Running head: LEARNING FROM TEXT

# Learning from text:
# Matching readers and texts by Latent Semantic Analysis

Michael B. W. Wolfe, M. E. Schreiner, Bob Rehder, Darrell Laham,
University of Colorado, Boulder
Peter W. Foltz
New Mexico State University
Walter Kintsch, and Thomas K Landauer
University of Colorado, Boulder

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

**Correspondence to:**
Michael B.W. Wolfe
Campus Box 345 - Cognitive
University of Colorado
Boulder, CO 80309
(303) 492-7299
Email: wolfem@psych.colorado.edu

Abstract

This study examines the hypothesis that the ability of a reader to learn from text depends on the match between the background knowledge of the reader and the difficulty of the text information. Latent Semantic Analysis (LSA), a statistical technique that represents the content of a document as a vector in high dimensional semantic space based on a large text corpus, is used to predict how much readers will learn from texts based on the estimated conceptual match between their topic knowledge and the text information. Participants completed tests to assess their knowledge of the human heart and circulatory system, then read one of four texts that ranged in difficulty from elementary to medical school level, then completed the tests again. Results show a non-monotonic relationship in which learning was greatest for texts that were neither too easy nor too difficult. LSA proved as effective at predicting learning from these texts as traditional knowledge assessment measures. For these texts, optimal assignment of text on the basis of either pre-reading measure would have increased the amount learned significantly.

# Learning from text: Matching readers and texts by Latent Semantic Analysis

Much of what we learn, as students and throughout life, we learn from reading. Learning from text, however, is not the same as remembering the text. Kintsch (1994) argued that a central feature of learning from text is linking up the textual information with prior knowledge. The new information must be integrated with prior knowledge both for current comprehension and for later use in new situations. Thus, learning presupposes suitable prior knowledge to which the to-be-learned information can be linked. If there is no relevant knowledge basis at all, this integration cannot take place - learning fails. Although texts can be memorized, they will remain isolated memory episodes, inert knowledge that is not available for new tasks.

Texts that are too distant from a reader's knowledge base are, therefore, not suitable as instructional texts. Trivially, texts that contain only information already known to the student are also useless for learning. Hence there exists an intermediate zone-of-learnability, where texts are just distant enough from what the student already knows but are not too remote. To optimize learning, readers - like Goldilocks, when she found herself in the house of the Three Bears - need texts that are neither too easy nor too hard, but instead are just right. We test this “Goldilocks” hypothesis and describe a practical method for determining zones of learnability for different readers and texts.

Ample demonstrations exist that background knowledge facilitates learning from text (e.g., McKeown, Beck, Sinatra, & Loxterman, 1992; Means & Voss, 1985; Moravcsik & Kintsch, 1993;

Schneider, Körkel, & Weinert, 1990; Spilich, Vesonder, Chiesi, & Voss, 1979). These studies frequently involve having a group of high-knowledge and low knowledge participants in some domain read a text. Results typically show that the high-knowledge participants can remember more text information and write better summaries than the low-knowledge participants. Text comprehension can be considerably enhanced by rewriting texts for which students do not have adequate preparation, making them easier and more explicit (Britton & Gulgoz, 1991). But texts can be too easy for students with more than adequate prior knowledge, and making the texts more challenging can improve learning outcomes for such students (McNamara, E. Kintsch, Songer, & W. Kintsch, 1996; Voss & Silfies, 1996). For optimal results, instructional texts at a particular level of difficulty ought to be matched with students at a particular level of background knowledge.

Teachers try to intuitively choose instructional texts that are suitable for their students, and in a well developed curriculum, they probably succeed quite well, at least on average. However, students sometimes face the task of selecting optimal instructional texts without the help of a teacher, perhaps because the teacher does not know enough about the domain in question; because the teacher does not know enough about the student's knowledge; or perhaps because no teacher is available, as in choosing texts from the World Wide Web. Other times, teachers or curriculum planners might theoretically profit from computer-based aids that would help to identify appropriate textual materials among large bodies of previously uncalibrated resources. Crude, global reading level measures leave much to be desired for such purposes, especially according to the present hypothesis, because they do not measure the conceptual overlaps

between particular student knowledge and particular text content, but only general vocabulary and syntactic complexity. Thus, a method of text selection based on theoretical principles would be desirable. Kintsch (1994) has described such a method, based on determining the links between a student's knowledge representation and the propositional representation of a text. While important theoretically, this method is totally impracticable. In this article, we show that Latent Semantic Analysis (LSA) can provide a practical alternative for matching students differing in background knowledge with instructional texts of varying topical sophistication.

## DESIGN OF THE STUDY

As a knowledge domain, we selected the functioning of the human heart. This is a topic about which college students have some background knowledge and in which large variations in background knowledge might be expected. As instructional texts we chose four texts, each of which was 17 paragraphs long. These texts varied widely in sophistication. Text A, the easiest text, was a chapter from a children's book about the human body. Text B was taken from an adult-level general introduction to the human circulatory system. Text C was written for undergraduate introductory anatomy students. Text D, the most difficult text, came from a medical school text and contained information a lay person couldn't ordinarily be expected to know. Each participant was randomly assigned to one of these texts and the amount of learning was determined for each student by comparing performance on pre- and post reading knowledge assessment tasks designed to reflect any knowledge gained by reading. Knowledge was assessed in two ways: by means of a short-answer questionnaire, and by means of an essay the student wrote on the functioning of the heart.

One question of interest was to what extent these two different methods would yield comparable results? The post-reading knowledge assessment was done either immediately after reading or 2 days later.

We had three primary goals:

(1) To see how well we can measure knowledge in a domain.

(2) To determine empirically how variations in the students' background knowledge, as assessed by conventional methods, affected their ability to learn from the different texts.

(3) To determine how well LSA was able measure background knowledge and predict learning.

## LATENT SEMANTIC ANALYSIS

LSA is a fully automatic computational technique for representing the meaning (content) of a text as a vector in a high-dimensional semantic space. The rationale and method for LSA have been described in Deerwester, Dumais, Furnas, Landauer, & Harshman, (1990) and Landauer & Dumais (1996) and are summarized in Landauer, Foltz & Laham (1998/this issue). The first step in employing LSA is to construct a semantic space based on a large body of relevant text. In this case, the space was constructed from all articles in a student encyclopedia that had anything to do with the functioning of the heart. There were 36 such articles, containing a total of 17,880 word tokens and 3,034 unique word types. A semantic space of 100 dimensions was constructed from this input material.

The instructional texts themselves, as well as the pre- and post reading essays each student wrote on the functioning of the heart, could then be represented as vectors in this space. The vector for a text is simply the vector sum or average of the vectors of its component words. The cosine between any two vectors provides a measure of their

semantic relatedness and was the measure of primary interest in the present study. Specifically, we hypothesized that if the cosine between a student's pre-reading essay on the functioning of the heart and the instructional text were too low, insufficient links would exist between the student's knowledge and the text, and reading would produce relatively little learning. If the cosine were too high, the student would already know the contents of the text, and reading would be ineffective as well. Learning should be greatest for intermediate values of the cosine measure. Does this cosine measure (which is easy and objective to compute) predict learning as well as more traditional empirical measures (in this case, a short-answer questionnaire, which is expensive to develop, and an essay, which is difficult to score)?

## EXPECTED RESULTS

According to the zone-of-learnability hypothesis, the amount learned from a text should increase as a function of prior knowledge up to a point and then decrease, and the zone of learnability should be at a lower level of prior knowledge for an easy text than for a more difficult one, approximately as in Figure 1. For concreteness, we have assumed in Figure 1 that the shape of the function relating prior knowledge and efficiency of learning is approximately Gaussian.

Domain knowledge was measured in three ways: as the score a student received on the questionnaire prior to reading the text (pre-questionnaire), the grade on the essay written prior to reading (pre-essay), and the cosine between the LSA representations of the student's prior essay and a standard college-level textbook chapter on the functioning of the heart (cosine pre-essay.standard). We chose Text C as a standard text, since it appeared to be most representative for the heart knowledge of college students, though Text B could have been used

equally well, with very similar results (Text A was too easy, Text D too difficult, which resulted in the range of cosines being more restricted).

Amount of learning was operationally defined in two ways: as the proportion of possible improvement in the scores on the questionnaire from before to after reading (learn-questionnaire), and as the proportion of possible improvement in the grades the student's essays received before and after reading (learn-essay).

Another way of looking at the relationship between background knowledge and learning efficiency is to plot the latter as a function of how closely related a particular student's knowledge is to the text he or she actually read. The LSA cosine between the essay a student wrote and the text he or she read provided such a measure (cos pre-essay.textread). The expected relationship between learning efficiency and the relatedness between a student's prior knowledge and an instructional text might look somewhat like Figure 2.

