowledge establish relationships between ideas presented in the text. Protocol statements of this type would have low degrees of similarity to text sentences, because students are not basing their processing on what is presented in the text and are not relating different ideas from the text. This contrasts with causal self-explanations and elaborations that connect to other ideas from the text. Students at this *unrelated processing* extreme would also be expected to have relatively lower reasoning scores on their responses to why Rome fell than would those who elaborate.

Finally, consistent with the prior literature, we would expect more complex reasoning about why Rome fell from those students who engaged in elaborative processing that used relevant prior knowledge and generated appropriate connections to other sentences or elaborative comments. This processing pattern should produce intermediate levels of similarity between each text sentence and protocol comment. Many self-explanations will be in this intermediate range of semantic similarity. In other words, relevant elaborations build on the semantic content of the text and, thus, should yield higher similarity than the *unrelated processing* extreme, but lower similarity than the *paraphrase processing* extreme.

With this in mind, we calculated the cosine between text sentence and protocol comment pairs and constructed an average for each student over both texts. This average LSA similarity index indicates the extent to which students tend to add meaning to the sentences they process, with similarity decreasing the more they add and the more unrelated it is. We then used the average similarity scores to predict the reasoning scores. The results are shown in Figure 3. As was predicted by our foregoing analysis of the relationship between processing and reasoning, the nonlinear relationship was significant [$F(1,41) = 9.22$, $p = .004$], resulting in 18% of the variability in reasoning scores being accounted for (see Table 1). The fit of the nonlinear trend line is shown in Figure 3. Students whose processing of sentences involved elaborating upon the meaning of the text information in terms of causal relationships and relevant prior knowledge had higher reasoning scores. Students whose processing was either too similar to the meaning of the texts or too different from the meaning of the texts had lower reasoning scores.

We derived another LSA-based index that captures the tendency to connect ideas across texts. More connections across texts suggests that a student is generating connections between arguments in the two texts during comprehension. In other words, they are comparing and contrasting the two different explanations for the fall of Rome. The LSA index of cross-connections compares the set of protocol statements from one text with the set of text sentences in the other text. These LSA similarity scores indicate the extent to which protocol comments made while one text is read are semantically related (connected) to the other text.[1] For each student, two *opposite-text* cosines were computed, and the average was taken. These average opposite-text cosines were then correlated with the reasoning scores. The data are presented in Figure 4. Students whose think-aloud comments had a higher degree of semantic relatedness to the text they were not currently reading had more complex reasoning scores [$r(43) = .40$, $p = .007$]. This factor accounts for 16% of the variability in the complexity scores, a correlation about equal to the correlation of any of the processing indices (see Table 1).

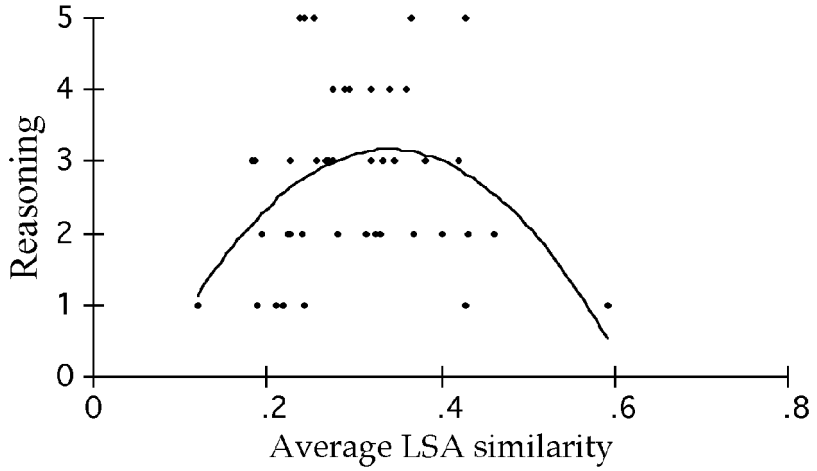

Figure 3. Average latent semantic analysis (LSA) similarity of each text sentence to its corresponding protocol statement, by student reasoning score.