There are two major factors that might distort the idealized function shown in Figure 2. First, the match between a student's background knowledge and the instructional text is only one factor that determines learning; it limits what can be learned, but does not guarantee that all of what could be learned actually will be learned. Motivation is another obvious variable, for instance, that might be responsible for a student's sub-optimal performance. Second, the idealized function of Figure 2 will be obtained only if students are distributed over the whole range of background knowledge. However, it is likely that for a given text, most or all students would fall into a relatively narrow band on the prior knowledge scale. Hence, The empirical relationships for different texts could take many forms. For example, the relationship could be nonmonotonic (e.g., a negative

quadratic) if the range of students corresponded to the range A shown in Figure 2. The relationship could be monotonically increasing (range B in Figure 2) or decreasing (range C), if the text was either too hard or too easy for most of the students. Or there could be no relationship at all, for instance if a text is much too hard (range D) or much too easy (range E) for all the students, or if all students tend to be sampled from the top of the curve (range F). There could also be positive quadratic relationships, such that most students know too little to learn anything, but the very best ones fall within the zone of learnability (range G), or most students know too much to learn anything, but the least knowledgeable ones fall within the zone (range H). Because of the variety of empirical relationships that may be obtained between prior knowledge and learning, it may appear that any results are consistent with our zone-of- learnability hypothesis. However, because the four instructional texts employed in this study were ordered according to difficulty, (both intuitively and by measures reported here and in Rehder et al, 1998/this issue) the variety of possible empirical relationships across all the texts between our measures of prior knowledge and learning is considerably constrained. Specifically, as the texts move from easiest to hardest, the relatedness between prior knowledge and text as described in Figure 2 should decrease. As a result of the decrease in relatedness with increasing text difficulty, we should find learning functions that correspond to the shift in relatedness. Moreover, observation of any single range with a convincing non-monotonicity will serve to substantiate both the theory and methods of measurement and analysis.

## METHOD

### Participants

One hundred six people participated in this experiment. Ninety four were undergraduate students at the University of Colorado, Boulder, and 12 were in their 2nd, 3rd, or 4th year of medical school at the University of Colorado. For pragmatic reasons, forty seven of the undergraduates received course credit for Introductory Psychology and forty seven were paid $25. The paid participants responded to fliers posted around the campus. This manipulation was not of interest and did not interact with any variables of interest, and thus will not be discussed further. The medical students were paid $50.

Materials

There were four different heart texts used in this experiment, as well as several different tasks that the participants were asked to perform. All of the materials used will be described in turn.

Essay task. A one paragraph instruction sheet explained that participants were to write an essay of approximately 250 words. The topic of the essay was the anatomy, function, and purpose of the human heart and circulatory system (see Appendix for verbatim instructions). The instruction sheet also stated that participants could take as much time as they needed to complete this task.

Card sorting task. Twenty words or phrases associated with the human heart were written, one on each of a set of standard index cards. Some examples of the card labels are: right ventricle, valves, oxygen (O2), and capillaries (see Appendix for the complete list of words).

General knowledge test. Seventeen questions about the anatomy, function and purpose of the human heart and circulatory system were written in a three page questionnaire. Written instructions for the general knowledge questionnaire appeared at the top of the first page. The instructions stated that participants should

answer each question in complete sentences, including as much information as they could. The questions increased in their specificity and complexity through the questionnaire. Thus, the answers to some of the earlier-appearing questions could be found in the latter-appearing questions. For this reason, the order of the question presentation could not be counterbalanced and all participants saw the questions in the same order. Therefore, the instructions stated that participants should respond to the questions in order, to neither jump ahead to questions nor return to questions once they had responded to them. There were 17 questions on the general knowledge questionnaire. The correct answer to some of the questions required only a single piece of information, and were thus worth only one point (e.g. "Blood returning from the body enters which chamber of the heart first?"; Answer: the right atrium). Correct responses to other questions required more than one piece of information, and each correct piece of information was scored independently. Thus, such questions were worth more than one point (e.g. "What is the protein which makes quick oxygen/carbon-dioxide transfer possible? How many molecules of oxygen can each such protein carry?"; Answer: hemoglobin, 4 molecules). In all 17 questions there were a total of 40 possible points.

The general knowledge test was developed to test the participants' general heart knowledge. The questions were selected by first gathering several texts on the heart and circulatory system. From these texts, common topics were identified, and a larger set of 39 questions was developed. Twenty six people with varying levels of background knowledge were pre-tested on this larger set of questions. On the basis of this distribution of responses, 17 of the 39 questions were intuitively identified as both comprehensive in evaluating

general heart knowledge and discriminating in identifying the responder's knowledge level. These 17 questions constitute the general knowledge test.

Cued passage recall questionnaire. Four cued passage recall sheets were developed, one for each of the four texts. Instructions appeared at the top of the page asking the participants to recall, in as much detail as possible, a specific passage of the texts they had read. For example, the instructions for the cued passage recall associated with Text 1 were: "In as much detail as possible, please write down the part of the text you studied which described the red blood cells."

Text specific questionnaire. Four text specific questionnaires were developed, one for each of the four texts. These four questionnaires were each made up of five different questions. The instructions for the text specific questionnaire were identical to the instructions for the general knowledge questionnaire. These instructions appeared on a separate sheet, and were understood by the participants to pertain to both the text specific questionnaire and the heart inference questionnaire (which immediately followed the text specific questionnaire). The five questions for each text pertained to information unique to the corresponding text. Moreover, they involved memory of specific words used by each text. Each of the four questionnaires was one page long.

Heart inference questionnaire. Seven questions were developed for which the correct answers could not explicitly be found in any of the four texts. Correctly responding to these questions involved conceptual understanding of the human heart and circulatory system, rather than memory for specific sections of the text. A sample question is: "What effect, if any, would there be on the efficiency of blood

circulation if for some reason the valve between the right atrium and ventricle were unable to close completely?"

Texts. The four texts chosen discussed the anatomy, function, and purpose of the human heart and circulatory system. These four text were written for different audiences: Text A was written for elementary school children (Silverstein & Silverstein, 1983); Text B was written for adult readers as an introduction to the human heart (Davis & Park, 1981); Text C was written for undergraduate introductory anatomy students (Basmajian, 1982); Text D was written for medical students studying introductory level heart pathology (Gould, 1983). All illustrations and references to such illustrations were edited from these 4 texts. Each text was rewritten to be 17 paragraphs and approximately 4 type-written pages long. The number of words in the four texts were, 1,633, 1,533, 1,651, and 1,672, for Texts 1 to 4, respectively, with an average of 1,618. Each paragraph of the text was numbered.

## Design

The experiment employed a 4 (text condition) X 2 (delay condition) between subjects design. The delay condition was included to test whether any of our learning results were dependent on participants being tested within a certain period of time. There were 6 dependent measures: essay grade, LSA essay cosine, questionnaire score, sorting score, inference score, and text-specific questionnaire score. The essays were graded by professional graders at the Educational Testing Service. Two graders independently graded each essay on a five point scale. The graders were not informed which essays were pre- or post-essays, or which text the student had read. The essays were also evaluated using LSA, producing a 100 dimensional vector for each essay and its cosine with each relative target text. The sorting score was

derived by comparing the student's arrangement of the sort items to the average arrangement of experts. The questionnaire score, the inference score and the text-specific questionnaire scores were simply items correct on the respective questionnaire. For each dependent measure, please refer to the results section where that measure is discussed for more details.

Procedure

Each participant was randomly assigned to one of the four texts, and performed the posttext tasks either 5 minutes or 2 days after reading the text. Participants were tested individually or in groups of 2. All of the tasks were self-paced with the exception of the text study period.

The first task of the experimental procedure was the essay task. Participants were seated in front of a Macintosh personal computer with a standard word processing program running. They were orally instructed to write an essay of approximately 250 words about the anatomy, function, and purpose of the human heart and circulatory system. Participants were also given the printed essay task instruction sheet and asked to read it before starting. Participants were instructed on the use of the computer word counting function and told that they could use it to keep track of their progress with respect to the length of their essay. Participants were given as much time as they needed to complete the essay task.

Upon completion of the essay, participants performed the card sorting task. This task was performed in isolation from other participants. First the participants were given the 20 randomly ordered sorting cards and asked to read each of them aloud. Then they were orally instructed to place the cards in groups or piles such that ideas or

concepts that they believed belonged together appeared in the same pile, and that ideas or concepts that did not belong together did not appear in the same pile. The participants were told that there could be as many or as few groups or piles as they wanted, and that as many or as few cards as they wanted could appear in each pile. They were also told that they could re-arrange the cards as much as they wanted, and that they had as much time as they needed to arrange the cards.

The third experimental task was the general knowledge test. After the general knowledge test, each participant read one of the four texts, which were randomly assigned. The oral instructions were to study the text for 20 minutes, and do whatever they would normally do to study for a test. A blank piece of paper was also provided on which the participants could take notes if they wished to do so. Participants studied the text for the entire 20 minutes, which included going over material again if they finished before the time expired. A five minute notice was given when there were five minutes remaining in the task. At the end of the 20 minute study period, the text and any notes that were taken were collected.

Following the study phase, each participant received a break of either 5 minutes or 2 days. All of the posttext tasks were self paced. Participants first completed the cued passage recall questionnaire. The participants then responded to the appropriate set of text specific questions, followed by the inference questions.

Next, participants answered the general knowledge test questions again. This procedure was identical to the pre-text administration of the test. Following the general knowledge test, the same sorting task, and then the same essay task were performed. Finally, the participants filled out a demographic survey which

contained questions about their age, gender, school level, major, and questions about how much experience they had with the topic and how difficult they thought the various tasks were.

## RESULTS

We first analyze how well the questionnaire measures of knowledge, the pre-questionnaire score and the pre-essay grade, predict how much a student learns from the four different instructional texts. Then we describe the relation between the LSA measures and the questionnaire measures of background knowledge. In the final two sections we address the ability of the LSA measures to predict learning from text.

For all these analyses, the data were pooled over the two delay conditions. There was a small, though statistically significant main effect of delay for the questionnaire scores (participants had a higher proportion improvement score when they took the post test after a 5-minute delay ($\underline{M}$ = 0.32) than when they took the post test after a 2-day delay ($\underline{M}$ = 0.22), $\underline{F}(1, 86) = 5.36$, $\underline{p}$ = .02), but delay did not interact significantly with text or level of background knowledge. For the essay scores there were no significant delay effects at all ($\underline{M}$ = 0.35 for the 5-minute delay, $\underline{M}$ = 0.26 for the 2-day delay). Delay, therefore, was of negligible importance in this experiment.

### Questionnaire Measures of Learning

Average learning scores - that is the proportion improvement between pre- and post- scores on the knowledge questionnaire (learn-questionnaire) and the essay grades (learn-essay) - are shown in Table 1 for the four instructional texts. There was an overall effect of text on learn-questionnaire ($\underline{F}(3, 90) = 3.4$, $\underline{p}$ = .02) as well as on learn-essay ($\underline{F}(3, 90) = 3.00$, $\underline{p}$ = .04). For comparison, a control group of 19 undergraduate

participants who took the pre- and posttest but did not read a text had a mean Learn-questionnaire of .01. These control participants performed the same pretext tasks as all other participants. After performing the pre-text tasks, they took a five minute break and answered the knowledge questionnaire again.

The correlation over all participants between Learn-questionnaire and Learn-essay scores was .43, ($p < .01$). Table 2 shows the correlations between the learner's background knowledge and the amount learned separately for the four texts. As shown, for the medical students, who read only Text A, a nonsignificant relationship was present between background knowledge and amount learned for the questionnaire measure, but there was a significant negative relationship for the essay measure. Medical students who had higher grades on the essay task improved less than those who got lower grades. For the undergraduate participants, no significant correlation was obtained between background knowledge and amount learned for the two easiest texts - Text A and Text B - either for the questionnaire measure or the essay measure (See Table 2). For Text C, background knowledge was positively correlated with amount learned for both measures. For the most difficult text - Text D - a significant quadratic correlation was obtained for the questionnaire measure: Low-knowledge participants showed little or no learning, but the participants with the highest preknowledge learned more. (The columns labeled cosine pre-essay.textread in Table 2 will be discussed later).

One interpretation of these results, consistent with the zone-of-learnability hypothesis, is that for the medical students, Text A was generally too easy, but some medical students who did poorly on the

pre-essay improved the second time. In other words, the knowledge of medical students relative to Text A corresponded to range H in Figure 2. For the two easier texts - Texts A and B - the undergraduate's prior knowledge was just right for these texts, so that they learned quite well (Table 1), but no significant relationship between background knowledge and amount learned emerged (Table 2) because most of the participants were sampled from near the maximum learning point for the relevant prior knowledge distribution (Range F in Figure 2). Participants did not have quite as much prior knowledge relative to the more difficult text - Text C - so those who knew more learned more than those who knew less (the linear function in Table 2, corresponding to Range B in Figure 2), but over-all learning was still quite good (Table 1). For the most difficult text - Text D - the undergraduate participants over-all lacked adequate prior knowledge which resulted in poor learning (Table 1), with only the very best students capable of profiting from this difficult text (the quadratic function for pre-questionnaire in Table 2, corresponding to range G in Figure 2).

LSA Measures

Text effects. The four instructional texts were represented as vectors in the LSA space. The cosines between these vectors are shown in Table 3. The texts were selected in such a way that their expected difficulty increased from A to D. These judgments, which were based on human intuition, were corroborated by the LSA results: In every case, the cosines between texts we had judged to be more similar were indeed higher, and the overall pattern (taking 1 - cosine as an estimate of distance) is quite close to placing the four texts on a single line with

approximately equal spacing between the texts, Texts B and C being slightly closer to each other than to Texts A and D, respectively.

Relation between LSA and questionnaire measures. Over all texts, the three measures of knowledge used here - pre-questionnaire, pre-essay, and cosine pre-essay.standard (which we use as the LSA measure of background knowledge) - were all quite highly correlated: pre-questionnaire to pre-essay, $r = .74$; pre-questionnaire to cosine essay.standard, $r = .68$; and pre-essay to cos essay.standard, $r = .63$, all $p$s $< .01$. For comparison, the correlation between the two professional graders who scored the essays was .77. Thus, one can say that the LSA measure of knowledge is about as good as our questionnaire measures, and correlates with human graders almost as well as the human graders correlate among themselves.[1]

Pre and posttext effects. Figure 3 shows the average value of the cosines between the students' pre- and post- essays and the texts they actually read. Pre-cosines differed significantly between texts, with the cosines being highest for the easiest text and decreasing regularly as the texts increased in difficulty (overall $F(3,90) = 25.17$ , $p < .01$: For the linear component $F(1,90) = 23.28$, $p < .01$, assuming the four texts to be equally spaced on the underlying dimension, presumably one of text difficulty or content sophistication). Post-cosines also differed significantly among texts, $F(3,90) = 36.84$, but as Figure 3 shows, the difference was confined to the most difficult text. With the three easier texts - Texts A, B, and C - the students closely approached the text they read in their post reading essays, but they were not as successful with the most difficult text - Text D.

Text-specific and inference questions. The LSA measure was also related to the text specific questions and the inference questions. For

each text individually, cos pre-essay.textread and the text-specific questions had an average correlation of .27, while the correlation across texts of cos pre-essay.textread and text specific questions was .42, $p<.0001$. For the inference scores, the average correlation of cos pre-essay.textread and the inference questions for the texts individually was .27, whereas the correlation across texts of cos pre-essay.textread and inference questions was .32, $p<.01$. The cosine of each pre-essay with Text C, which we use as a measure of background knowledge, correlates with the text specific questions ($r = .25$, $p<.05$), as well as with the inference questions ($r = .41$, $p<.0001$).

Matching Readers to Texts: Correlational Analyses

The average cosine between the students' essays and the text they read can be used to predict the proportion improvement scores for the general knowledge test and the essay grades, learn-questionnaire and learn-essay. The data are shown in Figure 4 for the four groups of college students who read Texts A, B, C, and D, respectively, as well as for the medical students who read only Text A. The latter were included in this analysis because none of the texts was obviously too easy for the college students; to test the zone-of-learnability hypothesis we needed a group of learners who read a text that was clearly too easy for them. The curves fitted to the points in Figure 4 are second-order polynomials. Figure 4 confirms the zone-of-learnability predictions. It seems noteworthy that the functions estimated for Learn-questionnaire and Learn-essay are almost identical in shape.

As with the questionnaire measures (pre-questionnaire and pre-essay), we also examine the ability of the LSA data - the cosine measure - to predict learning (learn-questionnaire and learn-essay) within each text.[2] The data are also shown in Table 2, together with the

questionnaire measures. The cosine measures in Table 2 are the cosines between the pre-essay a student wrote and the text he or she read. The point to be emphasized here is that in general the cosine correlations mirror the correlations obtained with the hand coded measures. There are no significant correlations between background knowledge and learning for the two easiest texts - Texts A and B - however measured. We interpret this to mean that with respect to those two texts the students' prior knowledge was in an intermediate range, where learning is good, but one would not necessarily expect to find a strong relationship between the amount of prior knowledge and the amount of learning (range F in Figure 2). For Text C, there was a positive relationship between background knowledge and learning for all four measures, presumably because this more difficult text fell into the range where learnability rapidly increases as a function of prior knowledge (range B in Figure 2). For the most difficult text, the results were less consistent, but still explainable within the zone-of learnability framework if one assumes that for most learners this text was beyond the zone of learnability. The cosine pre-essay.textD measure fell sufficiently within the learning zone to result in a positive relationship between knowledge and learning for the learn-questionnaire measure (range B from Figure 2), and a positive quadratic relationship between knowledge and learning for the learn-essay measure (Range G from Figure 2). The significant positive quadratic relationship between pre-questionnaire and learn-questionnaire confirms this interpretation; the lowest knowledge learners showed no relationship between knowledge and learning and little or no learning, but the few most knowledgeable students did manage to learn something, yielding a non-linear function (corresponding to Range G in Figure 2).

Matching Readers to Texts: Theoretical Analyses

The data analyses presented above are consistent with the theoretical expectations that were described at the beginning of this article. However, there are a number of factors that complicate the relationship between the theory and our data. One problem is that learning scores are, of course, only imperfectly predictable from prior knowledge, however measured (by questionnaire score, essay grade, or an LSA cosine). The other problem is the restriction of range in prior knowledge scores. We do not have students whose prior knowledge is evenly distributed over the whole range, from very low to very high. Thus, we rarely can observe the whole function, and usually can see only sections of the expected functions. These appear to fit into the general picture, but it is not a clear and obvious picture that emerges from these data. One way to get around this dilemma is to analyze the data as if they were indeed samples from Gaussian functions, as assumed in Figures 1 and 2.