As with the hand-coded indices, we conducted multiple regression analyses to determine the independent contribution of the two LSA indices and the total variability accounted for by the two. The extent to which students elaborated on the meaning of the sentences they read (the elaborative processing measure, as reflected in the linear and quadratic components shown in Figure 3) was used along with the opposite-text similarity measure in a multiple regression analysis to predict reasoning scores. The results of this analysis indicate that the model as a whole accounts for 26% of the variability in reasoning scores [$F(3,40) = 4.72$, $p = .007$], which is more than either measure individually. Thus, the LSA-based indices, as compared with the hand-coded ones, account for a comparable (or slightly higher) proportion of the variance in students' reasoning.

Discussion

The quality of student reasoning about contradictory history texts was predicted by LSA (Landauer & Dumais, 1997) similarity indices that strategically compared student think-aloud data with the content of the texts the students processed. These LSA similarity indices predicted the students' reasoning slightly better than did hand-coded processing indices of the same think-aloud data. The process of generating these LSA-based indices illustrates two critical points researchers need to attend to when applying LSA to psychological processing issues.

First, the LSA indices need to be derived from a thorough understanding of the role that semantic relatedness plays in the psychological process of interest. In the present experiment, two LSA indices were generated from conclusions that were derived from extant empiri-

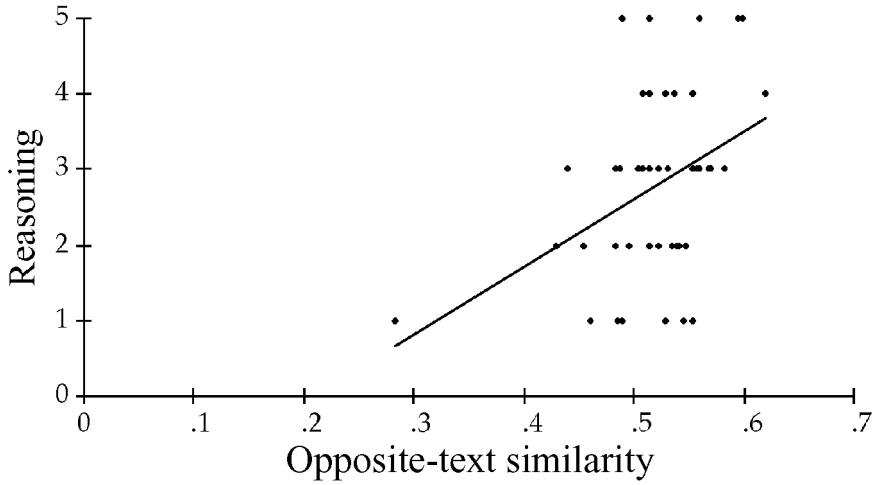

Figure 4. Average latent semantic analysis (LSA) similarity of protocol statements from one text compared with the opposite text, correlated with student reasoning score. The relationship remains significant with the removal of the apparent outlier.

cal literature and theoretical frameworks that have elucidated the relationship between comprehension processes, representation, and reasoning. In one index, the similarity of a text sentence to its corresponding think-aloud statement was predictive of student reasoning in a curvilinear fashion, with intermediate levels of similarity corresponding to the highest reasoning scores. It is important to note that this curvilinear similarity function does not have a direct analogue in the hand-coding scheme. Rather, it was derived from, and was consistent with, conclusions about psychological processing that arose from previous literature. The other LSA-based index, the extent to which think-aloud comments from one text were similar to the content of the *opposite* text, was also predictive of student reasoning. This index was derived from a consideration of how students might process the two texts in order to evaluate the relative merit of the two historians' arguments. A greater amount of comparative processing was hypothesized to lead to more complex reasoning in the students' own explanations for the fall of Rome.

The two LSA-based indices were also combined in multiple regression analyses. The derivation of these LSA-based indices did not require the laborious process of hand coding, and yet they accounted for slightly more variability in student reasoning than did the optimal combination of hand-coded categories. This result is interesting because coding of think-aloud data with the categories used here, or qualitatively similar categories, has been utilized by many researchers with excellent results in terms of characterizing the processing that takes place during comprehension and predicting memory and reasoning. The LSA advantage, albeit slight, may arise from the fact that semantic similarity scores are continuous, whereas coding categories are discrete.

The second issue that researchers need to attend to is that, for LSA to generate reasonable semantic relatedness judgments, the *semantic space* that is utilized must be created from a set of documents that approximates the knowledge that a human would need in order to make judgments under the same circumstances. For some applications, researchers wish to approximate the knowledge that the average novice would have about a topic. In such cases, it is appropriate to use a *general knowledge* semantic space. In other circumstances, the similarity judgments should approximate those of people who have relatively specialized knowledge about a topic, in which case it is more appropriate to use a *specialized* semantic space. In the present study, we wished to simulate judgments that would be made by someone with a reasonable amount of knowledge about ancient civilizations, so we created a specific space with documents relating to that topic. This was the most appropriate semantic space, because all of the participating adolescents had studied ancient Rome within the 4-month period prior to the study of reasoning. In order to evaluate the appropriateness of this semantic space, the text sentence to think-aloud statement similarity scores were correlated with both the extent to which the subjects tended to paraphrase sentences and the extent to which the subjects tended to elaborate on sentences. The correlations with paraphrases were quite high, whereas there was no correlation with the elaborations. In addition, our specialized Ancient Civilization space distinguished between these processing indices more effectively than did a general knowledge space. This particular method serves as an example of a means by which a semantic space can be evaluated. For other LSA applications, researchers need to devise tests specific to those applications in which LSA similarity indices are compared with judgments that would be made by humans who possess approximately the same knowledge, necessary to perform the task of interest.

REFERENCES

Burgess, C. (2000). Theory and operational definitions in computational memory models: A response to Glenberg and Robertson. *Journal of Memory & Language*, **43**, 402-408.

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in context space: Words, sentences, discourse. *Discourse Processes*, **25**, 211-257.

Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *Journal of the Learning Sciences*, **6**, 271-315.

Chi, M. T. H. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology* (pp. 161-238). Mahwah, NJ: Erlbaum.

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

Coté, N., & Goldman, S. R. (1999). Building representations of informational text: Evidence from children's think-aloud protocols. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 169-193). Mahwah, NJ: Erlbaum.

Coté, N., Goldman, S. R., & Saul, E. U. (1998). Students making sense of informational text: Relations between processing and representation. *Discourse Processes*, **25**, 1-53.

Foltz, P. W., Britt, M. A., & Perfetti, C. A. (1996). Reasoning from multiple texts: An automatic analysis of readers' situation models. In G. W. Cottrell (Ed.), *Proceedings of the 18th Annual Cognitive Science Conference* (pp. 105-115). Mahwah, NJ: Erlbaum.

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

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, **25**, 285-307.

Glenberg, A. M., & Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory & Language*, **43**, 379-401.

Goldman, S. R., Coté, N. C., & Saul, E. U. (1995). Paragraphing, reader, and task effects on discourse comprehension. *Discourse Processes*, **20**, 273-305.

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.

Kintsch, E., Steinhart, D., Stahl, G., LSA Research Group, Matthews, C., & Lamb, R. (2000). Developing summarization skills through the use of LSA-based feedback. *Interactive Learning Environments*, **8**, 87-109.

Kintsch, W. (2001). Predication. *Cognitive Science*, **25**, 173-202.

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.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, **25**, 259-284.

Magliano, J., Millis, K., Wiemer-Hastings, K., & McNamara, D. (2001, July). *Using LSA to reveal reader strategies*. Paper presented at

the conference of the Society for Text and Discourse, Santa Barbara, CA.

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 comprehension. *Discourse Processes*, **21**, 255-287.

Wolfe, M. B. W., Goldman, S. R., Mayfield, C., Meyerson, P. M., & Bloome, D. M. (2000, July). *Middle school students' processing of multiple accounts of an historical event*. Paper presented at the meeting of the Society for Text and Discourse, Lyon, France.

Wolfe, M. B. W., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). Learning from text: Matching readers and texts by latent semantic analysis. *Discourse Processes*, **25**, 309-336.

NOTE

1. The order in which the texts were read was not considered in this analysis, because each text makes explicit reference to the theory put forth in the other text. As a result, students know at the beginning of the first text that there are opposing theories that will be addressed and what the theories are. Thus, the students could integrate comments about the theory put forth in the second text while they processed the first text, in addition to commenting on the theory put forth in the first text while they processed the second text.

(Manuscript received February 22, 2002;
revision accepted for publication September 14, 2002.)