A standard Gaussian curve has three free parameters: the peak of the function ($\theta$), the width of the function ($\sigma$), and the area under the curve ($\underline{k}$). For the purposes of this qualitative analysis of the data, we are primarily interested in how the peak of the proposed zone of learnability shifts as a function of the difficulty of the text, as it interacts with the learners' prior knowledge. In other words, as the conceptual difficulty of the target text increases, we expect the zone of learnability to shift to center on higher pre-knowledge readers. This predicted shift is manifest in shifts of the peak of the function, or changes in $\theta$. Therefore, in fitting Gaussians for each of the four text conditions, we allowed $\theta$ to vary, while keeping $\sigma$ and $\underline{k}$ constant.

Using a standard hill-climbing algorithm, we first fit a Gaussian curve to the pre-questionnaire by Learn-questionnaire relationship for the Text B condition alone. Text B was selected for this purpose because it suffered less from the restricted-range problem than other texts, allowing us to estimate the shape parameters of the Gaussian function. The values of $\sigma$ (11.77) and $\underline{k}$ (5.70), the two parameters other than the peak, were then used to constrain the shape of the fits for the remaining three text conditions.

As previously discussed, the peak of the function was allowed to vary and is the focus of our qualitative interpretation of this analysis. The estimated peaks of the four text conditions were found to arrange themselves in the predicted order, as shown in Figure 5. The main effect of text on the peak is highly significant, $\underline{F}$ (3,87) = 24.61, $\underline{p}$<.0001 (note that these analyses do not include the medical students). Thus, we confirm our hypothesis that the zone of learnability shifts to higher ranges of pre-questionnaire as the difficulty of the text increases.

A similar analysis can be conducted using the cosine between the participant pre-essays and the text they read to predict Learn-questionnaire, for each of the four text conditions. In this case, however, the predictor variable (i.e., cos pre-essay.textread) is a relative measure, not an absolute measure as in the previous analysis. Thus, where in the previous analysis we expected the fitted Gaussians to shift such that the peak of the curves includes higher levels of pre-questionnaire for harder texts, in this analysis, the zone-of learnability hypothesis would predict that the four fitted Gaussians would all peak approximately over the same cosine value.

Using the same computational procedure as described above, the values of $\sigma$ (0.217) and $\underline{k}$ (0.092) were established from the Text B

relation, and were then used to constrain the Gaussian fits produced for the remaining three text conditions. Figure 6 shows that the peaks of the functions for the four texts have moved much more closely together in comparison with Figure 5, where absolute rather than relative knowledge was plotted on the abscissa. Whereas the peaks of the four text conditions were still significantly different, $F(3,87) = 3.24$, $p<.05$, they are clustered more closely than they were in the pre-questionnaire analysis. This clustering of peaks suggests that, consistent with the zone-of-learnability hypothesis, that when using a relative measure of learner knowledge as the basis for predictions of learning, one optimal zone of learnability emerges. [3] This optimal zone of learnability can be estimated explicitly by fitting the learning data from all four groups simultaneously to one Gaussian function. The results are presented in Figure 7.

The Gaussian fit of the cosine data shown in Figure 7 could be used to assign texts to students in order to optimize learning. The process of matching readers to texts would involve having each student write an essay about the heart, and then calculating the cosine between the essay and each of the four texts. Each cosine score would be placed on the Gaussian fit in Figure 7, and the text with the highest ordinate value would be predicted to be the text that the student would learn the most from. Using this method, we predict that the learning scores of our participants would have improved 53%, compared to the actual situation where participants were randomly assigned texts to read.

It is important to consider how the results of the theoretical model fitting presented in this section generalize beyond this particular study. Consider four particular aspects of the current study. First, would

our results generalize to new instructional texts about the heart? In other words, can the optimal zone of learnability in Figure 7, which includes the suggestion that a cosine of 0.54 produces optimal learning, be expected to apply to other texts? Our answer is a cautious yes, because when a relative measure of pre-knowledge was employed (cosines), roughly the same zone of learnability function emerged for all four instructional texts (Figure 6). This supports the view that there is a common scale of knowledge, and distance from a text on that scale predicts learning independently of the absolute position on the scale. Of course, the possibility exists that a new instructional text may differ from the four texts used in this study in ways other than level of knowledge (e.g., style of writing). These differences may get reflected in the cosine between student essays and the new text, and the empirical relationship presented in Figure 7 may no longer apply.

Second, can our results be expected to generalize to different subject populations? Once again, the answer is a cautious yes, because the theoretical model we assume includes an individual difference parameter, namely, the prior level of knowledge of a subject. Thus, subject populations whose level of domain knowledge differs from the population used in this study will be placed at the appropriate position in the zone of learnability according to the model presented in Figure 7. However, we have identified a problem that arises when subjects possess more knowledge than an instructional text. Because a cosine is an undirected measure of relatedness, a cosine between a subject's essay and an instructional text is unable to reliably determine whether the subject is above or below the knowledge level of the text. This problem does not arise in the current study because all our subjects possessed less knowledge than that contained in the four instructional texts. (For

a complete discussion of this problem, which we have come to call the directionality problem, see Rehder, et al. (1998/this issue). Another reason Figure 7 may not generalize to other subject populations is that a new population may differ from the present one in ways other than level of knowledge (e.g., style of writing) in ways that change the relation between cosine and level of knowledge, making the model inapplicable to the new population.

Third, would our results generalize to new experimental instructions that changed the contents of peoples' pre-essays, for example by making them more or less specific to the topic? Here the answer is probably not, because it is easy to imagine how the cosines between our subjects' pre-essays and the instructional texts would have increased if we had been more specific about what we asked our subjects to write, eliminating from the essays discussions about irrelevant topics such as exercise, cholesterol, etc. Obviously, this increase in cosine does not reflect an increase in our subjects' knowledge about the heart, and, therefore, the empirical relationship expressed in Figure 7 would no longer apply.

Fourth and finally, would the specific results generalize to different content domains? The answer is probably not, because a different content domain will require a different singular value decomposition (SVD) space, and there is no reason to believe that the cosines have the same meaning across different SVD spaces.

Sort Data

For each participant, we created both a pre-sort and a post-sort co-occurrence matrix in which items which were grouped into the same pile during the sorting task received a score of 1 and items which were not grouped together received a score of 0. These sorting matrices were

then averaged across all participants and also across all participants within each of the four text conditions.[4] Thus, this sorting measure implicitly estimates the clustering of related terms within the participants' representations of the target topic.

Two additional methods were used to assess the alternate clusterings of the sort terms: cosines and expert ratings. To estimate the clustering of the sort terms within the LSA space, cosines between all 171 possible pairs of 19 of the 20 sort terms were produced.[5] To estimate the clustering of terms in an expert-level representation of the target topic, we had five medical professionals (three medical doctors and two nurses) rate, on a scale from 0 to 10, the similarity of each of the 171 pairs of terms. The five ratings for each pair were then averaged to produce one expert rating for each of the 171 pairs.

The correlations to the cosine-clustering and to the expert-ratings for the pre-sort and the post-sort averages for both the undergraduates and the medical students are given in Figures 8 and 9. Overall, the correlation to the expert ratings for both the Medical students and the undergraduates were low, rs range from 0.33 to 0.49. This may be due to task differences. The Medical students' average sorting pattern hardly changes between pre-sort and post-sort tasks ($r = 0.93$). The shifts the undergraduates make from the pre-sort to the post-sort ($r = 0.88$) tend to make them look more like the experts, or at least look as much like the experts as the medical students do. Interestingly, the opposite is true for the comparisons to the cosine clustering. Across both pre-sort and post-sort tasks, the Undergraduates' clustering tends to be more similar to the cosine-similarity than is the Medical students' clustering. Furthermore, both the undergraduates' and the Medical students' clusterings become less like the cosine-similarity in the post-

sort task. One interpretation of this pattern is that the representation or grouping of information in the LSA space is at a relatively basic or introductory level. Indeed, the correlation between the cosine-similarity and average expert similarity rating is relatively low, $r = 0.27$, $p<.001$.

This pattern of starting out like the LSA space, but becoming more like the experts is particularly apparent when comparing the undergraduates' pre-sort to post-sort shifts for each of the 4 text conditions. The values of these correlations, along with the differences between these correlations are given in Table 4. On average, the undergraduates in each of the text conditions changed their sorting patterns in such a way as to become more like the experts, particularly for those who read Texts B and C. Conversely, in all four text conditions, the undergraduates' sorting patterns became slightly less like the cosine-clustering after having read the text, particularly for Text B. Recall that in these studies, LSA's representation (it's "knowledge") is based on analysis of just 36 rather heterogeneous and idiosyncratic encyclopedia articles with some heart related content, and a total corpus size of about 40 pages of text, and it does not include the text they read in the experiment.

## Discussion

That background knowledge, in general, facilitates learning has been demonstrated repeatedly (e.g., Schneider et al., 1989; Spilich et al., 1979). The present results both confirm and modify this conclusion. We found evidence of non-monotonic relations between the amount learned from a text and background knowledge. Students whose prior knowledge did not overlap enough with the contents of the text did not learn well, but neither did students whose knowledge overlapped

too much with the contents of a text. Our interpretation is that a text was best for learning when it offered the student enough hooks to link it to prior knowledge, but still contained enough new information to be acquired. Thus, we show that the relation between the content of an instructional text and the learner's background knowledge is a factor in learning from texts. In a separate series of experiments (McNamara, et. al., 1996), it was shown that, if the content of an instructional text is held constant, learners at different background levels require a different style of writing: An explicit, highly coherent style is best for low-knowledge students, while high-knowledge students profit more when they actively must fill in coherence gaps when they study a text (E. Kintsch & W. Kintsch, 1995; McNamara & Kintsch, 1996; McNamara et al, 1996; Voss & Silfies, 1996). Thus, although the general wisdom that background knowledge facilitates learning still holds, the relationship between learning and background knowledge is not a simple one, as it is modified by both content and style of an instructional text.

We were able to detect non-monotonic relationships in our data by examining domain knowledge as a continuous variable rather than a dichotomous variable. In addition, within each text, we found instances of both monotonic and non-monotonic relations that were mutually interpretable within the zone-of-learnability hypothesis. Thus, treating knowledge as a continuous variable may prove to be more informative in future research than dichotomizing participants into high- and low-knowledge groups.

This study focused upon three questions. How well can we measure domain knowledge? How well can we predict the success of learning from text on the basis of an assessment of the student's

domain knowledge? How well does LSA do as a measure of domain knowledge and a predictor of learning in comparison with questionnaire-type measures?

As far as knowledge assessment is concerned, all we can claim is that the three instruments we used achieved a high degree of reliability. The score on the questionnaire, the essay grade, and the LSA measure were measuring largely the same thing. The intercorrelation between the two professional essay graders ($\underline{r}$=.77) was of about the same magnitude as the correlation between the essay grade and the questionnaire scores ($\underline{r}$=.74), as well as between the LSA measure (the cosine between the pre-essay and Text C) and the essay grades ($\underline{r}$ = .63), and the LSA measure and the questionnaire scores ($\underline{r}$ = .68). Thus, one might hope that all these methods yield valid measures of how much a student knows about the functioning of the heart.

The learning measures are less highly intercorrelated ($\underline{r}$=.43) than the prior knowledge measures. This may in part be a consequence of our choice of proportion of possible improvement as the measure of learning: this measure is a function of two random variables (pre- and post-scores), and hence necessarily has a greater variance than either measure alone. Therefore, we are limited by the lower reliability of the learning measures in our effort to predict the success of learning from the student's prior knowledge. The expected function relating prior knowledge and learning depends on the level of prior knowledge as well as the difficulty of the instructional text. Thus, a complex but orderly pattern of relation was expected. The data (see Table 2) were generally in agreement with these expectations. One cannot make strong claims on the basis of such complex results, but on the whole they lend support to the predictions we derived from the zone-of-

learnability hypothesis. If we examine the learning data from each of the four texts for the undergraduates and medical students in terms of the theoretical curve plotted in Figure 2, we see an orderly pattern. At the high end of the scale of relatedness or prior knowledge and test difficulty, the medical students learned very little overall and in the case of the essay task, showed a negative slope between prior knowledge and learning. Moving down the axis of relatedness of prior knowledge and text, we found no significant relationships between knowledge and learning for Texts A and B, perhaps because they were optimal for learning and thus fell at the top of the curve. For Text C, the slightly harder text, there is an improvement in learning as a function of increasing background knowledge. Finally, Text D, the most difficult text, appears to fall on the low end of the relatedness scale, as evidence by only the highest knowledge students being able to learn from it.

Indeed, good arguments can be made that more unequivocal results could hardly be expected. One problem in a study using college students as participants is that it is highly unlikely that the whole range of prior knowledge, from next to nothing to professional expertise, would be represented among the students. We tried to ameliorate this problem by including a group of medical students as our participants, but this does not completely solve the problem. It would be difficult indeed to find a significant knowledge domain and a group of students that vary over the whole range of prior knowledge but are otherwise comparable.

A second problem in any experiment on learning as a function of prior knowledge is that prior knowledge is only one determinant of learning - an important one, but by no means the only one. Thus, we

find in our results over and over again students who learned very little, in spite of the fact that their prior knowledge appeared to be at an optimal level for learning. Obviously, how well a student learns depends on other factors, too, primarily, one might suppose, motivational ones. However, we did not find students who learned a lot when they knew either too much or too little according to our analyses. Thus, an appropriate level of prior knowledge relative to the instructional text appears to be a necessary but not a sufficient condition for learning.

The third question, how well LSA could do in comparison with other measures of knowledge, received a clear answer: very well indeed, just about as well as the reliability of the questionnaire measures allowed. The intercorrelation between the LSA derived measure of prior knowledge, the semantic relatedness of the essay a student wrote and the instructional text, are of the same order of magnitude (r=.63 and .68, respectively) as the intercorrelation of the other measures among themselves. Thus, automatically evaluating a student's essay by means of LSA gives results comparable to grading an essay by a human expert or using a knowledge questionnaire.

We have elsewhere explored automatic essay grading in greater detail (Landauer, Laham &, Foltz, 1998). We merely remark here on the considerable theoretical as well as practical interest of automatic essay grading. For theories of comprehension, it is notable that LSA achieves this level of performance without being able to distinguish true and false statements (as well as differences carried by syntax, such as argument order and scope, anaphor, etc.). The essays that received low grades did so, at least in part, one supposes because the graders noted logical errors and misconceptions in them. LSA cannot detect a logical

misconception; it is sensitive only to word usage. Thus, it appears students who make errors in their essays not only say wrong things, but do not use the right words in quite the right way. That is what LSA picks up, not the errors directly. LSA could easily be fooled by an expert who uses the language perfectly but turns it all into nonsense - for example, by writing an essay, then randomizing the words. But lying convincingly to LSA would be difficult. It will not be deceived by people who know very little, because these people will betray themselves by the way they write. Moreover, it would probably be extremely difficult even for an expert to write an essay with just the right combination of words without knowing enough to write a good essay by any other criterion. The results on automatic essay grading are also of interest for practical and research purposes. It is very easy to compute the cosine between a student's essay and a standard comparison text, even if the teacher or researcher does not understand much about either. But it is hard work to grade an essay, requiring appropriate expertise and intense intellectual effort. Questionnaires are easy to score, but to design a discriminating questionnaire in the first place requires a great deal of work and pre-testing and demands a great deal of expertise.

The zone-of-learnability hypothesis was supported most clearly by a plot of the average learning scores for the four texts as a function of the average cosine between the students' essays and the text they read, including a group of medical students who read Text A. The plot in Figure 4 is clearly non-monotonic. Thus, by combining information from all groups the basic relationship emerges, while within each group only segments of this underlying relationship can be observed (as attested by the pattern of correlation's in Table 2).

Further support for the zone-of-learnability hypothesis was provided by fitting Gaussian functions to the plots of learning from text as a function of prior knowledge for the four instructional texts used in this study (as assessed by the questionnaire score). In this analysis, four distinct functions are obtained, one for each text. For low scorers, the easiest Text A is optimal, for most students either Text B or C, which are of medium difficulty is optimal, and Text D is too hard for everyone (Figure 5). If, on the other hand, prior knowledge is assessed by a relative measure, the cosine between the essay vector and the text vector, these four curves tend to merge (Figure 6), suggesting that a cosine around .5 between a student's prior essay and the instructional text often yielded the best results. We mimicked using the Gaussian curve fits to assign texts to students by calculating the cosine of a students' essay to each of the four texts, and returning the text that has the highest ordinate value. This analysis suggests that if students had not been assigned randomly to the four different instructional texts, but had read the text so chosen to match the student's level of prior knowledge, learning could have been improved by about 53%. This is a huge potential improvement, and although our calculations are, of course, specific to the present student and text populations, practical applications of LSA for the purpose of matching learners and text, especially in the context of information rich environments such as the WWW, might be well worth pursuing.

This study has shown LSA to be a practical tool for characterizing the information contained in students essays based on their word usage and for measuring the relatedness of different essays based on their content. As such, these results lend strong support to the view of LSA as a statistical method for capturing the underlying relationships

between documents. These results also suggest LSA to be an good tool for any aspect of discourse research in which experimenters wish to characterize the content of a set of words or documents. Our results are also consistent with the view of LSA as a theory of knowledge representation. According to this view, the essays students write are reflections of their understanding of the material. What LSA represents is the mental structure that is latent in the essay. By comparing this representation to the representations of the information contained in the texts, LSA is providing a measure of the underlying semantic similarity of the two documents. This process can be viewed as akin to the method a teacher might use in reading essays and assigning students to texts of varying difficulty based on the content of the essays. The teacher could make a general assessment of the knowledge of the student based on their essay, and compare that assessment to the information content of the different texts. For LSA, the vector would represent the knowledge of the student about the topic, and would be compared to the vectors that represent the information contained in the texts.

References

Basmajian, J. V. (1982). Primary Anatomy. Baltimore/London: Williams & Wilkins.

Britton, B. K., & Gulgoz, S. (1991). Using Kintsch's computational model to improve instructional text: Effects of inference calls on recall and cognitive structures. Journal of Educational Psychology, 83, 329-345.

Davis, G. P., & Park, E. (1981). The Heart: The Living Pump. Washington, D. C.: U.S. News Books.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 41, 391-407.

Gould, S. E. (Ed.) (1983). Pathology of the Heart. Springfield, IL: Charles C Thomas.

Kintsch, E., & Kintsch, W. (1995). Strategies to promote active learning from text: Individual differences in background and knowledge. Swiss Journal of Psychology, 54, 141-151.

Kintsch, W. (1994). Text comprehension, memory, and learning. American Psychologist, 49, 294-303.

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

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

Landauer, T. K, Laham, D, & Foltz, P. W., (1998). Computer-based grading of the conceptual content of essays. Manuscript in preparation.

McKeown, M. G., Beck, I. L., Sinatra, G. M., & Loxterman, J. A. (1992). The contribution of prior knowledge and coherent text to comprehension. Reading Research Quarterly, 27, 79-93.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. Cognition and Instruction, 14, 1-43.

McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. Discourse Processes, 22, 247-288.

Means, M., & Voss, J. (1985). Star Wars: A developmental study of expert and novice knowledge structures. Memory and Language, 24, 746-757.

Moravcsik, J. E., & Kintsch, W. (1993). Writing quality, reading skills, and domain knowledge as factors in text comprehension. Canadian Journal of Experimental Psychology, 47, 360-374.

Rehder, B., Schreiner, M. E., Wolfe, M. B. W., Laham, D. Landauer, T. K, & Kintsch, W. (this volume). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. Discourse Processes.

Schneider, W., Körkel, J., & Weinert, F. E. (1990). Expert knowledge, general abilities, and text processing. In W. Schneider, & F. E. Weinert (Ed.), Interactions among aptitudes, strategies, and knowledge in cognitive performance (pp. 235-251). New York: Springer Verlag.

Silverstein, A., & Silverstein, V. B. (1983). Heartbeats: Your Body, Your Heart. New York: J. B. Lippincott.

Spilich, G. J., Vesonder, G. T., Chiesi, H. L., & Voss, J. F. (1979). Text processing of domain related information for individuals with high and low domain knowledge. Journal of Verbal Learning and Verbal Behavior, 18, 275-290.

Voss, J. F., & Silfies, L. N. (1996). Learning from history text: The interaction of knowledge and comprehension skill with text structure. Cognition and Instruction, 14, 45-68.

# Appendix

Instructions for the essay task:

In this part of the experiment we would like you to write down what you know about the human heart and circulatory system. Your essay should be approximately 250 words. We would like for you to be as specific as possible is discussing both the anatomy, function, and purpose of the heart and circulatory system. If you feel that you would ideally write either more or less than 250 words, please keep in mind that it is essential that all students write approximately the same amount. therefore please be as detailed as you can in that amount of space. 250 words is approximately one full double-spaced page.

Sort task word

| | |
|---|---|
| hemoglobin | oxygen ($O_2$) |
| carbon dioxide ($CO_2$) | capillaries |
| alveoli | SA node |
| pacemaker | contractions |
| heart rate | one-way |
| valves | right atrium |
| right ventricle | pulmonary artery |
| lungs | larger |
| left atrium | left ventricle |
| aorta | body |

## Footnotes

[1] By another technique, to be reported elsewhere, the scores on the essays were calculated by LSA on the basis of each essay's cosine similarity to essays previously scored by human experts. With this method, LSA scores were as highly correlated with the experts as they were with each other, and slightly better predictors of pre-questionnaire.

[2] In addition to using the questionnaire measures to determine learning, we examined the ability of LSA to provide a measure of learning. For a discussion of these results, see Rehder et al., (1998/this issue).

[3] A similar analysis of the pre-essay and learn-essay scores could not be performed because the range of pre-essay scores was so severely restricted that it was not possible to estimate the shape of the Gaussian distribution to be used.

[4] Because the individual sort data were quite variable and binary-to-continuous correlations had to be used, only group data could be analyzed.

[5] The term alveoli did not appear in the LSA space. As a result, the term alveoli was eliminated from all of the sorting analyses, resulting in 19 rather than 20 sorting terms.

## Author Note

Michael B. W. Wolfe, M. E. Schreiner, Bob Rehder, Darrell Laham, Walter Kintsch, and Thomas K Landauer, Department of Psychology, University of Colorado, Boulder. Peter W. Foltz, Department of Psychology, New Mexico State University.

This research was supported by a grant from the National Institute of
Mental Health, MH -15872 to W. Kintsch and a contract from ARPA-CAETI to T. Landauer and W. Kintsch.

Correspondence concerning this article should be addressed to Michael B.W. Wolfe, Department of Psychology (Cognitive), Campus Box 345, University of Colorado, Boulder, CO, 80309. Electronic mail may be sent via internet to mwolfe@clipr.colorado.edu.

Table 1

Mean learning scores for questionnaire and essay for the four instructional texts.

| | Learn-questionnaire | Learn-essay |
|---|---|---|
| Text A (easiest) | .26 | .39 |
| Text B | .37 | .32 |
| Text C | .27 | .36 |
| Text D (hardest) | .19 | .17 |

Table 2

The correlations between two measures of learning (Learn-questionnaire & Learn-essay) and two questionnaire predictors (Pre-questionnaire & Pre-essay) and the cosine measure from the LSA analysis for the four texts. Quadratic correlations are in parentheses.

| | | Learn-questionnaire | | Learn-essay | |
|---|---|---|---|---|---|
| Participants | Text | vs. Pre-questionnaire | vs. cos pre-essay.textread | vs. Pre-essay | vs. cos pre-essay.textread |
| med. student | A (easiest) | .06 | -.20 | -.88** | -.42 |
| undergrad | A | .00 | -.16 | -.31 | .30 |
| undergrad | B | .01 | -.13 | -.21 | -.09 |
| undergrad | C | .59** | .43* | .73** | .58** |
| undergrad | D (hardest) | (.46*) | .54** | -.22 | (.43*) |

Table 3

Cosines between the four instructional texts.

| | Text A | Text B | Text C | Text D |
|---|---|---|---|---|
| Text A (easiest) | - | .82 | .70 | .52 |
| Text B | | - | .80 | .63 |
| Text C | | | - | .81 |
| Text D (hardest) | | | | - |

Table 4

Correlations of average clustering frequencies of heart terms by undergraduates with expert judgments of similarity and LSA cosine measures of similarity.

| | Expert-rated similarity | | | Cosine-similarity | | |
|---|---|---|---|---|---|---|
| | Pre: | Post: | Diff: | Pre: | Post: | Diff: |
| Text A (easiest): | 0.354 | 0.392 | 0.038 | 0.751 | 0.736 | -0.014 |
| Text B: | 0.334 | 0.469 | 0.135 | 0.784 | 0.633 | -0.151 |
| Text C: | 0.280 | 0.472 | 0.192 | 0.747 | 0.716 | -0.031 |
| Text D (hardest): | 0.316 | 0.408 | 0.092 | 0.763 | 0.684 | -0.079 |

Figure Caption

Figure 1. Theoretical relationship between background knowledge and learning.

Figure 2. Theoretical relationship between the prior knowledge/text match and learning. This non-monotonic function could be borne out in the data by a number of different specific functions. The letters denote different potential relationships between prior knowledge/text match and learning, as discussed in the text.

Figure 3. Average pre- and post-cosines between the students' essays and the text they read.

Figure 4. Average learning scores (learn-questionnaire and learn-essay) for four groups of college students who read texts A-D, and a group of medical students who read Text A as a function of the average cosine between the students' prior essays and the text they read. For each cosine value, there are two learning scores, Learn-questionnaire and Learn-essay. $A_{med}$ is the essay-based improvement score for the medical students, who read only text A.

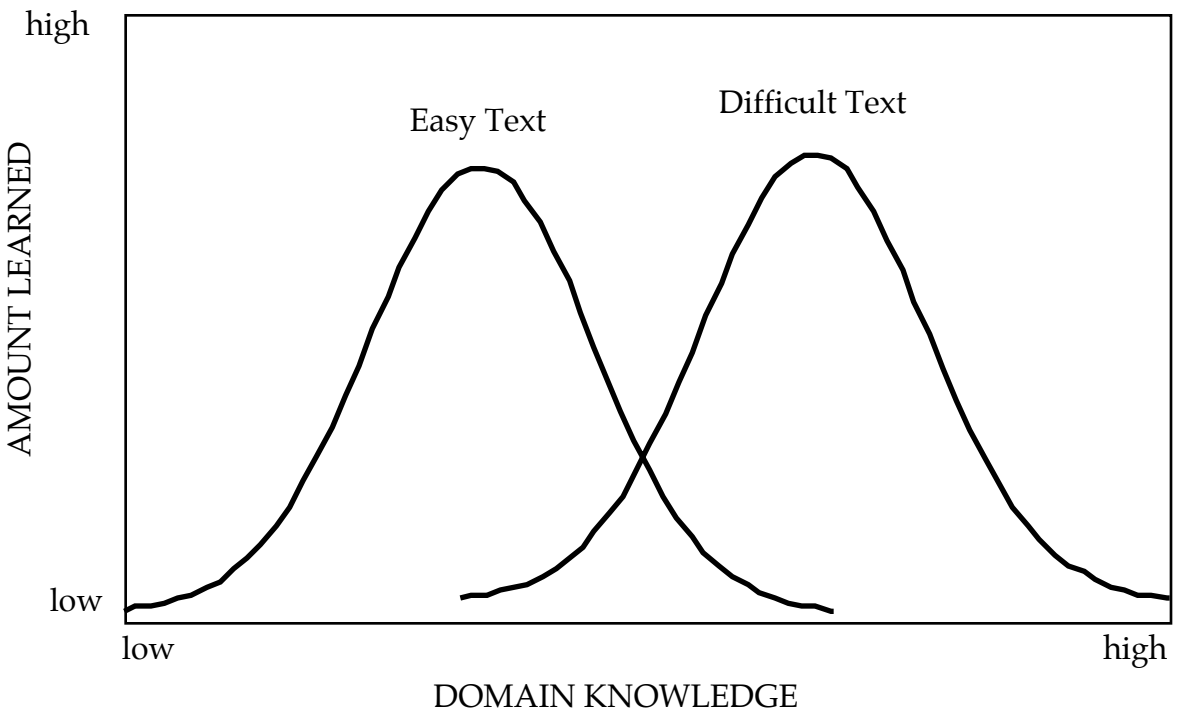
Figure 5. Fitted Gaussian curve of the learn-questionnaire and pre-questionnaire relationship for each of the four text conditions.

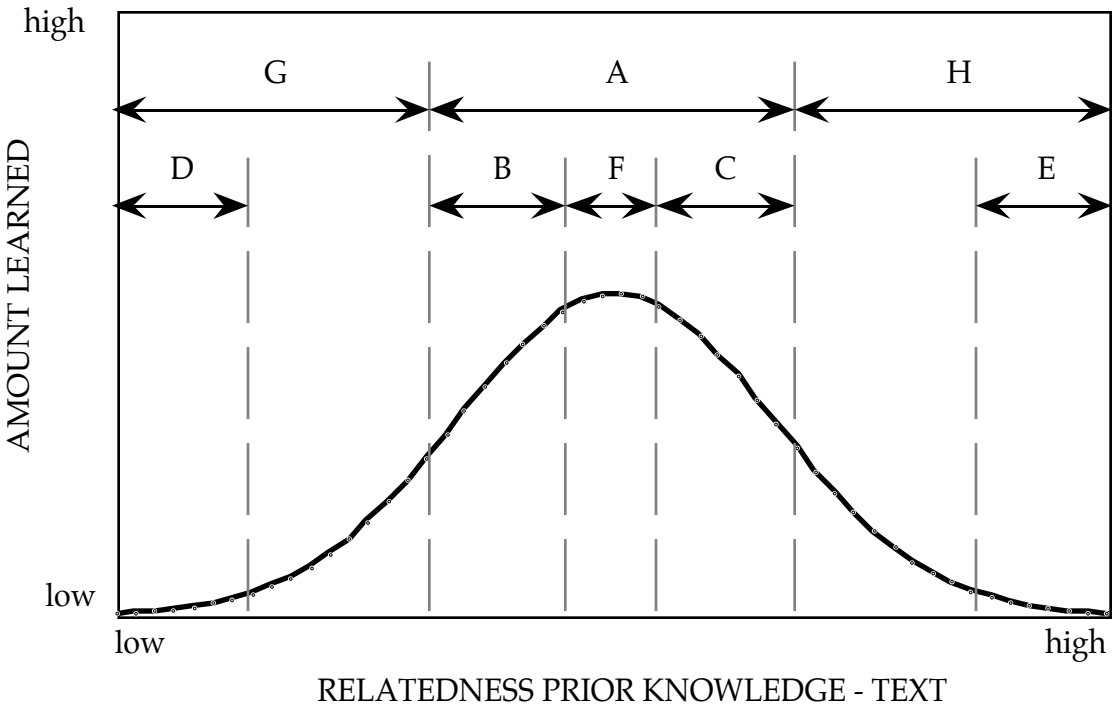
Figure 6. Fitted Gaussian curves of the learn-questionnaire and cos pre-essay.textread relationship for each of the four text conditions.

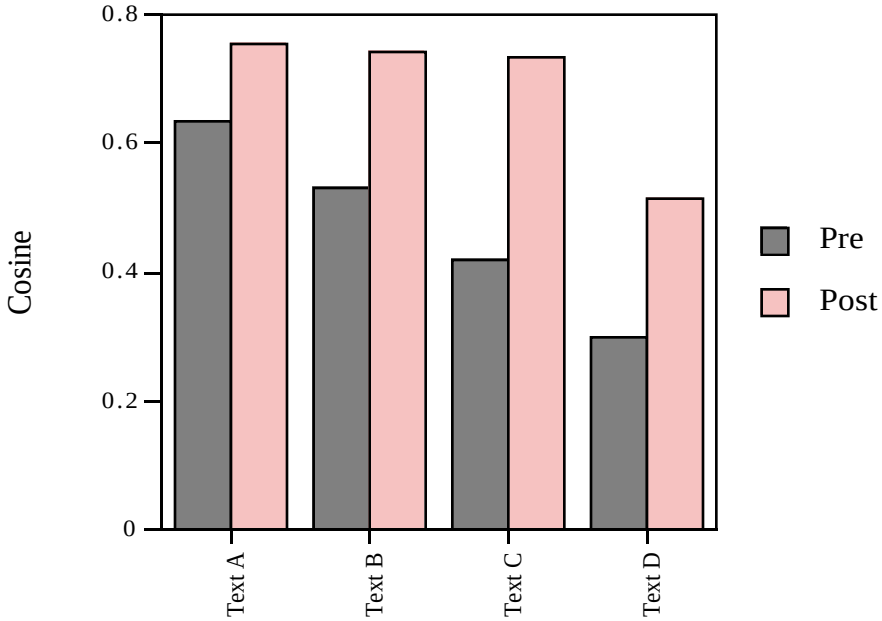
Figure 7. Fitted Gaussian curve of the learn-questionnaire and cos pre-essay.textread relationship over all four text conditions.

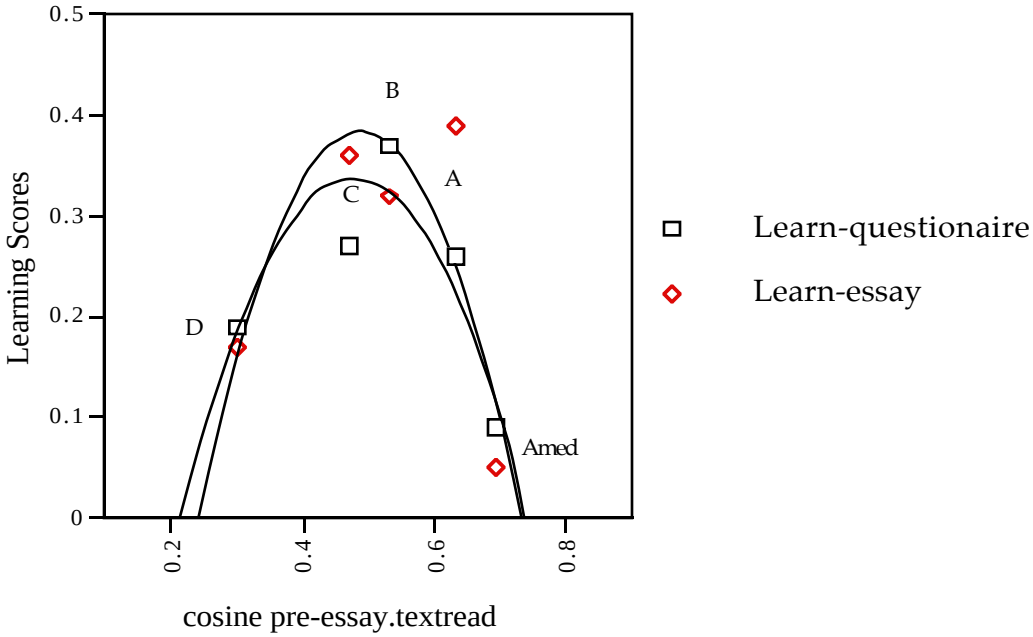
Figure 9. Correlations with expert ratings for average pre and post sort patterns of medical students and of undergraduates.

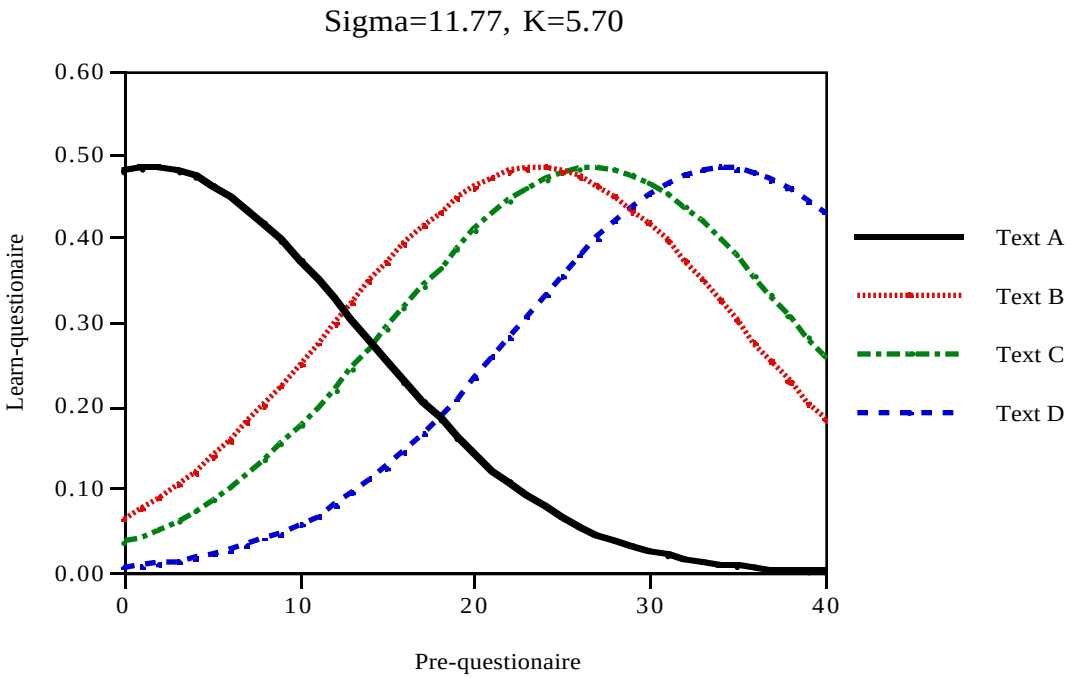
Figure 9. Correlations with cosine similarity for average pre and post sort patterns of medical students and of undergraduates.

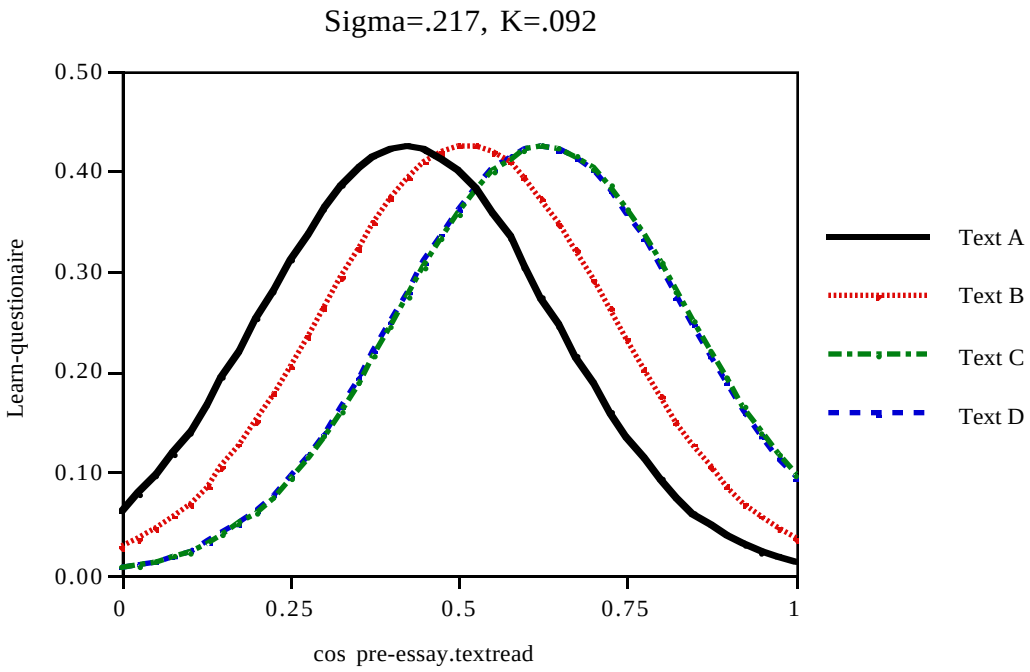

high
low
AMOUNT LEARNED
Easy Text
Difficult Text
low
high
DOMAIN KNOWLEDGE

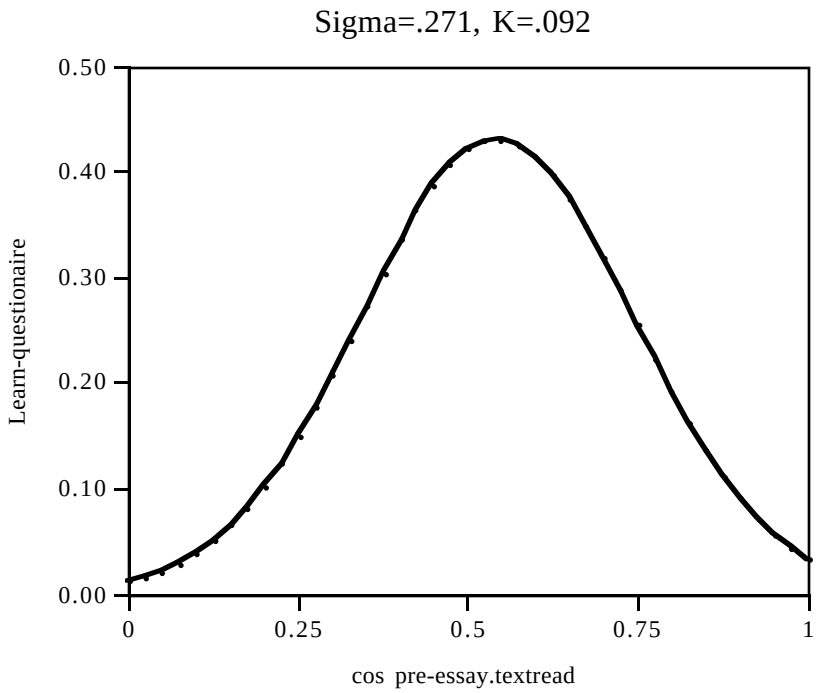

high
G
A
H
D
B
F
C
E
AMOUNT LEARNED
low
low
high
RELATEDNESS PRIOR KNOWLEDGE - TEXT

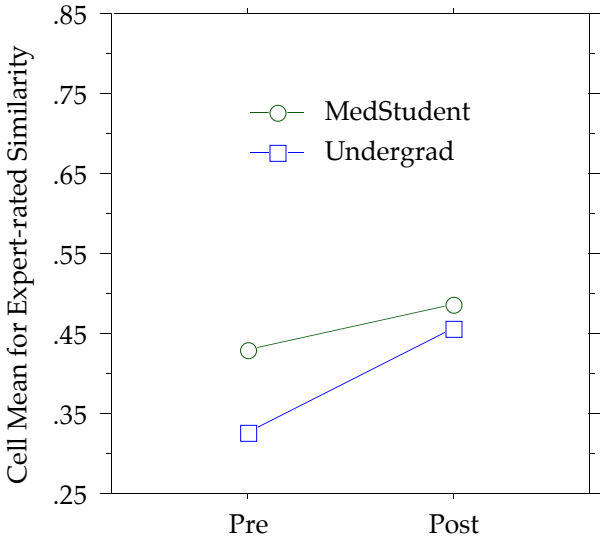

0.8
0.6
0.4
0.2
0
Cosine
Text A
Text B
Text C
Text D
Pre
Post

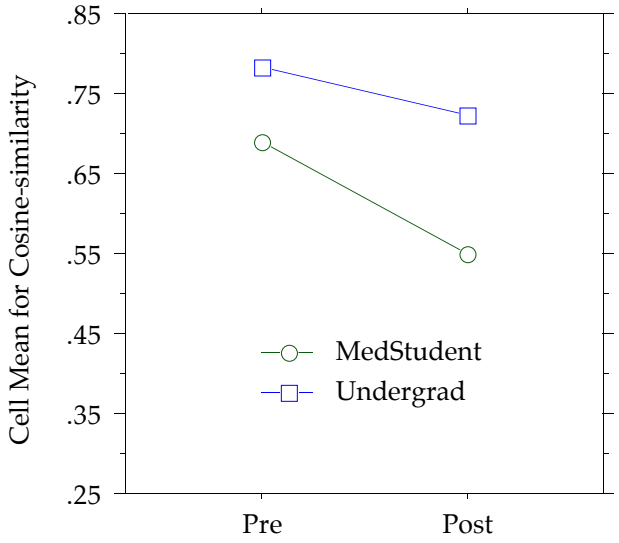

0.5
0.4
0.3
0.2
0.1
0
Learning Scores
B
A
C
D
Amed
0.2
0.4
0.6
0.8
cosine pre-essay.textread
Learn-questionaire
Learn-essay


Sigma=11.77, K=5.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Learn-questionaire
0
10
20
30
40
Pre-questionaire
Text A
Text B
Text C
Text D


Sigma=.217, K=.092
0.50
0.40
0.30
0.20
0.10
0.00
Learn-questionaire
0
0.25
0.5
0.75
1
cos pre-essay.textread
Text A
Text B
Text C
Text D


Sigma=.271, K=.092
0.50
0.40
0.30
0.20
0.10
0.00
Learn-questionaire
0
0.25
0.5
0.75
1
cos pre-essay.textread


Cell Mean for Expert-rated Similarity
.85
.75
.65
.55
.45
.35
.25
MedStudent
Undergrad
Pre
Post


Cell Mean for Cosine-similarity
.85
.75
.65
.55
.45
.35
.25
Pre
Post
MedStudent
Undergrad